THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Evaluation of the obesity paradox in diabetes: a longitudinal case control study

Ebenezer Senyo Owusu Adjah

MSc, Epidemiology and Biostatistics
BSc, Biochemistry

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2019

Faculty of Medicine

Abstract

In the general population, obesity is associated with significantly higher cardiovascular disease (CVD) and all-cause mortality risk compared to normal weight. Among patients with type 2 diabetes mellitus (T2DM), some studies reported significantly higher mortality risk for those with normal weight at the time of diagnosis compared to their obese counterparts – indicating the presence of the obesity paradox. However, a detailed exploration of the possible reasons for the obesity paradox in patients with T2DM has not been conducted.

The clinical-epidemiological aim of this thesis was to conduct an extensive exploration of the potential role of weight change before the diagnosis of T2DM and ethnicity in the association between BMI and CVD / mortality risk in patients with T2DM, using a large nationally representative patient-level electronic medical record (EMR) database. Given the methodological and analytical challenges in using such databases to design and conduct epidemiological outcome studies, the methodological aims were to compare and generalise (1) statistical methodological approaches for the robust extraction of a disease cohort and (2) methods for imputation of missing longitudinal risk factor data.

This thesis used the patient-level primary care EMR database from the United Kingdom –The Health Improvement Network (THIN) database. A robust methodological framework that incorporates several biostatistical methods was used to address the aims of this thesis. First, an extensive machine learning (ML) classification algorithm was used to identify and extract a cohort of patients with T2DM from the THIN database. Second, an exact matching algorithm was developed and used to match four non-diabetic controls to each patient with T2DM based on age, sex, and ethnicity. Longitudinal measurements of anthropometric, cardiovascular, and glycaemic risk factors were extracted and arranged in 6-monthly non-overlapping windows. Third, the predictive mean matching technique of multiple imputation was used to impute missing longitudinal cardiovascular and glycaemic risk factor data. These applied methodological tasks were conducted to ensure the ability to draw robust inferences on the epidemiological aims of this thesis, including the use of different study designs, inclusion, and exclusion criteria. Generalised linear model under general estimating equations setup, with unstructured covariance was used to evaluate body weight trajectories before and after diagnosis of T2DM while multivariate stratified Cox proportional hazards regression was used to assess the association of BMI at diagnosis with mortality risk in patients with T2DM.

For large EMR databases like THIN (n=~11 million patients), the use of extensive data mining / ML algorithms are required to robustly identify patients with a disease of interest. Furthermore, multiple imputation of missing longitudinal risk factor data was a valid approach as the distributions of imputed data over 24 months post diagnosis of T2DM were similar longitudinally compared to that of the unimputed data. While patients with T2DM had a significantly higher mean BMI levels and prevalence of comorbidities at diagnosis compared to non-diabetic controls, similar prevalence of cardiovascular multi-morbidity was observed among White European, African-Caribbean, and South Asian patients who were normal weight at diagnosis.

Weight trajectory analysis among patients with T2DM and no established comorbidities at diagnosis, showed that normal weight and overweight patients experienced a small but significant reduction in body weight six months before diagnosis, followed by significantly increasing trend post-diagnosis. For patients in all obese categories, consistently increasing body weight was observed six months before diagnosis followed by a decreasing trend after diagnosis. Furthermore, a paradoxical association of BMI with mortality risk was observed among patients who did not lose body weight before diagnosis – where normal weight patients had 35% significantly higher adjusted mortality risk compared with the grade 1 obese patients. However, among patients experiencing weight loss before diagnosis, BMI at diagnosis was not associated with mortality risk. The obesity paradox was further observed among White Europeans and South Asians where those with normal body weight at diagnosis were significantly more likely to die earlier by 0.6 years and by 2.5 years respectively, compared to their respective obese patients.

The findings of this thesis add to the evidence base that patients with T2DM, who were normal weight at the time of clinical diagnosis have significantly higher mortality risk compared to those who were obese, and this may partially be driven by different cardiovascular and glycaemic risk profiles of different ethnic groups. Empirical results from this thesis suggest that there was no evidence of pre-existing latent or severe disease conditions being overrepresented in normal weight patients. In fact, dynamic changes in body weight before clinical diagnosis of T2DM were independent of pre-existing latent or severe disease conditions. The increased mortality risk in the normal weight group may reflect differences in the aetiology of diabetes in normal weight people and emphasises the importance of addressing risk factors for excess mortality in this group

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, financial support and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my higher degree by research candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis and have sought permission from co-authors for any jointly authored works included in the thesis.

Publications included in this thesis

The body of this thesis contains five primary results Chapters, including five published papers, as follows:

- **Owusu Adjah ES***, Montvida O*, Agbeve J, Paul SK. Data Mining Approach to Identify Disease Cohorts from Primary Care Electronic Medical Records: A Case of Diabetes Mellitus. *The Open Bioinformatics Journal* 2017;**10**:16-27. *Joint first authorship.
- Paul SK*, **Owusu Adjah ES***, Samanta M, Patel K, Bellary S, Hanif W, Khunti K. Comparison of body mass index at diagnosis of diabetes in a multi-ethnic population: A case-control study with matched non-diabetic controls. *Diabetes, Obesity and Metabolism* 2017;**19**(7):1014-1023. *Joint first authorship
- **Owusu Adjah ES**, Bellary S, Hanif W, Patel K, Khunti K, Paul SK. Prevalence and incidence of complications at diagnosis of T2DM and during follow-up by BMI and ethnicity: a matched case-control analysis. *Cardiovascular Diabetology* 2018;**17**(1):70.
- **Owusu Adjah ES**, Samanta M, Shaw JE, Majeed A, Khunti K, Paul SK. Weight loss and mortality risk in patients with different adiposity at diagnosis of type 2 diabetes: a longitudinal cohort study. *Nutrition & diabetes* 2018;**8**(1):37.
- **Owusu Adjah ES**, Ray KK, Paul SK. Ethnicity-specific association of BMI levels at diagnosis of type 2 diabetes with cardiovascular disease and all-cause mortality risk. *Acta Diabetologica* 2018;**56**(1):87-96.

Submitted manuscripts included in this thesis

No manuscripts submitted for publication.

Other publications during candidature

Peer-reviewed publications

- **Owusu Adjah ES**, Agbemafle I. Determinants of domestic violence against women in Ghana. *BMC Public Health*. 2016; **16** (1): 1-9.
- **Owusu Adjah ES***, Montvida O*, Agbeve J, Paul SK. Data Mining Approach to Identify Disease Cohorts from Primary Care Electronic Medical Records: A Case of Diabetes Mellitus. *The Open Bioinformatics Journal* 2017;**10**:16-27. <u>*Joint first authorship</u>.
- Paul SK*, **Owusu Adjah ES***, Samanta M, Patel K, Bellary S, Hanif W, Khunti K. Comparison of body mass index at diagnosis of diabetes in a multi-ethnic population: A case-control study with matched non-diabetic controls. *Diabetes, Obesity and Metabolism* 2017;**19**(7):1014-1023. <u>*Joint first authorship.</u>
- **Owusu Adjah ES**, Bellary S, Hanif W, Patel K, Khunti K, Paul SK. Prevalence and incidence of complications at diagnosis of T2DM and during follow-up by BMI and ethnicity: a matched case-control analysis. *Cardiovascular Diabetology* 2018;**17**(1):70.
- **Owusu Adjah ES**, Samanta M, Shaw JE, Majeed A, Khunti K, Paul SK. Weight loss and mortality risk in patients with different adiposity at diagnosis of type 2 diabetes: a longitudinal cohort study. *Nutrition & diabetes* 2018;**8**(1):37.
- **Owusu Adjah ES**, Ray KK, Paul SK. Ethnicity-specific association of BMI levels at diagnosis of type 2 diabetes with cardiovascular disease and all-cause mortality risk. *Acta Diabetologica* 2018;**56**(1):87-96.

Conference abstracts

- **Owusu Adjah ES**, Paul SK. Evaluating the obesity paradox in diabetes: A longitudinal case-control study. Presented at the *QIMR Berghofer Early Career Research Seminars,* 23 March 2018, Brisbane, Queensland, Australia.
- **Owusu Adjah ES,** Montvida O, Khunti K, Paul SK. Interactive changes in cardiovascular risk factors and the long-term cardiovascular risk differ by adiposity levels in incident type 2 diabetes patients: Real World Study. Diabetologia; 2017: Springer 233 Spring St, New York, NY 10013 USA. Presented at the *European Association for the Study of Diabetes Annual Scientific Meeting,* 14 September 2017, Lisbon, Portugal.

- **Owusu Adjah ES,** Ray K, Paul SK. Association of adiposity level at diagnosis of type 2 diabetes with cardiovascular and mortality risk: Ethnicity-specific real-world study. *Presented at Melbourne Health Research Week*, 23 June 2017, Melbourne, Australia.
- **Owusu Adjah ES**, Paul SK. Evaluating the Obesity Paradox in Type 2 Diabetes. Presented at the *QIMR Berghofer Early Career Research Seminars,* 24 March 2017, Brisbane, Queensland, Australia.
- **Owusu Adjah ES**, Paul, SK. Evaluating the Obesity Paradox in Type 2 Diabetes with Real-world Data from Primary Care System. Presented at the *Australian Diabetes Society and the Australian Diabetes Educators Association, Annual Scientific Meeting*, 24-26 August 2016, Gold Coast, Queensland, Australia.
- **Owusu Adjah ES**, Samanta M, Shaw JE, Majeed A, Khunti K, Paul SK. Longitudinal changes in body weight before and after diagnosis of type 2 diabetes by BMI categories at diagnosis and the associated mortality risk. Presented at the *QIMR Berghofer Annual Student Symposium,* 15 July 2016, Brisbane, Queensland, Australia
- **Owusu Adjah ES,** Agbeve J, Klein K, Paul SK. Challenges in Relational Database for Clinical-Epidemiological Research. Oral poster presented at *QIMR Berghofer Biennial Student Retreat*, 17-18 September 2015, Gold Coast, Queensland, Australia.

Contributions by others to the thesis

- General thesis advice editing, clinical biostatistician training and advice: Prof Sarah Medland, Prof Sanjoy Paul, Dr Penelope Lind, Prof Greg Rice, Dr Mayukh Samanta
- Data access and Ethics (IRB) applications: Prof Sanjoy Paul,

Statement of parts of the thesis submitted to qualify for the award of another degree

No works submitted towards another degree have been included in this thesis

Research Involving Human or Animal Subjects

The QIMR Berghofer Medical Research Institute holds the license to extract and analyse data from The Health Improvement Network (THIN) database for research and evaluation purposes, with Professor Sanjoy Paul as the custodian and Principal Investigator for the THIN database related studies. Following the directive of National Research Ethics Service of the National Health Services of government of United Kingdom (Reference: 07/H1102/103, dated 6th March 2008), there is no requirement for any local research ethics committees to be informed about specific study(s) to be conducted using this database or for site-specific assessment to be carried out at each study site.

Any research project or program of studies using this database, resulting in publication, has to have a formal scientific protocol that needs to be evaluated in terms of research question and scientific quality, and approved by the dedicated Scientific Review Committee managed through the Licensee. The protocol for this study was approved by the Scientific Review Board (15THIN030, 17th August 2015: Appendix E).

Acknowledgements

I am grateful to Almighty God for His mercies and countless blessings in my life.

I would like to express my sincerest gratitude to all my supervisors: Professor Sarah Medland, Professor Sanjoy Paul, and Dr Penelope Lind. Thank you for your encouragement, constructive feedback, support and guidance throughout this research project. I also acknowledge general thesis guidance and support I received from my former supervisors Prof Greg Rice, Dr Mayukh Samanta.

Special thanks to goes to members of the Psychiatric Genetic group at QIMR Berghofer Medical Research Institute, I appreciate your support, contribution and encouragement during the journey to present my PhD dissertation. I am also grateful for the various support I received from the members of the former Clinical Trials and Biostatistics Unit of QIMR Berghofer Medical Research Institute. My Appreciation also goes to the staff of Melbourne Epicentre at the Royal Melbourne Hospital. Thank you for all your support and encouraging words.

I thank QIMR Berghofer Medical Research Institute for providing an excellent environment for undertaking a PhD training and acknowledge financial support through the QIMR Berghofer International PhD Scholarship and The University of Queensland International Scholarship. To my fellow PhD students and friends – thanks for the stimulating conversations and academic discourse.

Finally, I wish to express my heartfelt gratitude to my family, especially my wife, Sandra Akoley Sowah for patience, support and encouragement during difficult times and provided feedback on my PhD dissertation. To my mother Janet and siblings -thanks for the continual love and support throughout this PhD journey.

Financial support

This research was supported by QIMR Berghofer International Ph.D. Scholarship, QIMR Berghofer Ph.D. Top-up Scholarship, QIMR Berghofer Student Travel Award and The University of Queensland International Scholarship. I also acknowledge research project funding from the National Health and Medical Research Council of Australia (Project Number: GNT1063477) and European Association for the Study of Diabetes (EASD) Student Travel Grant.

Keywords

obesity paradox, type 2 diabetes mellitus, longitudinal case-control, electronic medical records, all-cause mortality, cardiovascular diseases, weight trajectory

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 111706, Epidemiology, 40%

ANZSRC code: 010402, Biostatistics, 60%

Fields of Research (FoR) Classification

FoR code: 1117, Public Health and Health Services, 30%

FoR code: 0104, Statistics, 70%

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AC	African-Caribbean
ACE	Angiotensin-converting enzyme
ACM	All-cause mortality
ADA	American Diabetes Association
ADD	Antidiabetic drug
AFR	African region
AMR	Region of Americas
ARB	Angiotensin II receptor blocker
AUC	Area under the curve
BMI	Body mass index
CA	Central Adiposity
CABG	Coronary artery bypass graft
CHD	Coronary heart disease
CI	Confidence interval
CKD	Chronic Kidney Disease
CPU	Central processing unit
CV	Cross-validated
CVD	Cardiovascular disease
CVD-M	Cardiovascular disease mortality
DM	Diabetes mellitus
DEXA	Dual-energy x-ray absorptiometry
EMR	Electronic medical record
EUR	European Region
FCS	Fully conditional specification
FBG	Fasting blood glucose
FU	Follow-up
GAD	Glutamic acid carboxylase
GLP-1	Glucagon-like peptide-1 agonist
GP	General practice
HbA_{1c}	Glycated haemoglobin
HES	Hospital episode statistics
HF	Heart failure
HR	Hazard ratio
ICD	International classification of diseases
IDF	International diabetes federation
IRB	Institutional Review Board
IRR	Incidence rate ratio
LBW	Lost body weight
MI	myocardial infarction
ML	Machine learning
MODY	Maturity-onset diabetes of the youth
MP	Multilayer Perceptron
MREC	Multi-centre research ethics committee
NHANES	National Health and Nutrition Examination Survey
NHS	National health service
NSAID	Nonsteroidal anti-inflammatory drugs
NWL	No weight loss
OAD	Oral antidiabetic drug

PAD	Peripheral artery disease
PMM	Predictive Mean Matching
PROactive	PROspective pioglitAzone Clinical Trial In macroVascular Events
PW	Percent body weight
p-yrs	Person-years
REACH	The REduction of Atherothrombosis for Continued Health
RBG	Random blood glucose
SD	Standard deviation
SBP	Systolic blood pressure
SA	South Asian
SEAR	South East Asia Region
SVM	Support vector machine
T1DM	Type 1 diabetes mellitus
T2DM	Type 2 diabetes mellitus
THIN	The Health Improvement Network
TNR	True negative rate
TPR	True positive rate
TRIAD	The Translating Research Into Action for Diabetes
UK	United Kingdom
USA	United States of America
USD	United States Dollars
WC	Waist circumference
WE	White Europeans
WHR	Waist to hip ratio
ZODIAC	The Zwolle Outpatient Diabetes project Integrating Available Care

Chapter 1: Introduction

The prevalence of both type 2 diabetes mellitus (T2DM) and obesity are continuing to rise despite efforts by the international community to halt their progress. Even though there is evidence supporting the link between obesity and an increased risk of developing T2DM, the exact mechanism still eludes the scientific community. Moreover, recent reports have also indicated that obesity offers a survival advantage to patients suffering from certain diseases like T2DM. This phenomenon is called the "obesity paradox in T2DM", where patients who were normal weight at the time of diagnosis, were found to have significantly higher mortality rates than their overweight or obese counterparts. My thesis further explores the obesity paradox in T2DM by identifying (1) longitudinal changes in body weight, in conjunction with dynamic changes in other cardiovascular and glycaemic risk factors, (2) the possible ethnicity-based differences in the risk paradigm, (3) exposure to various anti-diabetic therapies, and (4) specific causes of death in the association of obesity and mortality in patients with T2DM.

1.1 THESIS OUTLINE

There are nine chapters in this thesis. In the introductory chapter (Chapter 1), the aims and objectives are highlighted and a review of the global distribution of diabetes and the current treatments available, as well as the risk factors and management of obesity, are provided. The obesity paradox in T2DM is also discussed in the context of statistical and methodological limitations of existing research. Finally, the implications from the literature and the conceptual framework for the study are discussed. Chapter 2 describes the design adopted to achieve the aims and objectives of this thesis, while Chapter 3 discusses the different approaches used to robustly identify and extract a cohort of T2DM patients from a relational database called The Health Improvement Network (THIN) database. In Chapter 4, an exact matching method is used to match non-diabetic patients to patients with T2DM in a pre-specified ratio. Furthermore, Chapter 4 addresses the difference between complete and imputed longitudinal data on different outcomes. Chapters 5, 6, 7 and 8 contain the primary results of this thesis. These chapters cover the relationship between BMI and cardiovascular risk, death due to cardiovascular diseases in patients with incident T2DM. Finally, a discussion and conclusion that puts the findings of this study into perspective are provided in Chapter 9.

1.2 BACKGROUND

Type 2 diabetes mellitus (T2DM) is a chronic disease that is mainly associated with an increasingly sedentary lifestyle and high prevalence rates of obesity, along with other factors [1,2]. With an exponentially increasing prevalence worldwide [3], the implications for the lifetime complications of the disease are enormous because of possible long-term damage to multiple organs [4]. Among various risks factors, higher body weight (obesity) has played a central role in the history and development of T2DM. The most commonly used measure of obesity is the body mass index (BMI) due to its simplicity and reproducibility. It examines body weight relative to height and is calculated by dividing a person's weight in kilograms by the square of height in metres. The World Health Organisation (WHO) defines an obese person as having BMI greater than or equal to 30 kg/m², while overweight, normal, and underweight persons have BMI measurements in the ranges of 25-29.9 kg/m², 18.5-24.9 kg/m², and <18.5 kg/m² respectively [5]. Despite the significant health burden of both diabetes and obesity, the relationship between them remains a complicated one [6]. While obesity is a standard feature of T2DM patients, likely due to insulin sensitivity and resistance, obesity is also known to precede T2DM [7].

Obesity is considered a major risk factor for cardiovascular disease, hypertension, and diabetes. A recent meta-analysis has reinforced the fact that compared to normal weight, obesity is associated with significantly higher cardiovascular and all-cause mortality risk in the general population [8]. However, in some specific clinical populations, obese or overweight patients appear to have a better survival prognosis compared to normal weight patients in a phenomenon referred to as the “obesity paradox” [9-11]. Patients with T2DM are among clinical populations in which the paradoxical association between body weight, measured as BMI and mortality have been observed [12]. This paradox is also seen in heart failure [13], coronary heart disease [14], hypertension [15], and chronic kidney diseases [16,17]. This leads to the challenge of exploring the optimum adult body weight that best advances health, minimizes the risk of chronic disease like diabetes, and promotes longevity. This quest for optimum adult body weight has recently engaged the interest of the clinical investigators and public health professionals since weight loss is so frequently a focus of management of T2DM. Some recent observational studies have evaluated the cardiovascular and mortality risks in normal weight and overweight patients compared to obese patients with incident diabetes [12,18-20]. Patients with T2DM who were normal and overweight at the time of diagnosis had 60%, and 10% increased mortality risk respectively, compared to their counterpart obese patients. Other investigations into the obesity paradox in T2DM have produced conflicting results. Some studies did not observe lesser mortality in obese or overweight participants compared with normal weight [21-24], while others have shown that both normal weight and obese patients had significantly higher mortality outcomes than overweight patients (U-shape association) [25,26].

To the best of our knowledge, the exact mechanism of this observed obesity paradox in patients with T2DM is not fully explained. It is possible that unrecognized underlying comorbidities (pre-existing diseases) are overrepresented in the normal weight group, leading to weight loss in this group and subsequent increased risk of death (reverse causation). In addition, mortality risk may be reduced for individuals in the overweight/obese category because of treatment (e.g., more aggressive therapy, use of metformin), or adiposity in the overweight range is genuinely healthy. Also, given the interplay between ethnicity, BMI and mortality [27-30], ethnicity could play a role in explaining the obesity paradox in patients with T2DM.

While weight loss as a treatment has provided compelling evidence for diabetes control [31-33], lack of full understanding of the observation of higher mortality risk in normal weight patients with diabetes significantly limits the ability to provide appropriate weight targets for patients. Therefore, there is a need for further research to focus on potential mechanisms of this observed paradox in T2DM. I have addressed this issue by conducting a set of extensive clinical-epidemiological evaluations in terms of risk factor dynamics, and cardiovascular and mortality outcomes, using the patient-level primary care database from the United Kingdom. In this comparative longitudinal case-control study based on retrospective real-world data, the dynamics of obesity paradox would be explored by evaluating (1) body weight trajectory before diagnosis and (2) the trajectories of body weight and clinical risk factors following diagnosis of diabetes in different BMI categories, after adequately taking care of confounding factors and on-going treatment regimens.

1.2.1 Hypotheses

1. Patients who were normal weight at the time of diagnosis of T2DM may have higher mortality risk compared to overweight or obese patients.
2. Presence of underlying illness, the influence of glycaemic and other cardiovascular risk factors, and the anti-diabetic treatment may modify the association of body weight with mortality risk in patients.

1.2.2 Aims and objectives

AIM 1: To compare and generalise statistical methodological approaches for the extraction of disease events and dealing with missing data issues from national electronic medical records (EMRs) containing large patient-level longitudinal data.

Specific objectives:

a. To develop robust data mining techniques that deal with the large and complex relational database containing longitudinal patient-level information.
b. To design a comparative longitudinal study of patients with T2DM and their matched non-diabetic controls.
c. To evaluate the difference between the association of complete and imputed longitudinal data on outcomes.

AIM 2: To evaluate the association of BMI at diagnosis of T2DM with long-term cardiovascular risk and mortality.

Specific objectives:

a. To conduct a systematic review of current studies evaluating the association of BMI with cardiovascular and mortality risks in patients with T2DM.
b. To investigate the association between BMI at diagnosis and mortality risk, accounting for weight change patterns before the diagnosis of T2DM.
c. To investigate the possible roles of ethnicity, collider-stratification bias, and reverse causation in explaining the obesity paradox in patients with T2DM.

1.3 LITERATURE REVIEW

1.3.1 Diabetes Mellitus: A Brief Epidemiological Review

Diabetes mellitus (DM) is a chronic metabolic disease of different origins, with a hallmark feature of sustained high plasma glucose resulting from defects in insulin secretion, insulin action or both [4,34]. Insulin deficiency or a defect in insulin secretion is central to all the pathogenic processes involved in the development of DM. Insulin is a hormone that is secreted by the β-cells of the islets of Langerhans found in the pancreas. It is the primary regulator of carbohydrate metabolism in the whole body and is secreted or degraded in response to nutritional or hormonal states. In the fed state, high plasma glucose is detected by the β-cells of the pancreatic islets, and in turn, an essential peptide hormone insulin is synthesised and secreted from the β-cells of the pancreatic islets mainly in response to glucose. With insulin, the cells of the body are now able to take up glucose. Glucose is then taken through the glycolytic pathway to produce adenosine triphosphate which is the primary energy currency of the body [35].

When insulin is not produced, or the amount produced is not enough to enable glucose uptake into the cells from the blood, glucose is left to circulate in the blood resulting in the state of hyperglycaemia, a major feature of DM. Metabolic pathways of major macromolecules like carbohydrate, lipid, protein, and fats are affected by sustained hyperglycaemia which subsequently leads to long-term damage, failure or dysfunction of several organs [4,34]. There are many forms or manifestations of DM but the American Diabetes Association (ADA) and a report from WHO consultation, have grouped diabetes cases into two broad categories, based on clinical symptoms. The two major categories are type 1 diabetes mellitus (T1DM) and T2DM. In addition to these are gestational diabetes and "other specific types" that do not fall under either the T1DM, T2DM, or gestational diabetes, mainly due to the process that led to the particular type of diabetes and their clinical manifestation [4,34].

In T1DM, there is complete destruction of the β-cells of the pancreatic islets resulting in an absolute deficiency in insulin production and secretion. Patients with this form of diabetes are usually characterised by the presence of anti-glutamic acid decarboxylase (GAD), islet cell or insulin antibodies which identify the autoimmune processes that lead to beta-cell destruction and hence require insulin for survival. On the other hand, patients with T2DM are predominantly insulin resistant with relative insulin deficiency to a predominantly secretory defect with or without insulin resistance. Here the degree of hyperglycaemia sufficient to cause pathologic and functional changes in various target tissues, but without clinical symptoms, may be present for an extended period before diabetes is detected [4,34]. Gestational diabetes occurs when any degree of glucose intolerance appears during pregnancy, regardless of whether different treatment was used or whether the condition persists after pregnancy [2].

In addition to T1DM, T2DM, and gestational diabetes, other types of diabetes have been identified. This is where diabetes is associated with other conditions, for example, diabetes secondary to diseases of the exocrine pancreas, drugs and other endocrinopathies [36]. As such, patients may require oral agents or insulin depending on the ability of the pancreas to produce insulin. Also, monogenic defects in β cells can cause maturity-onset diabetes of the youth (MODY), a form of diabetes that is characterised by hyperglycaemia at an early age (25 years) and impaired insulin secretion with minimal or no defects in insulin action.

Globally, the prevalence of DM in all age groups is increasing exponentially. The WHO predicted that 366 million people will be living with DM by 2030 [37], but given the latest survey by the International Diabetes Federation (IDF), it is clear that the WHO may have underestimated the prevalence of DM – because by 2017 there were already 425 million people (equivalent to 1 in 11 adults) with DM [3,38], far above the predicted estimate for 2030. The IDF projects the prevalence of DM to rise to 642 million by 2040. If prevalence data is not available for a country, it is extrapolated from another country using regional data, World Bank income, ethnicity and language [38]. Given the fact that these extrapolations are less reliable, the current IDF estimates may still be underestimated [39].

The prevalence of DM is disproportionately high in some ethnic groups and socioeconomically deprived societies (e.g., South Asians, African-Caribbean). South Asians develop DM earlier and at lower BMI levels, compared to White Europeans [40,41]. In India alone, 72 million individuals (~5%) were living with DM in 2017, with a projected rise to 123.5 million by 2040 [3]. A population-based survey conducted in China in 2010 suggests that about 12% of the adult population had diabetes and about 50% of total population had pre-diabetes (impaired glucose tolerance, defined as 2-hour oral glucose tolerance levels 7.8–11.0 mmol/l, and impaired fasting glucose, defined as fasting glucose levels 6.1–6.9 mmol/l) [42].

It is estimated that more than 75% of people with DM live in low and middle-income countries, and more than 70% of them are in the working age of 20-64 years. It is known to also occur more in females than males [36] with the estimated number of females (20-79 years) living with diabetes in 2017 being 204 million [3]. Accordingly, the associated cost of managing DM and its related complications worldwide is also increasing exponentially [43] (i.e. 12% of global health expenditure ~ USD 730 billion).

T2DM accounts for 90% of diabetes cases worldwide and occurs mostly in adults but is also seen in younger patients. With the average age of onset dropping, the incidence of T2DM among adolescents has increased 15- to 20-fold since 1982 [4,34,44]. Advances in epidemiological research on T2DM has shown that the determinants of T2DM consist of many contrasting and interacting genetic, epigenetic and lifestyle factors [39]. Several reasons have been postulated for the escalating epidemic of T2DM. These include population ageing, economic development, urbanization, unhealthy eating habits and sedentary lifestyles [43]. Since adverse lifestyle changes primarily cause T2DM, populations that have undergone radical changes from traditional to western lifestyles relating to poor nutrition have very high adult prevalence. For example, Aboriginal people of Australia, North American Indians, and Pacific Islanders are known to have a high prevalence of T2DM than their surrounding communities [43,45]. High-calorie diets that lead to excess body fat, hypertension, and dyslipidaemia are considered to be a major contributor to the disease burden. People with a history of diabetes in first- and second-degree relatives have increased the risk of developing T2DM.

T1DM, which mostly occurs in children, accounts for 5-10% of DM cases. With a peak incidence during adolescence, it is estimated that over 500 million children are living with T1DM [43]. Developed countries have the highest prevalence of T1DM with Finland, Denmark, Norway, and Sweden taking the lead. Japan has the lowest incidence among developed countries. The United Kingdom (UK) alone has seen a doubling of the incidence of T1DM in persons under the age of 16 years in recent years. Moreover, because T1DM can affect persons of any age group, about 20% of patients initially diagnosed with T2DM are eventually found to have evidence of autoimmune activity typical of T1DM. This form of clinical manifestation is called latent autoimmune diabetes in adults [2].

Gestational diabetes, a pregnancy complication defined as glucose intolerance with onset or first recognition during pregnancy, significantly influences T2DM risk in exposed women and their offspring. The prevalence of gestational diabetes varies depending on the diagnostic criteria used and the study population. According to IDF 2017 estimate, about 16% of live births had some form of hyperglycaemia in pregnancy, and 1 in every 7 births was affected by gestational diabetes. Women with gestational diabetes had a seven-fold increased risk of developing T2DM compared to those without the condition [46-48]. In the children of women with gestational diabetes, exposure to intrauterine hyperglycaemia was found to be associated with an 8-fold risk of developing diabetes/prediabetes at 19-27 years of age [49].

Patients with T1DM are treated with insulin as they require it for survival, but there is an extensive range of oral antidiabetic drugs available for the treatment and management of T2DM. Based on their mode of action, they are classified as agents that: (1) stimulate insulin secretion (e.g. sulphonylureas), (2) reduce hepatic glucose production (e.g. biguanide such as metformin), (3) delay digestion and absorption of intestinal carbohydrate (e.g. α-glucosidase inhibitors), and (4) improve insulin action (e.g. thiazolidinediones). As a rule of thumb, patients with T2DM are usually treated with lifestyle interventions therapies including exercise, dietary modifications, smoking, and alcohol cessation. Oral antidiabetics are then initiated at low doses and titrated upwards according to the glycaemic response as measured by glycated haemoglobin (HbA_{1c}) levels. Insulin is introduced when lifestyle interventions and other oral antidiabetic medications fail to achieve the desired glycaemic targets [50,51].

1.3.2 Obesity

Obesity is a condition characterised by excessive fat accumulation which poses adverse risks to health [35,52]. Simply, intake of food more than the human body can use leads to accumulation and storage of excess energy as fat. This phenomenon of a positive energy balance and weight gain is compounded by (1) lack of physical activity (most cases), (2) genetics, (3) mental illness, (4) lack of sleep, and (5) endocrine disruptors [53-58]. Excess body weight is now one of the most important risk factors contributing to the overall burden of disease worldwide. In fact, the WHO describes obesity as one of the most visible, yet most neglected public health problems that threaten to overwhelm developed and underdeveloped countries [52,59].

Measurement and classification of obesity

A variety of techniques are available for the measurement and accurate approximation of body fatness and include methods like underwater weighing, dual-energy X-ray absorptiometry (DEXA), total body water, total body electrical conductivity, total body potassium, body average density management, whole body air displacement plethysmography, bioelectric impedance analysis and computed tomography. However, these methods have limitations such as cost and complexity of use. For example, DEXA has the ability to distinguish between bone minerals from non-bone fat-free and fat soft tissue and measure the whole body as well as individual segments [60]. However, use of DEXA as a gold standard remains to be fully evaluated as its main limitation is that DEXA scanners have a weight and scanning bed area limit of about 136 kg and 60 cm respectively, making it impossible to measure obese patients. Despite these challenges, an approach of estimating total body composition from half body scans have been explored and validated in two separate studies. The results from both studies showed that half-body scans can accurately predict whole-body per cent [60]. Anthropometry-based tools for the clinical evaluation of body fatness has proven beyond doubt to be most useful because of their simplicity and low cost of operation [61] [62,63]. Examples of anthropometry-based tools used in assessing body fatness are (1) body fat percentage (2) body mass index (BMI), (3) waist circumference (WC), (4) hip circumference and (5) waist-to-hip ratio (WHR).

Body Fat percentage

Measurement of body composition is increasingly becoming important in clinical practice with the gold standard for body composition analysis being cadaver analysis [64] in which the cadaver is divided into parts and analysed completely. However, a close to accurate measure of body fat (essential fat and storage fat) is body fat percentage which is calculated as total body fat mass divided by the total mass, multiplied by 100.

Body Mass Index (BMI)

BMI is calculated mathematically as weight in kilograms divided by the square of height in meters [5]. As one of the clinically accepted methods of assessing total body fat, the WHO defines an obese person as a patient whose measurement is greater than or equal to 30 kg/m². It follows that overweight, normal and underweight have BMI measurements in the ranges of 25-29.9 kg/m²,18.5-24.9 kg/m², and <18.5 respectively. Also, grade 1, grade 2, and grade 3 obesity are defined by BMI in the ranges 30-34.9 kg/m^2, 35-39.9 kg/m^2, and ≥ 40 kg/m^2 respectively [5]. Though BMI is a crude measure of body fat [65,66], it has been recommended as an essential component of the initial clinical assessment of obesity due to its simplicity and reliability. Its main limitation is that BMI cannot distinguish between lean and fat mass so when age and sex-related differences in body composition come into play, it cannot correlate well with body fat in some age, sex, and ethnic groups [5].

Waist Circumference (WC)

Though BMI may prove to be simple and easy to use method of assessing obesity in the clinical setting, there are other measurements of adiposity that can also be used to define obesity in a more accurate and specific manner than BMI. In assessing central obesity, WC has been shown to be more effective and associated with a high risk of CVD and mortality [63]. It is easily measured while the patient is standing or in expiring position, but it has not been fully accepted in clinical practice. This is due to the variability in measurement of waist circumference mainly due to the subjective nature of measuring site. For instance, evidence from literature has revealed 8 different sites for WC: (1)

halfway between the lowest rib and the iliac crest (midpoint), (2) point of minimal circumference, (3) immediately above the iliac crest, (4) umbilicus, (5) 1 inch above the umbilicus, (6) 1 cm above the umbilicus, (7) at the lowest rib, and (8) point of largest circumference around the waist [62,67]. As a result of this variability in measurement locations, different health authorities have recommended the use of different locations for the estimation of WC. While the WHO recommends the use of midpoint WC measurement, the National Institutes of Health and National Heart, Lung, and Blood Institute and the American Heart Association recommend the iliac crest. The iliac crest has the advantage of being easily found even in patients with lots of body fat because it is a bone. With good training and a measuring tape, using the iliac crest as the measurement location may yet prove to be a standardised method for WC measurement [62].

Waist-to-Hip Ratio (WHR)

WHR is defined as waist circumference divided by hip circumference with the hip circumference measured as the largest circumference around the buttocks. Both WC and WHR have been shown to independently predict the incidence of CVD, T2DM and mortality [68]. In a meta-regression analysis, de Koning and others [68] pooled data from over 250,000 participants followed over 6 years with more than 4000 cardiovascular endpoints and found significantly increased risk cardiovascular events (defined as myocardial infarction, ischemic heart disease, and coronary artery disease). In particular, a 1 cm increase in WC was associated with a 2% increase whereas a 0.01cm increase in WHR was associated with 5% increased risk of a cardiovascular event after adjusting for confounders.

Brief epidemiology of obesity

As a modifiable risk factor for death due to cardiovascular diseases and all-cause mortality worldwide, obesity affects people of all ages. Despite efforts by the international community to halt obesity rates to those of 2010, the global burden of obesity is still enormous. Current epidemiological indices of obesity have doubled since 1980, with 13% of the world's population being obese as of 2014 [69]. Estimates of regional trends indicate that the Americas, Europe, and Eastern Mediterranean regions have the highest burden of obesity whereas South East Asia has the lowest burden (Figure 1.1) [69]. However, considering that Asians are at higher risk of obesity-related complications at relatively lower BMI, a small increase in prevalence translates into millions of cases of chronic diseases [70]. Furthermore, the developing world is now set to endure a double burden of disease and will soon equal or overtake the developed world regarding obesity prevalence mainly due to the abandonment of traditional lifestyles in exchange for western lifestyles [71].

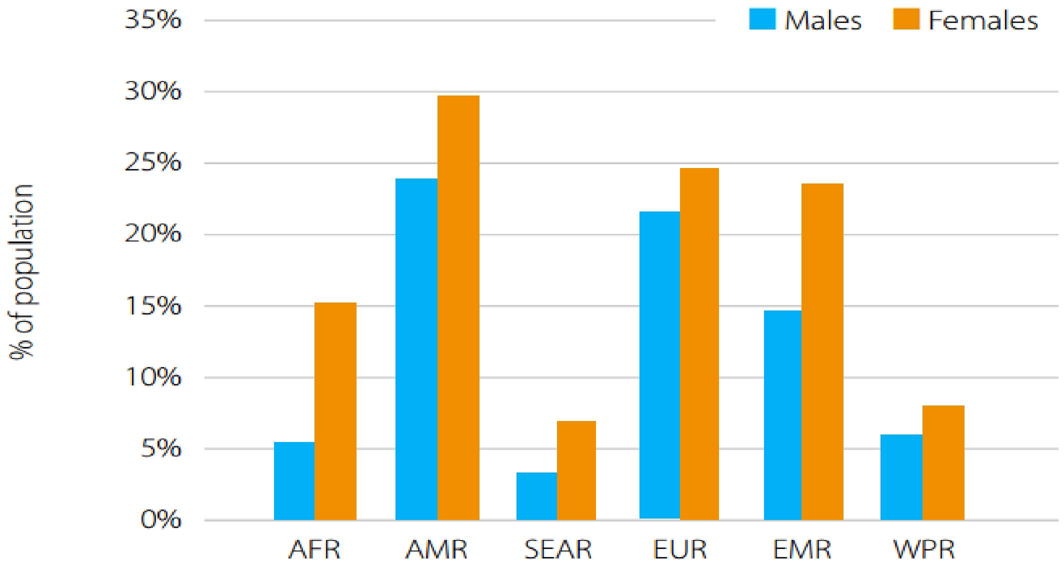


Figure 1.1: Age-standardised prevalence of obesity in adults aged 18 years and over (BMI ≥ 30 kg/m^2), by WHO regions

(*Picture adapted from Global status report on non-communicable diseases, 2014; AFR=African Region, AMR=Region of the Americas, SEAR =South-East Asia Region, EUR=European Region, EMR=Eastern Mediterranean Region, WPR=Western Pacific Region*).

While the prevalence of obesity in Western Pacific countries is low, the actual rate of obesity in this region far outweighs that of the USA. Nonetheless, the USA alone has about one-fifth of the worldwide cases and the highest obesity rates, and none of its 50 states has a prevalence of less than 20%. The associated medical cost of obesity was $147 billion in 2008 alone and with increasing trends in obesity; this cost is projected to increase [69]. Currently, the mean BMI for men and women in the UK are 27.6 kg/m^2 and 27.1 kg/m^2 , respectively and the prevalence of overweight and obesity among adults in the UK is at a staggering 27% [72].

1.3.3 The obesity paradox

Obesity and overweight have been implicated as risk factors in most disease conditions including heart-related diseases [10]. However, the effect of weight, most often measured as BMI, on mortality in certain disease states remains unclear, as several studies have yielded contrasting results ranging from a direct association, no association, to U or J shaped associations [20-26]. So far, findings have indicated associations suggesting high mortality in individuals with BMI in the overweight and obese ranges. In the general population, a recent meta-analysis, based on a sample of more than 2.88 million individuals with more than 270,000 deaths, has reinforced the fact that compared to normal weight, obesity was associated with significantly higher all-cause mortality [8]. The risk of death ranged from 18% in the overall obesity category to 29% for grade 2 and 3 obesity, although grade 1 obesity was not associated with higher mortality. In fact, being overweight was protective and associated with significantly lower all-cause mortality by 6% compared with normal weight individuals.

Consequently, mounting evidence in recent studies points to an inverse relationship rather than a direct relationship, suggesting that overweight and obesity confers a survival advantage [9-11]. This inverse relationship is a phenomenon that has been observed in many clinical conditions and has been termed the "obesity paradox". The obesity paradox has been observed in patients with T2DM; however, results have been controversial. While some studies found evidence in support of the obesity paradox [7,12,19,25,26,73-82] others did not [7,21-24,83-91]. A summary table of 31 studies investigating the phenomenon of the obesity paradox in patients with T2DM is presented in Table 1.1.

Obesity paradox in T2DM

Twenty-four studies found evidence supporting the obesity paradox in patients with T2DM [7,12,19,25,26,73-82,92-100] (Table 1.1). Obese and overweight participants in the Translating Research Into Action for Diabetes (TRIAD) [80] and the PROactive trial [93] experienced lower mortality compared to persons who were either normal weight or those who lost weight in the course of the trial. Although the findings from these studies supported the obesity paradox, there was no data on the duration of diabetes of the participants. Furthermore, participants in the PROactive study also had cardiac diseases before enrolling in the study. These factors limited the findings of the study as cardiovascular diseases and variations in diabetes duration could independently influence mortality in patients with diabetes.

Even when normal weight and underweight categories were collapsed into one (BMI<24.9 kg/m²), Dallongville and colleagues [76] reported better outcomes with increasing obesity among diabetic patients who had established atherosclerotic arterial disease in the REACH registry [76]. This study was limited by the shorter follow-up period of 2 years and the fact the patients were already undergoing lipid-lowering therapy, thus a true causal inference cannot be drawn. Secondly, it is impossible to evaluate if the adverse effect found in the lower BMI categories as defined in this study was due to underweight associated complications like death due to malnutrition and anaemia.

Table 1.1: Summary of studies reporting evidence in favour of or against the obesity paradox in patients with T2DM

	Author	**Exposure groups**	**Adiposity measure**	**FU (yrs.)**	**Outcome**	**Risk model**	**Finding**
1	Sasaki, [81]	T2DM only (n=1,939)	% BW	9.4^{α}	ACM	Survival, Logistic regression	Obesity paradox
2	Balkau, [74]	T2DM (n=1,005), Non-diabetic (n=6,161)	BMI, HTR	15.6^{α}	ACM	Cox proportional regression	Obesity paradox
3	Chaturvedi, [75]	T2DM (n=2,690)	BMI	13	ACM	Cox proportional regression	Obesity paradox
4	Zoppini, [7]	T2DM (n=3,398)	BMI	$10^{\ddagger}$	ACM	Cox proportional regression	Obesity paradox
5	McEwan, [80]	T2DM (n=8,733)	BMI	3.7^{α}	ACM	Cox proportional regression	Obesity paradox
6	Khalangot, [78]	T2DM (n=89,443)	BMI	2.7^{α}	ACM	Cox proportional regression	Obesity paradox
7	Weiss, [82]	T2DM (n=122)	BMI	3.7^{α}	ACM	Cox proportional regression	Obesity paradox
8	Carnethon, [19]	T2DM (n=2,625)	BMI	27,125 p-yrs.	ACM	Cox proportional regression	Obesity paradox
9	Dallongeville, [76]	T2DM (n= 19,579)	BMI, WC	2^{α}	ACM, CVD-M	Cox proportional regression	Obesity paradox
10	Doehner, [93]	T2DM (n=5202)	BMI	2.9	ACM, CVD-M	Cox proportional regression	Obesity paradox
11	Kokkinos , [79]	T2DM (n=4,156)	BMI	$7.5^{\#}$	ACM	Cox proportional regression	Obesity paradox
12	Ma [95]	Diabetic (n=1,712), IGT (n=2,545), Non-diabetic (n=11,791)	BMI	$9.4^{\ddagger}$	CVD-M, CHD	Cox proportional regression	Obesity paradox
13	Logue,[25]	T2DM (n= 106,640)	BMI	4.7^{α}	ACM	Cox proportional regression	Obesity paradox
14	Jackson, [77]	Diabetic (n=4,740), Non-diabetic (n=69,970)	BMI	$9^{\ddagger}$	ACM	Cox proportional regression	Obesity paradox
15	Perotto [96]	T2DM (n=1475)	BMI, WHR	$10.2^{\ddagger}$	ACM, CVD-M	Cox proportional regression	Obesity paradox
16	Thomas, [12]	T2DM (n= 47,509)	BMI	$5^{\#}$	ACM	Cox proportional regression	Obesity paradox
17	Lajous, [92]	Free of diabetes	BMI	16.7^{α}	ACM	Cox proportional regression	Obesity paradox
18	Tseng, [94]	T2DM (89,056)	BMI	12	ACM	Cox proportional regression	Obesity paradox
19	Murphy [97]	T2DM (n=637)	BMI	$6.1^{\ddagger}$	ACM	Cox proportional regression	Obesity paradox
20	Zhao, [26]	T2DM (n= 34,832)	BMI	8.7^{α}	ACM	Cox proportional regression	Obesity paradox
21	Badrick [73]	T2DM (n=10,464), Non-diabetic (n=31,020)	BMI	$8.7^{\#}$	ACM	Cox proportional regression	Obesity paradox
22	Lee [98]	Diabetic (n=546,232), IFG(n=2,505,235), Non-diabetic (n=9,403,894)	BMI	10.5^{α}	ACM	Cox proportional regression	Obesity paradox
23	Xu [99]	T2DM (n=52,488)	BMI	6^{α}	ACM	Cox proportional regression	Obesity paradox
24	Jenkins[100]	T2DM (n=23,842)	BMI, WC, WHR		ACM	Cox proportional regression	Obesity paradox
25	Pettitt [88]	Diabetic (n=499), Non-diabetic (n=1,968)	BMI	23,608 p-yrs.	ACM	Survival model	No paradox
26	Rosengren,[89]	Diabetic (n=232) *, Non-diabetic (n=6,665)	BMI	7.1^{α}	CHD, ACM	Logistic regression	No paradox
27	Ford, [23]	Diabetic (n=602) *, Non-diabetic (12,562)	BMI	10^{α}	CHD, ACM	Cox proportional regression	No paradox
28	Ross, [90]	T2DM (n=373)	BMI	$14^{\ddagger}$	ACM	Cox proportional regression	No paradox
29	Cho, [83]	T2DM (n=5,897)	BMI	57,909 p-yrs.	Fatal CHD	Cox proportional regression	No paradox
30	Church, [21]	Diabetic (n=2,196) *	BMI	32,161 p-yrs.	ACM	Cox proportional regression	No paradox
31	Mulnier, [87]	T2DM (n=28,725), Non-diabetic (n=15,505)	BMI	$7^{\ddagger}$	ACM	Cox proportional regression	No paradox
32	McAuley, [86]	T2DM (n=831)	BMI	4.8^{α}	ACM	Cox proportional regression	No paradox
33	Sluik, [91]	Diabetic (n=5,434)	BMI, CA	$9.3^{\#}$	ACM	Cox proportional regression	No paradox
34	Tobias, [24]	T2DM (n= 11,427)	BMI	15.8^{α}	ACM	Cox proportional regression	No paradox
35	Bozorgmanesh [101]	Diabetic (n=1322)	BMI, WC, WHR	$9.1^{\ddagger}$	ACM	Parametric survival model	No paradox

36	Costanzo, [22]	T2DM (n= 10,568)	BMI	10.6 $^{\#}$	ACM	Cox proportional regression	No paradox
37	Kuo, [85]	T2DM (n=2,161)	BMI	5.6 $^{\alpha}$	ACM	Cox proportional regression	No paradox
38	Edqvist [84]	T2DM (n=149,345), Non-diabetic (n=734,097)	BMI	5.5 $^{\#}$	ACM	Cox proportional regression	No paradox

$^{\alpha}$: mean; $^{\#}$: median; ‡: maximum;
p-yrs.: person-years;
*authors included patients with diabetes (no distinction between types provided);
FU: Follow up duration;
ACM: All-cause mortality;
CHD: coronary heart disease;
CVD-M: CVD mortality;
CA: Central adiposity;
%BW: Percent body weight;
WC: Waist circumference;
WHR: Waist-to-hip ratio;
IGT: Impaired glucose tolerance;
IFG: Impaired fasting glucose;

In evaluating the possible association between BMI and mortality among patients with diabetes, sex and ethnicity were important modifying factors, as in some instances, mortality risk was more pronounced in men than women or in one ethnic group compared to others. Using the diabetes duration of within one year from diagnosis, Logue and colleagues [25], evaluated the association of BMI with the risk of cause-specific mortality in Scottish cohort of 106,640 participants. They showed a U-shaped association of BMI with mortality where both normal weight and obese patients had significantly higher mortality outcomes compared to overweight patients. Notably, mortality risk was 22% and 32% higher in normal weight (20-25kg/m²) men and women respectively compared to their overweight counterparts. Although the study was based on a large sample size with over 9,000 deaths, the authors were limited by their inability to evaluate the differential risk among patients with and without the history of cardiovascular disease. Nevertheless, evidence supporting the obesity paradox in T2DM have also been reported in studies that used male subjects only where an inverse association of quintiles of BMI with mortality have been reported [7,79,82,91].

Furthermore, a similar U-shaped association of BMI with all-cause mortality was observed by Zhao and colleagues [26], in their prospective cohort study of 19,478 African-Caribbean and 15,354 White European patients with T2DM. The authors found a significantly increased risk of all-cause mortality among African-Caribbeans with BMI <30 kg/m^2 and ≥ 35 kg/m^2 and among White Europeans with BMI < 25 kg/m^2 and ≥ 40 kg/m^2 compared with patients with BMI of 30 to 34.9 kg/m^2. In a Taiwanese population of diabetes patients, higher mortality from all causes, cancer, and DM complications were reported when patients with BMI < 18.5 kg/m² were compared to patients with BMI in the range of 18.5-22.9 kg/m². Given that BMI classifications in the purely Asian population are slightly different from the WHO accepted classification, Tseng and colleagues [94] may have actually found an obesity paradox in an Asian (Chinese) population using a relatively large sample of about 89,000 diabetes patients. These studies highlight the importance of evaluating potential differences in risk at different BMI levels in subgroups of patients defined by ethnicity or sex.

In a study to investigate the relationship between weight status and mortality, Carnethon and colleagues [19] used adults with new-onset T2DM to mitigate the possibility of participants developing any diabetes-related complication which is likely to influence the weight status and mortality and subsequently influencing the study findings. The researchers showed that mortality was higher in adults with normal weight at the time of incident diabetes than obese or overweight cohorts [19]. Furthermore, another group of researchers independently found that adults with normal weight at diagnosis of T2DM have significantly higher mortality risk compared to those who were obese [12].

Using data on 47,509 patients from the UK general practice (GP) database with onset T2DM, Thomas and colleagues [12] showed that among incident T2DM patients without prior CVD, overall mortality risk was 47% higher for normal weight patients compared their counterpart, obese patients. The findings by Carnethon and colleagues[19] and Thomas and colleagues [12] have abridged the limitations of potential baseline disparities in diabetes durations as was the case in the TRIAD study as well as the findings from Logue and colleagues [25].

No obesity paradox in T2DM

Fourteen studies found no evidence to support the obesity paradox in patients with T2DM [7,21-24,83-91,101] (Table 1.1). Ford and colleagues [23] used data on 602 diabetic participants from NHANES epidemiologic follow-up study and reported no association of obesity with overall mortality but a direct association with coronary heart disease (CHD) mortality. Using data on men with diabetes from the Aerobic Center Longitudinal Study and the Veteran Exercise Test Study respectively, Church and colleagues [21] and McAuley and colleagues [86], reported no difference in total mortality with increasing BMI in men with diabetes. They found that reduced exercise capacity or cardiorespiratory fitness was to blame for adverse effects of BMI on mortality in persons with DM. Furthermore, even when a limited set of clinical factors (smoking, diabetes duration, sex, insulin use and metformin use) were adjusted for, Landman and colleagues [20] found no association between BMI and cause-specific mortality like cancer mortality among diabetes outpatients enrolled in the ZODIAC trial in the Netherlands. The authors, having previously observed an inverse relationship between weight status and cancer mortality, now argued that this trend disappeared because of the relatively large sample size and longer follow-up duration of about ten years. More so, Tobias and colleagues, [24] did not observe any lesser mortality in obese or overweight participants than others with normal weight. Nonetheless, the cohorts in this study were free of any cardiovascular diseases or cancer at the onset of diabetes diagnosis as was observed in the PROactive trial [93]. Therefore, any possible bias that could have been introduced by these conditions was reduced.

In a more recent study, Costanzo and colleagues [22] observed an inverse association in mortality and overweight patients with T2DM but not in their obese counterparts, with median diabetes duration ranging from 1 to 3 three years over the BMI categories at baseline. It appears that weight / BMI measured at the first visit was used as a baseline data and for all the risk analyses. Without the knowledge of anti-diabetic and weight-modifying medications, it would be difficult to ascertain the possible changes in body weight from diagnosis of diabetes to the study baseline. Also, without information on the longitudinal changes in weight / BMI post baseline, this essentially sets the baseline BMI as a random measure, with the potential to bias the inference. While the obesity paradox was reported in the context of BMI measured at the time of diagnosis of T2DM, Paul and colleagues [102] further reported the longitudinal changes in the BMI compared between those who died and who remained alive. This clearly suggested that the longitudinal measures of BMI for those who died were consistently lower on average by 2.4 kg/m^2 ($p<0.01$) during two years of follow-up, compared to those who remained alive. A 3 kg/m^2 higher BMI during follow-up was also associated with 3-14% reduced likelihood of mortality ($p=0.025$), adjusting for longitudinal measures of blood pressure, lipids, concomitant anti-diabetic and cardio-protective medications, and the competing risk of cardiovascular and renal events. The protective effect of higher BMI trajectory on mortality risk was evident irrespective of co-morbidity status, including cardiovascular and renal disease [103].

1.3.4 Methodological limitations of existing research

The limitations of existing studies can be broadly grouped into design and analytical issues. The design issues include (1) the use of any available measure of BMI as the baseline measure, (2) non-inclusion of pre-diabetes weight change scenario before the clinical diagnosis of diabetes, and (3) no consideration on the history or prevalence of cardiovascular and other diseases at the time of baseline assessment. The analytical issues include (1) inconsistent BMI classifications, (2) missing BMI and risk factor data, (3) non-proportional risk for the five BMI categories, and (4) the potential for confounding by medication use. The detailed explanation of the design and analytical limitations of existing studies are discussed below.

Measurement of BMI at diagnosis of T2DM

One of the primary weaknesses of most of the existing studies is that the time of clinical diagnosis of diabetes (or close time-window around it) was not considered as the baseline or index date for the follow-up risk evaluation. The time at which BMI was measured has the potential to impact on the estimates obtained from risk assessment models and the resulting clinical inferences. Several time points at which BMI was measured have been used in the literature and these include measurement at adulthood, menarche, diagnosis, and entry into study or registry. By the definition of the obesity paradox in T2DM, the evaluation of the association of BMI with mortality or cardiovascular risk should use BMI measured at diagnosis as the baseline measure. However, only 5 of the 24 studies that reported an obesity paradox used a BMI measure obtained around the time of diagnosis of diabetes [12,19,25,26,73]. Of these, only studies by Carnethon and colleagues [19], Thomas and colleagues [12], and Zhao and colleagues [26] obtained BMI status at diagnosis (exact), within 3 months, and within 6 months of diagnosis respectively as their baseline BMI measurement. Similarly, only 2 of the studies that reported no evidence for the obesity paradox in T2DM used BMI measured around diagnosis as the baseline measure in their mortality risk assessment [24,87]. These studies by Tobias and colleagues [24] and Mulnier and colleagues [87] used BMI obtained within 11 months before diagnosis and within 3 years of diagnosis respectively. The other 31 studies either used a random measure of BMI or used BMI obtained at the entry into the study. In the context of the obesity paradox in T2DM, the use of BMI measured at random, study entry, and more than 6 months before and after diagnosis may be misleading as these have different clinical impacts. Studies investigating the obesity paradox in T2DM should include in their design, BMI measured as close as possible to the time of diagnosis (i.e. ± 3 months).

Inconsistent use of BMI Classifications

While the WHO has provided a uniform classification for BMI categories [104], the BMI categories reported in the 31 studies included in this review have been inconsistent. Quintiles of BMI [7,76], modified cut off points (e.g. 20-25 kg/m^2 for normal weight instead of 18.5-24.9 kg/m^2) [87] and addition of extra BMI cut off points (e.g. 18.5-22.9 kg/m^2, 23-24.9 kg/m^2) [26] are among the different classifications used. This limits the ability to conduct a thorough comparison of effect estimates across studies. Also, some studies included underweight patients or lump together underweight and normal weight patients. Being underweight is associated with complications like death due to malnutrition and anaemia, so in the context of evaluating the obesity paradox, studies that included patients in the underweight category may not be able to provide robust inferences [76]. Even in the meta-analysis by Kwon and colleagues [105], an approach that fixed the reference BMI to 18.5 kg/m^2 was used. In this study, most of the risk was shown to be low for BMI < 40 kg/m^2 [all hazard ratios (HR) ≤ 1 compared to 18.5 kg/m^2] and this is a misleading synthesis of studies on the obesity paradox in T2DM.

Missing BMI and longitudinal risk factor data

It is clear from Table 1.1 that the most common study design employed was a prospective cohort study, where exposure was defined as diabetes status (i.e., diabetic vs. non-diabetic). Both prospective and retrospective study designs are prone to missing or incomplete information which can impact the generalisability of study findings. Despite the availability of standard statistical techniques for addressing the problem of missing covariate data in longitudinal studies, only a few studies have provided sufficient information on the distribution of missing longitudinal data on body weight or BMI and the methods used for imputing missing data [106]. A review of the 38 studies that have evaluated the obesity paradox in T2DM reveal that only 17 of these studies mentioned missing values either on BMI or on some other covariate [21,23-26,73,74,76,80,84,86,87,91,98,99,107,108]. Of this, only 2 reported the imputation method used [80,84]. The other 15 of the 17 studies that reported missing adiposity measures simply excluded patients based on missing values [21,23-26,73,74,76,86,87,91,98,99,107,108] (Table 1.1).

Studies investigating the obesity paradox in T2DM require a measure of adiposity in the form of BMI, waist circumference, or weight. Therefore, it is not out of place to exclude patients with missing adiposity measures. While such exclusions did not result in smaller sample sizes in these studies, it would be prudent to apply the exclusion criteria after imputing for missing data. Furthermore, in studies that will incorporate longitudinally collected covariate data, multiple imputation of missing data could prove beneficial as single imputation does not reflect the uncertainty about the prediction of unknown missing values and the resulting estimated variance of the parameter estimates obtained will be biased towards zero [109].

Pre-existing disease conditions

Reverse causality has been proposed as one of the possible reasons for the obesity paradox in chronic diseases [110]. The term itself was traditionally used to denote the event where an outcome precedes and causes the exposure [111]. However, the applicability of this concept to studies investigating the relationship between BMI and mortality has been questionable. In particular, studies that investigated the obesity paradox have been meticulous in defining the outcome of the study as all-cause mortality or cardiovascular mortality (Table 1.1). In such settings, reverse causality by its traditional definition—where death would precede weight/BMI status is not applicable. Nonetheless, several modified and inconsistent definitions of reverse causality with regards to the relationship between BMI and mortality has been proposed [111], and one of such definitions is that some chronic diseases or latent diseases lead to weight loss before diagnosis of a disease and such weight loss before diagnosis would have an impact on the association between BMI and mortality [111,112].

Technically, this can be considered as confounding by pre-existing disease and some studies investigating the obesity paradox in patients with T2DM have taken steps like exclusion of patients with prevalent disease at baseline [12], limiting analysis to patients with more than one year of follow-up [25], and adjusting for prevalent diseases [86]. While it was demonstrated by Thomas and colleagues [12] that the obesity paradox in T2DM exists irrespective of existing cardiovascular diseases before or after diagnosis of T2DM, no other study has investigated the impact of pre-existing diseases on weight levels before diagnosis. Addressing such a question would further our understanding of possible mechanisms of the obesity paradox in patients with T2DM.

Non-proportional risk

The obesity paradox can be holistically evaluated only with observational data, as randomised control trials would be long, expensive, unethical or impractical to randomise a patient to a BMI category. However, given the observational nature of all the studies included in this review (Table 1.1), significant differences in risk factors between groups being compared could impact any inferences obtained. Most of the risk analyses reported were based on multivariate Cox regression model, the validity of which depends on the proportional hazards assumption. This is unlikely to be true for patients with incident T2DM under different adiposity levels. To account for the inherent differences in risk factors between the defined BMI categories and the fact that risk may not be proportional, survival time treatments effects modelling approach can be used to provide robust inferences [113-116]. This modelling approach uses the potential outcomes in a counterfactual framework to allow comparison of survival time for CVD and all-cause mortality for patients with different BMI categories. Primarily this novel methodological approach allows us to balance the categories of comparisons on the basis of global risk paradigm within the cohort, generally using the weighted propensity-score type adjustments. To the best of our knowledge, no study investigating the obesity paradox in T2DM has adopted robust approaches to proving inferences.

The potential for confounding by medication use

There is an extensive range of oral anti-diabetic drugs available for the treatment and management of T2DM. Despite helping maintain good glycaemic control, some anti-diabetic treatments are known to cause an increase (sulphonylureas, insulin, thiazolidinedione), or decrease (GLP-1 receptor agonists and SGLT-2 inhibitors) in body weight, while others have neutral effects (metformin, α-glucosidase inhibitors, DPP-4 inhibitors) on body weight [117,118]. Thus, the weight loss or gain after diagnosis and exposure to anti-diabetic therapy may have different effects on mortality and cardiovascular risk. This necessitates the need to explore the possible association of weight changes with cardiovascular or mortality outcomes in patients treated with different anti-diabetic medications. To date, none of the 31 studies that have investigated the obesity paradox in T2DM has conducted a dedicated analysis of the effect of drug classes on the association of BMI at diagnosis with mortality or cardiovascular risk.

1.3.5 Significance of the study

The question of optimal BMI and the effects of being underweight or overweight on the risks of cardiovascular diseases and mortality remain controversial [8,119-121]. This has led to the challenge of exploring the optimum adult body weight that best advances health, minimizes the risk of chronic disease like diabetes, and promotes longevity since weight loss is so frequently a focus of management of T2DM.

The obesity paradox significantly challenges the centrality of weight reduction in diabetes management. The consequences of answering this question have profound health and socio-economic implications for individuals and the population. Therefore, a holistic evaluation of the complex relationship between weight change and long-term risk at the population level, based on a robust methodological framework, is required to evaluate the phenomenon of the obesity paradox in patients with T2DM. With the strength of extensive long-term longitudinal data on both diabetic and non-diabetic patients, this thesis is ideally poised to address this critical issue of clinical and public health importance.

Firstly, this thesis will use BMI classification cut points defined by the WHO in keeping with the majority of the previous studies. Second, to ensure maximum use of available information of longitudinal weight, BMI, blood pressure and lipids, multiple imputations of these risk factors will be conducted before any inclusion and exclusion criteria are applied. Third, to disentangle the contribution of pre-existing diseases to weight loss before the diagnosis of T2DM, I argue that it is necessary to exclude patients with prevalent diseases before the diagnosis of T2DM and then conduct an analysis of weight trajectory before diagnosis. Also, for further clarification, an evaluation of the impact of weight change (gain or loss) before the diagnosis of T2DM on the association between BMI at diagnosis and mortality risk is necessary. To date, no study investigating the obesity paradox in T2DM has performed such an analysis. Finally, it is necessary to evaluate pre-existing conditions and outcomes of interest in diabetes and non-diabetes populations. Therefore, in this thesis, a retrospective longitudinal study design is employed in which data is obtained on cohorts of patients with T2DM and their matched non-diabetic controls. This thesis will address the emerging challenge regarding diabetes and weight status, as the findings could directly inform timely prevention and management practices to reduce adverse outcomes in all patients with T2DM, especially in those with normal body weight, who may have a false sense of protection because they are not overweight or obese.

Chapter 2: Research Design

2.1 DATA DESCRIPTION

2.1.1 Data Source

The data used for this thesis was extracted from The Health Improvement Network (THIN) database, a nationally representative individual patient-level primary care database from the UK. In the UK Department of Health system, patients are registered with a GP, while secondary care treatment can be provided elsewhere. Under terms specified by the UK's National Health Service (NHS), GPs contribute data to THIN, and the database is updated continuously within the centralised data capture system. Although data collection for THIN's robust scheme started in 2002, primary care EMRs in THIN are available for some patients since 1987 [122]. The database is linked to other sources of hospital and national statistics data and is demographically representative of the UK population in terms of age, sex and patients with T2DM [123-125]. Currently, the THIN database contains comprehensive longitudinal data on more than 17 million patients from over 700 GP centres from across England, Wales, Scotland and Northern Ireland. For this thesis longitudinal data on about 13 million patients from 1990 till September 2014 was used, of which 85% were identified to have records that are considered valid and acceptable for research (decided by THIN). The majority of these patients were registered with practices within England (79%) and Scotland (12%) (Table 2.1). The UK primary care databases (THIN and Clinical Practice Research Database) are considered the most exhaustive collection of all possible demographic, clinical, laboratory, medications and event history data worldwide. A schematic representation of the THIN database is presented in Figure 2.1.

Table 2.1: Distribution of practices within the version THIN database used for this thesis.

Country	Health authority	Number of practices	Number of patients (%)
England	East Midlands	19	356,702 (3)
	East of England	40	815,454 (7)
	London	73	1,500,498 (14)
	North East	15	247,254 (2)
	North West	69	1,000,072 (9)
	South Central	57	1,442,637 (13)
	South East Coast	47	1,028,292 (9)
	South West	57	1,018,951 (9)
	West Midlands	49	940,088 (9)
	Yorkshire & Humber	20	342,094 (3)
Wales	Wales	49	768,162 (7)
Scotland	Scotland	89	1,289,931 (12)
Northern Ireland	Northern Ireland	27	267,890 (2)
Total		611	11,018,025 (100)

2.1.2 Demographic and anthropometric data

Demographic data in the THIN database includes age at registration with the practice, date of birth, gender, ethnicity, smoking, alcohol use, death dates for those who have died, and the transfer out dates for patients who have moved away. The ethnicity data in the UK primary care databases are limited, with only 35% of patients in our database having their ethnicity defined. Self-reported ethnicity was used to classify patients as White European, African-Caribbean, South Asian, other Asian, Middle Eastern, Mixed, and other. This primary care database provides a validated score on the socio-economic status of individuals, by estimating a socioeconomic "deprivation score" using four parameters: (1) unemployment (as a percentage of those aged 16 and over who are economically active); (2) non-car ownership (as a percentage of all households), (3) non-home ownership (as a percentage of all households), and (4) household overcrowding. Finally, a score from one to five is assigned with score 1 representing the most affluent and score 5 representing the least affluent [126].

The anthropometric data includes longitudinal data on body weight, height and BMI. The last known measures of these variables are made available at the time point of the updated database release. All longitudinal data contains the date of measurement. Behavioural data includes longitudinal information on smoking and drinking status. The most updated information on these variables is provided at the time of updated data release. However, this information is not necessarily collected at all GP visits for an individual. The basic demographic characteristics of patients in the THIN database are presented in Table 2.2. The median follow-up for patients in the THIN database was 13 years [median (Q1, Q3): 13 (6, 22)] and there were more females (52%) than males. Patients who self-identified as White European and mixed ethnicity formed 19% and 10% of the entire population respectively in the THIN database. The proportion of South Asian and African-Caribbean patients were 5% and 3% respectively (Table 2.2).

Table 2.2: The distribution of basic demographic characteristics of patients in the THIN database

Patients in THIN	**11,018,025**
Age at last date of collection $^{\alpha}$	45 (25)
Age at last date of collection $^{\#}$	
<=20	1,876,724 (17)
21-30	1,478,108 (13)
31-40	1,792,669 (16)
41-50	1,779,530 (16)
51-60	1,246,761 (11)
61-70	972,119 (9)
70+	1,872,114 (17)
Sex $^{\#}$	
Female	5,715,579 (52)
Male	5,302,446 (48)
Ethnicity $^{\#}$	
White European	2,079,461 (19)
African-Caribbean	130,632 (1)
South Asian	180,873 (2)
Other Asian	82,305 (1)
Middle Eastern	11,824 (<1)
Mixed	1,049,975 (10)
Other	277,915 (3)
Missing	7,205,040 (65)
Deprivation $^{\#}$	
Lowest	1,554,601 (14)
Lower	2,065,291 (19)
Middle	2,122,244 (19)
Higher	2,001,463 (18)
Highest	2,229,501 (20)
Unknown	1,044,925 (9)
Nation $^{\#}$	
England	8,692,042 (79)
Wales	768,162 (7)
Scotland	1,289,931 (12)
Northern Ireland	267,890 (2)
Follow-up (years) *	13 (6, 22)

$^{\alpha}$: mean (SD); *: median (Q1, Q2); $^{\#}$ n (%)

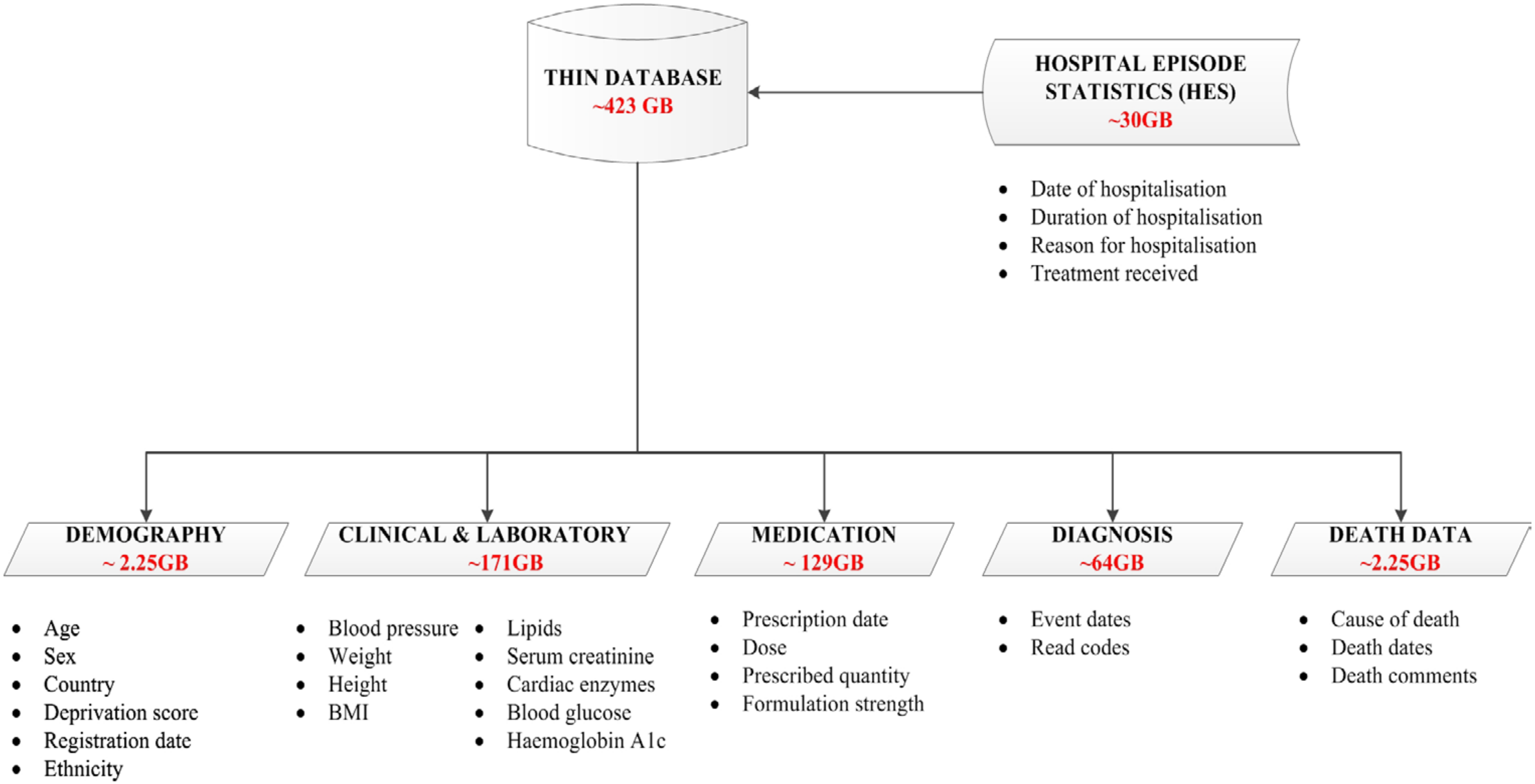


Figure 2.1: Schematic representation of The Health Improvement Network (THIN) database [GB: Gigabytes]

2.1.3 Clinical, laboratory, and prescription data

Individual-level longitudinal data on clinical and laboratory measurements are captured as and when such measures are obtained in the GP centre or laboratories. The clinical data include BMI, systolic and diastolic blood pressure. For patients with T2DM, the laboratory measures include HbA_{1c} (%), random and fasting blood glucose (mmol/L), lipids, serum and urine albumin, and creatinine levels. All clinical and laboratory measures contain dates of measurements. The lipid measures include total cholesterol, high- and low-density lipoprotein cholesterols, and triglycerides. Furthermore, data on the immunological status of infectious diseases, including hepatitis A, B and C, yellow fever, typhoid, influenza, rabies, and smallpox, are also available. Biological markers include cardiac enzymes, liver enzymes, urine biochemistry, sex hormones, HIV test, red blood cell count, shape and size, and adrenal autoantibodies.

Medication data is recorded for any prescription given to a patient from a nurse or GP. This includes information on dates of prescription, duration of the prescription, formulation, strength, dose, and quantity. The coding of medication data is based on both the British National Formulary (BNF) codes and Anatomical Therapeutic Chemical (ATC) codes. Medications captured in the THIN database include antidiabetic drugs, cardio-protective drugs, aspirin, and nonsteroidal anti-inflammatory drugs (NSAIDs), antibiotic, anti-parasitic, vaccines, and vitamin supplements.

2.1.4 Disease event data

Disease status of patients in the THIN database is recorded using the most comprehensive medical coding system in the world−Read codes. These classification codes are not used only for disease coding but also for history and symptoms, examination findings and signs, diagnostic procedures, preventive, operative, therapeutic, administrative procedures, drugs, appliances, occupations, and social information [127,128]. Event dates are recorded for each disease event experienced by patients and the occurrences of medical conditions such as myocardial infarction (MI), stroke, heart failure (HF), CHD, peripheral artery disease (PAD), angina, angioplasty, coronary artery bypass graft (CABG), neuropathy, retinopathy, and renal complications are available. Also, medical histories including amputations, atherosclerosis, coma, seizures, multiple sclerosis, fractures, revascularisations, and surgeries are also available. The data from the primary care is linked with hospital episode statistics (HES). The hospitalisation statistics include the reason(s) for hospitalisation, treatment (invasive and non-invasive) received, duration of hospitalisation, and the records on adverse events. The distribution of selected clinically diagnosed disease is presented in Table 2.3.

Table 2.3: The distribution of some clinically diagnosed diseases in the THIN database

	n (%)
Patients in THIN	11,018,025(100)
DM	444,148 (4)
T1DM	46,238 (0.4)
T2DM	379,657 (3.4)
Gestational diabetes	15,814 (0.1)
Angina	214,798 (2)
Coronary artery disease (CAD)	284,063 (3)
Heart failure (HF)	413,364 (4)
Hospitalisation for HF	209,541 (2)
Myocardial infarction (MI)	1,417,821 (13)
Peripheral artery disease (PAD)	229,472 (2)
Kidney disease	524,621 (5)
Stroke	497,426 (5)
Arrhythmia	362,453 (3)
Cancer	291,013 (3)
Rheumatoid arthritis	600,016 (5)
Retinopathy	13,388 (<1)
Neuropathy	98,788 (1)

2.1.5 Strengths

The major strength of the THIN database is the size, as the version of THIN used for this thesis includes data on over 13 million patients drawn from over 600 practices across the UK. The database also draws on long follow-up of individuals [median (Q1, Q3): 13 (6, 22) years] as one of its key features. In addition to GP consultation data collected from practices in the UK, additional data is also obtained from other healthcare professionals, and the database can be linked to external data sources like HES which is provided by Health and Social Information Centre (HSCIC). More than 75% of THIN practices are now electronically linked to pathology laboratories [122]. Compared to UK national Quality of Outcomes Framework (QOF) data, THIN provides similar estimates of crude prevalence for diabetes, chronic obstructive pulmonary disease (COPD), HF, epilepsy hypertension, mental health, cancer and asthma [124]. All these key advantages combined present researchers with a unique opportunity to investigate chronic as well as rare disease conditions, with long latency and the study of long-term outcomes.

2.1.6 Limitations

As with all observational studies in which longitudinal data is obtained on patients, the major limitations of the THIN database include (1) loss to follow-up, (2) missing data on specific variables, (3) misdiagnosis, misclassification and miscoding, and (4) unreliable data on some relevant variables.

Patients are lost to follow-up when they move to different locations or transfer out of practice. They either move to another participating GP or to a practice that does not participate in contributing data to THIN. Continuous recording of vital longitudinal information ceases for patients who moved to GPs that do not participate in THIN. For those who move to another participating practice, recording of longitudinal information is continued.

Missing communications from specialists, discharge summaries from hospitals, and test results from pathology laboratories among others are some of the reason missing values exist for some variables in the THIN database. This missing data usually follows a complex pattern and requires skilled expertise for analysis of any related data [129]. Due to the nature of GP settings, some variables are recorded more often than others. For example, systolic and diastolic blood pressure measurements may be recorded at every GP encounter because of the relative ease with which it can be measured. Also, the capture of the common adiposity measurements within the THIN database varies. Body weight is measured more often than waist circumference due to the simplicity and standard way it is measured—leading to many missing values on waist circumference.

The potential for misdiagnosis, misclassification, and miscoding of diagnostic codes in EMRs such as the THIN database cannot be understated [130-133]. However, according to a number of studies, based on extensive data mining and quality assessments, most of the diagnosis codes in the UK primary care database are well recorded [134-136]. The Read code system allows easy recording of clinical information on a computer by the GP, without advance knowledge of coding and classification. However, major types of data entry errors like omissions, typing or communicating errors usually result in a relatively small number of false positives, and larger numbers of false negatives patients identified by Read codes. This is usually a problem for disease conditions that are phenotypically heterogeneous. Often, expert domain knowledge, statistical, and programming skills are required to classify patients accurately and distinguish between prevalent and incident diagnosis [137]. Finally, there is also non-availability of complete and reliable data on ethnicity and smoking cessation during follow-up, information on diet, exercise or weight lowering medications.

2.2 STUDY DESIGN

2.2.1 Inclusion criteria

This thesis uses a large comparative longitudinal case-control design to address the main hypothesis that body weight category at the time of diagnosis influences survival outcome in patients with T2DM. Patients were considered for inclusion in this study if they were 18-90 years of age with at least one episode of care between January 1990 and September 2014, and complete data on gender. A cohort of patients with T2DM was identified using modifications of validated approaches [133,138] and machine learning algorithms. The machine learning algorithms used were; (1) Naïve Bayes [139,140], (2) Logistic regression [141], (3) Support Vector Machine (SVM) [142-144], (4) Multilayer Perceptron (MP) [145,146], (5) Decision Tree with J48 modification [147,148], and (5) One Rule [149]. The use of these clinical and machine learning algorithms to extract specific DM subgroups from the THIN database is discussed later in Chapter 3.

2.2.2 Control subjects

The control (non-diabetic) cohort was defined as patients without any diagnostic codes suggestive of diabetes or an antidiabetic medication or elevated blood glucose measurement or glycated haemoglobin (HbA_{1c}) measurement during the whole period of follow-up. Appropriate controls from the pool of non-diabetic control patients were matched to each T2DM patient in a ratio of 1:4 using an exact matching algorithm. The dynamic matching conditions were the year of birth, sex, and separately for ethnicity where ethnicity data were available. Furthermore, the index date for controls was defined as the date of the T2DM diagnosis for the matched cases (discussed in Chapter 4).

2.2.3 The arrangement of longitudinal covariate data

Longitudinal measures of body weight, BMI, waist circumference, systolic and diastolic blood pressure, HbA_{1c}, random blood glucose, fasting blood glucose, low-density and high-density lipoprotein cholesterols, triglycerides, serum albumin and creatinine, and glomerular filtration rate in the 36 months prior to the diagnosis of T2DM and 84 months following the diagnosis date were extracted and arranged in six-monthly windows. All available measures on or within three months before the diagnosis date were considered as the index date (date of diagnosis of T2DM) measures. If more than one measurement existed within this interval, the closest to index date was taken.

2.2.4 Extraction of longitudinal anti-diabetic drugs (ADD)

The complete list of generic and brand names used to extract ADDs from THIN database are presented in Appendix A, Table 4. Complete information on ADDs including prescription dates and classes of ADD was extracted for each patient with T2DM. Among patients who had at least two prescriptions of ADDs, time spent before first-line therapy, type of first-line therapy, those who remained on first-line therapy as well as those who added or switched to another ADD(s) were extracted. By comparing prescription initiation and cessation dates, the addition of a second anti-hyperglycaemic drug was defined if cessation date of the first drug is more or equal to the start date of the second drug. In contrast, if the cessation date of the first drug is less than the start of the second drug then this was defined as switching to a second drug (Appendix B, Table 5 and 6).

Subsequently, body weight, BMI, systolic and diastolic blood pressure, and HbA_{1c} measured at initiation of the first line and second line therapy was included on the basis of a 3-month window on or within the start of first- and second-line therapy respectively. Follow-up measures of body weight, BMI, systolic and diastolic blood pressure, and HbA_{1c} during 36 months were arranged longitudinally on the basis of non-overlapping 6-monthly intervals which were defined progressively from initiation of first and second line therapy respectively.

2.2.5 Other covariate data

Complete records on the prescriptions of different classes of antihypertensive drugs, weight lowering drugs, anti-depressant drugs, and lipid-modifying drugs were extracted along with the dates of prescriptions. Other covariate data included the date of registration with practice, health authority, nation, date of birth, deprivation score (a socioeconomic status measure based on residential address [126]), and ethnicity.

2.2.6 Outcome variables

The primary outcome was all-cause mortality and time to death was calculated as the time from diagnosis of T2DM to occurrence of death. Information on deaths with dates and possible cause of death were also extracted. Secondary outcomes of interest were comorbid diseases that occurred before or after diagnosis of T2DM. These were identified using Read codes and included CKD, cancer, angina, non-fatal MI, coronary artery disease (including bypass surgery and angioplasty), HF, bariatric surgery, depression, rheumatoid arthritis, and stroke. Time to a specific disease event or death was calculated as the time from the diagnosis date to the first occurrence of the disease event or date of death respectively. Patients who were still alive at the end of the study (September 2014) were censored on the end date or censored on drop out date.

2.3 STATISTICAL METHODS

2.3.1 Dealing with missing longitudinal measurements

Where necessary, missing clinical and laboratory data were imputed for under varying follow-up scenarios. A complete assessment of missing covariate data including the description of missing data mechanisms and patterns is provided in Chapter 4. Missing data patterns were explored using exploratory analysis and an assumption about the underlying missing data mechanism made. Subsequently, multiple imputation approaches were used to impute for missing longitudinal covariate data. Estimates of central tendency were used to assess the consistency of imputation.

2.3.2 The distribution of study variables

Checks for normality of continuous variables were performed using density plots and histograms. All variables were checked for outliers, and inconsistent values were put to missing before imputation. For categorical variables like smoking status, exercise, and ethnicity, dummy variables were created to indicate missing status. Summary statistics of the study population was summarised as number (percentage), mean (SD) or median (Q1, Q3), as appropriate. A two-sample t-test or Scheffe's multiple comparison post hoc ANOVA test were used to test for significant difference in means between two and more than two groups respectively, where appropriate. Similarly, a non-parametric Kruskal Wallis test was used to compare medians across groups of interest. Finally, the chi-square test was used to identify significant differences in different categorical study parameters across groups.

2.3.3 Presentation of longitudinal distribution of risk factors

For continuous longitudinal measurements, a generalised linear model under general estimating equations setup, with unstructured covariance fitted. Separate analyses were conducted for each BMI category, and the unadjusted and adjusted mean (95% confidence intervals, CI) of longitudinal 6-monthly measures of body weight before and after T2DM diagnosis were estimated respectively.

2.3.4 Analysis of disease event data: calculation of rates/risk (hazard ratios)

Disease events and mortality rates were calculated using standard life-table analysis technique, after calculating the time to such events from the index date. The event rates per 1000-person-years along with their 95% CIs were calculated.

For evaluating the possible association of exposure, including the BMI at diagnosis of T2DM, with the risk of cardiovascular diseases and all-cause mortality, different multivariate regression based risk models were used. These include the stratified multivariate Cox regression models and the novel treatment-effect models.

An example of the multivariate Cox regression model for investigating the association between BMI at diagnosis and all-cause mortality, with adjustments for covariates and confounders, is presented below:

i. ***Simple Model**: Risk of Event ~ function of (age, sex, smoking status, and baseline SBP, DBP, HbA$_{1c}$, oral ADDs, insulin and BMI categories)*

ii. ***Extended Model**: Risk of Event ~ function of (all components of Simple Model plus LDL, HDL and triglyceride measures at baseline)*

Separately inferences of HRs for each BMI category compared to the grade 1 obese category were obtained for patients with and without a history of diseases at diagnosis (defined as the occurrence of cardiovascular disease, cancer and renal diseases before the index date).

2.3.5 Treatment effects model

Given the observational nature of this study, significant differences in risk factors between groups being compared could impact any inference obtained. To account for the inherent differences in risk factors between the defined BMI categories, the novel "treatments effects" modelling approach was used to provide robust inferences. This modelling approach uses the potential outcomes or counterfactual framework to allow comparison of survival time for CVD and all-cause mortality for patients with different BMI categories. Briefly, given an observed outcome (Y_0), for a patient with normal weight, the potential outcome or the counterfactual (Y_1) for this same patient is the outcome if the patient had belonged to another BMI category and vice versa. Therefore, the average of the difference between the observed outcomes given a specific BMI category and the potential outcome is the average treatment effect [i.e., average treatment effect (ATE) = average (Y_1-Y_0)] [113-116]. Since the outcome of interest is survival time, a survival model with inverse-probability weight estimator was used to estimate ATE for each BMI category, with appropriate adjustments and balancing of confounders.

2.4 ETHICAL CONSIDERATIONS

The THIN data collection was approved by the NHS South-East Multi-Centre Research Ethics Committee (MREC) in 2003. Access to this database for research purposes is granted in the form of a sub-license which enables access to the entire dataset for the period of the sub-license. QIMR Berghofer Medical Research Institute has obtained formal access to this database. The protocol for this study was approved by the Scientific Review Board managed by the THIN database vendor company (15THIN030, 17th August 2015).

Chapter 3: Cohort Identification from Primary Care Database

This body of this chapter contains one published paper that discusses different approaches used to robustly identify and extract a cohort of type 2 diabetes (T2DM) patients using rule-based clinically guided algorithms and a machine learning algorithm from the THIN database. The citation of the published paper is as follows:

Owusu Adjah ES*, Montvida O*, Agbeve J, Paul SK. Data Mining Approach to Identify Disease Cohorts from Primary Care Electronic Medical Records: A Case of Diabetes Mellitus. *The Open Bioinformatics Journal* 2017;**10**:16-27. * Joint first authors

All the listed have agreed to the inclusion of this published scholarly work in this thesis and the statement of my contribution to the authorship of this published scholarly work is included below:

Contributor	Statement of contribution
Owusu Adjah Ebenezer S. (Candidate)	Responsible for the primary design of the study and the methodological developments. Conducted the data extraction from the THIN database. Responsible for data manipulation, aggregation, transformation in SAS. Performed data manipulation and built the data mining/machine learning workflow in WEKA. Conducted the statistical analyses in SAS and contributed towards the interpretation of results. Developed first draft and contributed towards finalisation of the manuscript.
Montvida Olga	Contributed to the primary design of the study. Evaluated the methodological approach, contributed towards the data extraction, building of data mining/machine learning workflow in WEKA, interpretation of results and finalisation of the manuscript.
Agbeve Julius	Evaluated the methodological approach and contributed towards finalisation of the manuscript.
Paul Sanjoy K	Conceived the idea, was responsible for the primary design of the study and the methodological developments and contributed towards finalisation of the manuscript.

3.1 ABSTRACT

Background

Identification of diseased patients from primary care based electronic medical records (EMRs) has methodological challenges that may impact epidemiologic inferences.

Objective

To compare deterministic clinically guided selection algorithms with probabilistic machine learning (ML) methodologies for their ability to identify patients with type 2 diabetes mellitus (T2DM) from large population-based EMRs from nationally representative primary care database.

Methods

Four cohorts of patients with T2DM were defined by deterministic approaches based on disease codes. The database was mined for a set of best predictors of T2DM, and the performance of six ML algorithms was compared based on cross-validated true positive rate, true negative rate, and area under receiver operating characteristic curve.

Results

In the database of 11,018,025 suitable research individuals, 379 657 (3.4%) were coded to have T2DM. Logistic Regression classifier was selected as the best ML algorithm and resulted in a cohort of 383,330 patients with potential T2DM. Eighty-three percent (83%) of this cohort had a T2DM code, and 16% of the patients with T2DM code were not included in this ML cohort. Of those in the ML cohort without disease code, 52% had at least one measure of elevated glucose level, and 22% had received at least one prescription for antidiabetic medication.

Conclusion:

Deterministic cohort selection based on disease coding potentially introduces significant misclassification problem. ML techniques allow testing for potential disease predictors, and under meaningful data input, can identify diseased cohorts holistically.

3.2 INTRODUCTION

Recent advances in the design and implementation of large patient-level electronic medical records (EMRs) from national primary care databases have created opportunities in clinical, epidemiological and public health research [150,151]. In a typical primary or ambulatory care setting, large volumes of data are generated as patients go through various phases of treatment. Individual patients' longitudinal data on demographics, lifestyle, disease and treatment history, clinical and laboratory parameters, hospitalisation statistics, and clinical events are typically organised and stored in the form of a relational database. Such databases present unique challenges in terms of efficient and effective extraction of data for various investigative interests [152]. One of the challenging aspects in this context is the identification of disease cohorts for retrospective or prospective clinical, epidemiological studies [133,153].

Diagnostic codes, such as the International Classification of Diseases (ICD) codes or Read codes [128], are generally used to identify disease cohorts from EMRs [153]. The reliability of diagnosis coding for various diseases has been extensively examined for many primary care databases including The Health Improvement Network (THIN) database from the United Kingdom [135] [154,155]. However, there are four specific issues in relation to identifying cohorts by diagnostic codes: (1) differentiating between disease subtypes from high-level codes, (2) overlapping codes of disease subtypes longitudinally at individual patient level, (3) absence of codes for diseased patients (false negatives), and (4) presence of disease-specific codes for patients without the specific disease (false positives).

With regards to diabetes mellitus (DM), identification and appropriate classification of different types of diabetes in the primary care databases are particularly challenging [130,131,133,156,157]. These challenges border mostly on inaccurate coding leading to misclassification, misdiagnosis, and undiagnosed diabetes [157]. Algorithms based on laboratory, clinical, and medication data have thus been proposed as tools for distinguishing between type 1 diabetes mellitus (T1DM) and type 2 diabetes mellitus (T2DM) [138,156,158,159]. However, the overall accuracy and reliability of derived disease cohorts based on diagnostic codes can be improved by implementing advanced machine learning (ML) or statistical data mining techniques and clinically guided cohort selection algorithms that robustly capture comprehensive patient-level information available in the EMRs [130,133,153,157].

Shivade and colleagues (2014) have conducted a systematic review of various techniques used for the identification of different disease cohorts from different sources of clinical databases [151]. Some of these proposed algorithms have been criticised for their appropriateness in the context of other studies [160]. While several studies compared or applied ML techniques to identify T2DM patients, to the best of our knowledge, there is no study that employed an extensive assessment of diagnostic codes, deterministic clinical selection algorithms, and ML algorithms simultaneously to identify T2DM cohorts from primary care EMRs.

The aims of this exploratory methodological study were to (1) explore technical challenges in the extraction of disease cohorts, (2) compare the ability of different clinically guided cohort selection algorithms to identify the disease cohorts, and (3) compare the disease cohorts identified by ML algorithms and clinically guided cohort selection algorithms using a large nationally representative primary care database from the UK.

3.3 METHODS

In this section, the challenges in identifying a cohort of patients with specific disease (i.e., T2DM), as well as an explanation of the clinically guided cohort selection algorithms, and the data mining and computational processes leading to the comparison of different supervised ML techniques are provided.

3.3.1 Challenges in identifying disease cohorts

THIN uses the UK's standard Read code classification system which is useful for hierarchical classification of patients' specific circumstances and lifestyles, thereby enhancing scalability and retrieval [128]. However, the Read coding system is complex as a disease or an encounter with a GP can be coded in several ways including the use of existing codes or by creating new user-defined codes [161]. This way, considerable variation, and inconsistency are introduced into the coding system as seen in the case of DM [131,138,162].

Differentiating between disease subtypes

Typically, many diabetes-related codes are available for a single patient, some of which are high-level codes (e.g., C10 - "Diabetes mellitus") or disease-related codes that are unspecific in the description of the diabetes type (e.g., C106.12-"Diabetes mellitus with neuropathy"). Common practice has been to exclude any high-level codes [138,163] which may lead to underestimation of the disease cohort. When it is impossible to identify disease subtype (type 1 or type 2 diabetes) from the diagnostic codes, data on surrogate markers (like glutamic acid carboxylase) could be useful, but such information is not available in the THIN database. Nevertheless, combinations of available biomarkers (such as age, weight or HbA_{1c}) and medication prescriptions have been used to distinguish types of diabetes in some studies [138,156].

Longitudinally overlapping disease subtypes

Patients may have different disease subtypes coded longitudinally as a result of data entry errors or the natural progression of the disease. While the former can lead to any combinations of subtypes, the latter may result in developing T1DM from T2DM or T2DM from gestational diabetes. To distinguish between contradictory codes, longitudinal exploratory techniques were applied in some studies [133]. Also, the techniques described above that deal with unspecific codes may be considered. To address the issue of contradictory diagnostic codes longitudinally, the following was adopted to distinguish between T1DM and T2DM:

i. Use of Read codes that uniquely distinguish between T1DM and T2DM
ii. In patients with unspecific codes or longitudinally overlapping subtypes, the following is used:
    a. If an oral antidiabetic drug is taken ≥ 2months, then T2DM.
    b. Otherwise, if age at first available diagnosis date ≤ 18 years and insulin initiated within 1 year, then T1DM.
    c. Otherwise, if age at first available diagnosis date > 18 years and insulin initiated within 3 months then T1DM.
    d. Else T2DM.
iii. Patients with codes for gestational diabetes and other forms of diabetes were excluded.

The absence of codes for diseased patients and the presence of codes for non-diseased patients

Data entry errors such as omissions, typing, communicating errors and patients' temporary loss of follow-up in EMRs usually result in a relatively small amount of false positive, and larger numbers of false negative patients identified by diagnostic codes. Earlier studies have addressed this complex issue by employing deterministic or probabilistic algorithms [151,158,159]. We further focus on this challenging aspect by comparing deterministic (clinically guided), and probabilistic (ML) cohort identification approaches.

3.3.2 Clinically guided cohort selection algorithms

Four separate cohorts were created by applying logical, clinically guided algorithms that select patients from those who have at least one record of Read code for T2DM (Figure 3.1). The complete list of Read codes used is presented in Appendix A, Tables 1, 2, and 3.

Specifically, the T2DM cohorts were selected from available records for T2DM as follows:

i. *Selection algorithm 1*: T2DM Read code (Cohort 1);
ii. *Selection algorithm 2*: Lifestyle modification advice + T2DM Read code (Cohort 2);

iii. *Selection algorithm 3*: At least one prescription for antidiabetic medication + lifestyle modification advice + T2DM Read code (Cohort 3);

iv. *Selection algorithm 4*: At least one prescription for antidiabetic medication or lifestyle modification advice + T2DM Read code (Cohort 4);

3.3.3 Supervised machine learning techniques

The process of selecting one most appropriate probabilistic algorithm to identify patients with T2DM is described below.

Feature selection

The THIN database was mined to detect the most frequent medications, comorbidities, laboratory and anthropometric measurements among patients with T2DM identified on the basis of Read codes. The resulting 280 variables were combined with current clinical considerations, practices, and guidelines for T2DM management [164], and 11 potential disease predictors were obtained through an iterative process (Table 1). Correlation-based Feature Selection (CFS) algorithm was applied to determine best of these predictors [165] [166]. This scheme independent attribute subset selection approach is particularly useful when attributes are correlated with one another, and with the class attribute. Bi-directional, forward and backward greedy search methods were applied using 10-fold cross-validation [146], and they all agreed on the same seven features described in Table 3.1.

Training dataset

From the 11,018,025 patients in the THIN database, a training dataset of 150,000 instances, containing an equal number of positive and negative representatives was extracted. Positive instances were randomly selected from patients with (1) available T2DM Read code, (2) at least one year of follow-up, and (3) 18-90 years old at the time of T2DM diagnosis. Negative instances were also randomly selected from those without Read code for any subtype of DM and at least one year of follow-up (Figure 3.2, training set).

Table 3.1: Features selected as best T2DM predictors.

	Feature name	**Feature type**	**Selected for ML**
1	Two measurements of HbA_{1c}>6% or fasting blood glucose > 7 mmol/l or random blood glucose > 11.1 mmol/l within 1 year	Binary	Yes
2	Any antidiabetic drug prescriptions for at least 6 months	Binary	Yes
3	Average BMI	Continuous	Yes
4	Hypertension diagnosis or antihypertensive drug use greater or equal to 6 months or beta blockers prescription for 6 months or more	Binary	Yes
5	Chronic kidney diagnosis	Binary	Yes
6	Retinopathy or neuropathy diagnosis	Binary	Yes
7	Average systolic blood pressure	Continuous	Yes
8	Lifestyle modification advice	Binary	No
9	Average HbA_{1c}	Continuous	No
10	Average Random Glucose	Continuous	No
11	Heart Failure or Myocardial Infarction or Stroke or Coronary Artery Disease	Binary	No

Classification algorithm selection

Keeping the selected subset of 7 robust predictors of T2DM, six classification algorithms were applied to the training set. Ten repeat 10-fold cross-validation was applied to calculate true positive rate (sensitivity), true negative rate (specificity), and area under receiver operating characteristic curve (AUC). Percent of correctly classified instances and required central processing unit (CPU) time for training the algorithms were also derived. The algorithms for comparison were: Naïve Bayes [139,140], Logistic regression [167], Support Vector Machine (SVM) [168,169], Multilayer Perceptron (MP) [145], Decision Tree with J48 modification [147], and One Rule [149].

One Rule algorithm performed significantly worse. Except for differences in CPU time, the performance of other algorithms was similar. Among them, Naïve Bayes had lower sensitivity misclassifying approximately 500 additional patients compared to other approaches. AUC was smaller for SVM and J48, while SVM and MP required significantly higher CPU time (Table 2). Interestingly, neither body mass index nor blood pressure contributed significantly to any model. Logistic regression was selected as the most appropriate model for predicting T2DM. The model obtained from full training dataset was applied to all THIN database patients with no record of Read code for diabetes diagnosis other than T2DM, and with available follow-up for at least one year (Figure 2, prediction set).

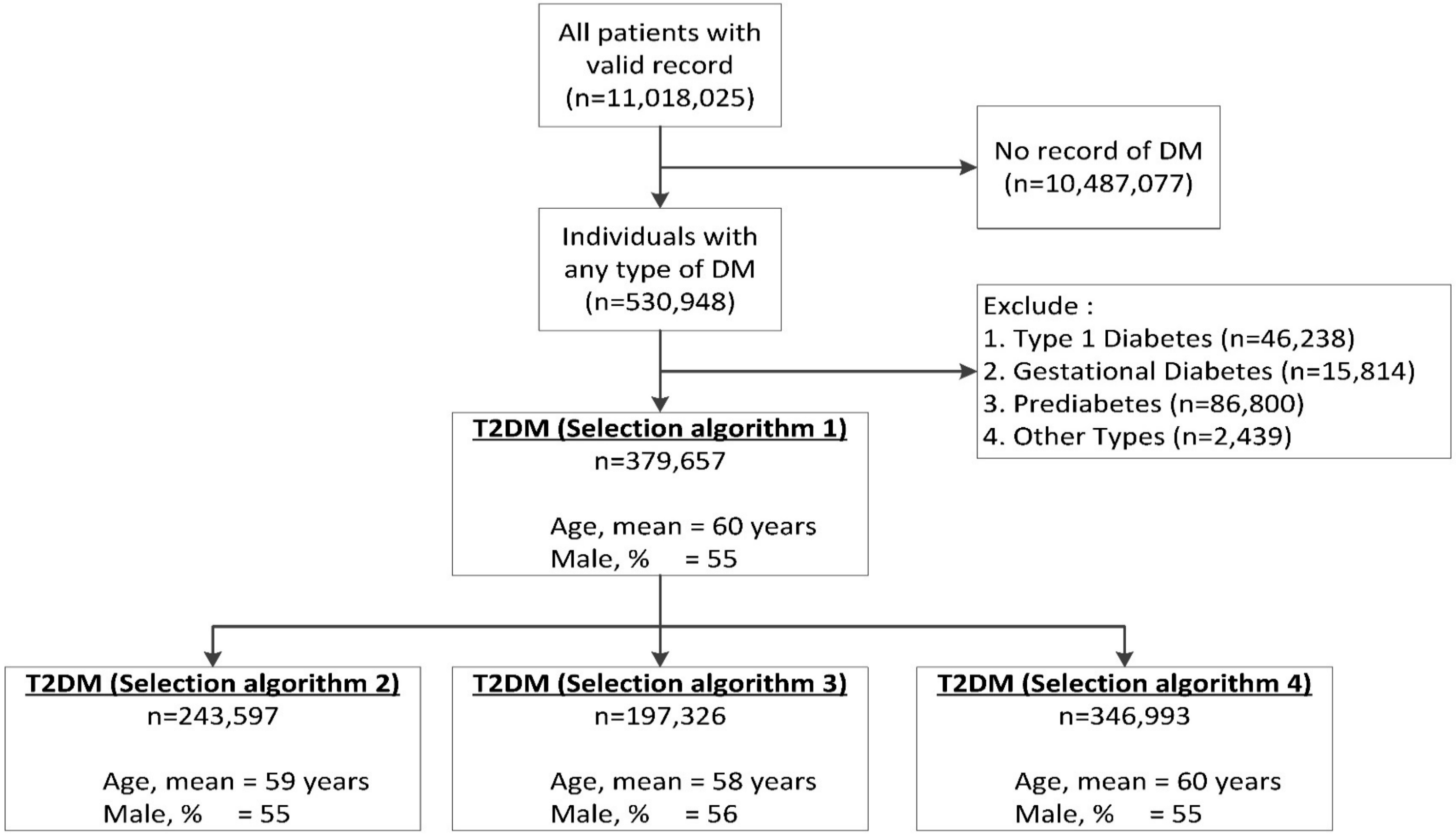

All patients with valid record (n=11,018,025)
No record of DM (n=10,487,077)
Individuals with any type of DM (n=530,948)
Exclude :
1. Type 1 Diabetes (n=46,238)
2. Gestational Diabetes (n=15,814)
3. Prediabetes (n=86,800)
4. Other Types (n=2,439)
T2DM (Selection algorithm 1)
n=379,657
Age, mean = 60 years
Male, % = 55
T2DM (Selection algorithm 2)
n=243,597
Age, mean = 59 years
Male, % = 55
T2DM (Selection algorithm 3)
n=197,326
Age, mean = 58 years
Male, % = 56
T2DM (Selection algorithm 4)
n=346,993
Age, mean = 60 years
Male, % = 55

Figure 1.1: Flowchart for the selection of type 2 diabetes (T2DM) cohorts by clinically guided algorithms.

[Selection algorithm 1: T2DM Read code only; Selection algorithm 2: T2DM Read code + lifestyle modification advice. Selection algorithm 3: T2DM Read code + antidiabetic medication + lifestyle modification advice. Selection algorithm 4: T2DM Read code + (antidiabetic medication or lifestyle modification advice)]

Table 1.2: Performance of machine learning algorithms on the training dataset.

	Naïve Bayes	**Logistic Regression**	**Multilayer Perceptron**	**Support Vector Machine**	**J48 Decision Tree**	**One Rule**
Percent correct	95.6	95.9	95.9	95.9	95.9	91.7
TPR	0.98	0.99	0.99	0.99	0.99	0.99
TNR	0.93	0.93	0.93	0.93	0.93	0.84
AUC	0.98	0.98	0.98	0.96	0.96	0.92
CPU time	0.09	3.36	68.03	191.9	1.78	0.21

TPR: True Positive Rate, TNR: True Negative Rate; AUC: Area Under receiver operating characteristic Curve; CPU: Central Processing Unit

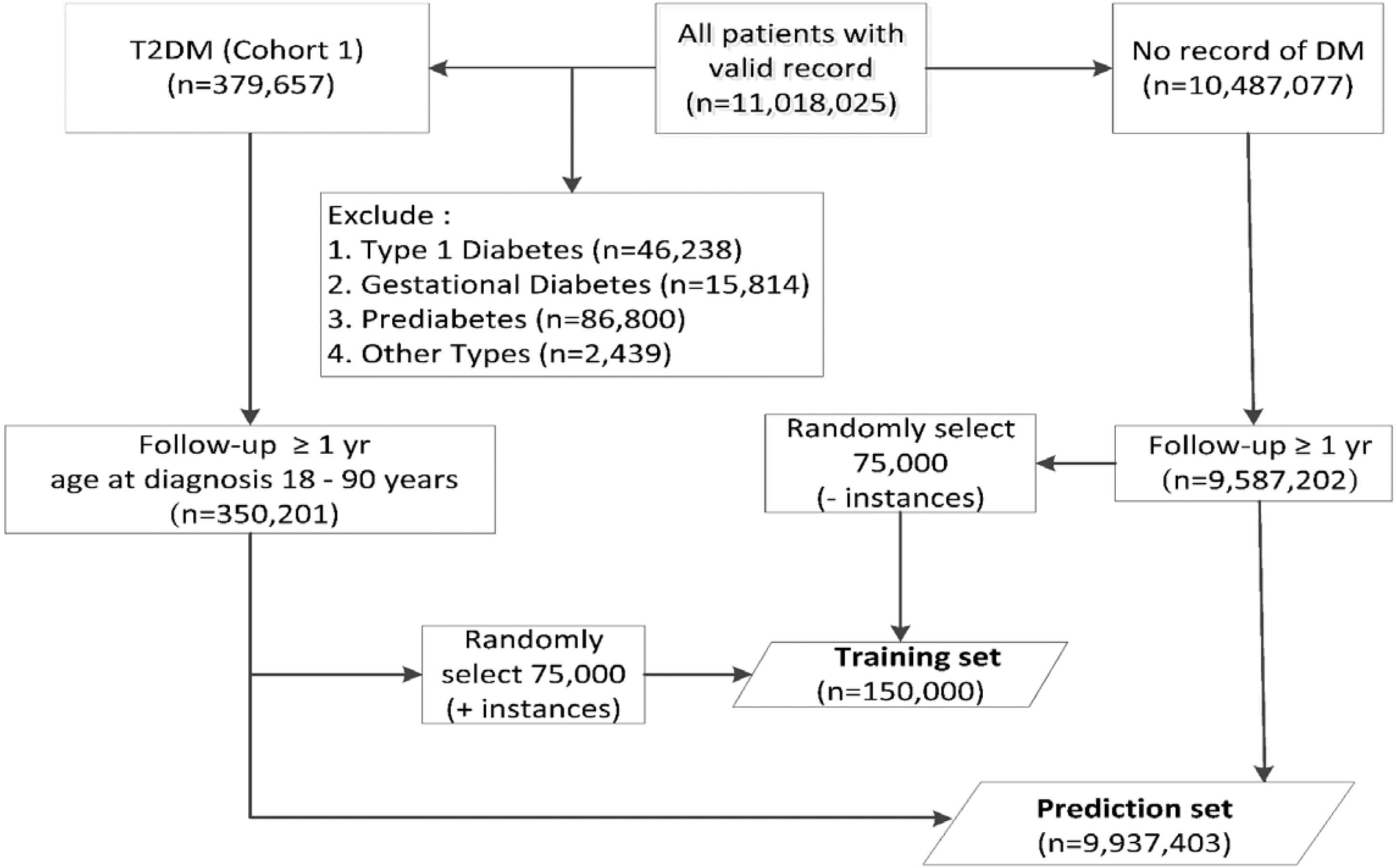


Figure 1.2: Flowchart for creating a dataset for machine learning training, and of the dataset for predicting diabetes status

3.4 RESULTS

The distributions of basic characteristics of patients identified by all four clinically guided algorithms and the ML algorithm were similar (Table 3.3). Clinically guided algorithms 1-4 and the ML algorithm resulted in cohorts of 379,657; 243,597; 197,326; 346,993; and 383,330 patients with T2DM respectively. For patients identified by the ML algorithm who did not have a Read code, the first available date of entry of the significant predictors was used as their date of diagnosis. At the time of diabetes diagnosis, identified patients were on average 60 years old, 86 kg in weight with 55% male. The proportions of those who had two elevated glucose level measurements within one year were 75, 86, 90, 79, and 82% of cohorts identified by selection algorithms 1-4 and ML respectively. With median 11 years of follow-up post-diagnosis, proportions of those who received at least one prescription for antidiabetic medication were 79, 81, 100, 87, and 75 % in cohorts identified by rules 1-4 and ML respectively. Among the cohort of T2DM patients identified by ML algorithm, 317,979 (83% of 383,330) patients had Read code for T2DM (Table 3.4). It is worth noting that 59,678 (16% of 379,657) patients with a record of T2DM Read code were not selected by ML approach. Almost a fifth (17% of 383,330) of the patients in the ML cohort were without a record of T2DM Read code. Of them, 52% had at least one measure of elevated glucose level, and 22% had received at least one prescription for antidiabetic medication (Table 3.4). To assess the proportion of patients that remain undetected by the algorithms used in this study, complement cohort-specific analysis was performed (data not shown). Among patients not selected by ML as T2DM, only 884 patients had at least two elevated glucose measurements (HbA_{1c} > 6% or fasting blood glucose > 7 mmol/l or random blood glucose > 11.1 mmol/l) within 1 year, compared to 32,039, 106,671, 137,796, and 42,583 patients not selected by selection algorithms 1-4.

Table 1.3: Baseline characteristics of T2DM patients identified by selection algorithms and logistic regression classifier (ML).

	Selection algorithm 1	**Selection algorithm 2**	**Selection algorithm 3**	**Selection algorithm 4**	**ML**
Patients, n	379,657	243,597	197,326	346,993	383,330
Age at diagnosis (years) $^{\alpha}$	60 (15)	59 (14)	58 (14)	60 (15)	59 (15)
Age at diagnosis (years) *	61 (50,71)	60 (50,69)	58 (49,67)	60 (50,70)	60 (50,70)
≤40	32,644 (9)	19,761 (8)	17,969 (9)	29,701 (9)	71,752 (19)
41-50	62,656 (17)	43,872 (18)	39,289 (20)	59,608 (17)	58,813 (15)
51-60	90,464 (24)	62,610 (26)	54,006 (27)	85,587 (25)	84,277 (22)
61+	193,893 (51)	117,354 (48)	86,062 (44)	172,097 (50)	168,488 (44)
Male $^{\#}$	208,155 (55)	134,393 (55)	110,178 (56)	191,107(55)	200,447 (52)
At least one prescription$^{\#}$	300,722 (79)	197,326 (81)	197,326 (100)	300,722 (87)	287,095 (75)
Prescription duration ≥ 6 months$^{\#}$	243,064 (64)	171,800 (71)	171,800 (87)	243,064 (70)	254,255 (66)
RBG (mmol/l) $^{\alpha\,\S}$	11.5 (5.1)	11.4 (5.1)	12.1 (5.3)	11.6 (5.2)	11.3 95.2)
RBG (mmol/l) $^{\alpha\,\ddagger}$	9.5 (3.4)	9.4 (3.3)	9.9 (3.4)	9.6(3.4)	9.1 (3.5)
FBG (mmol/l) $^{\alpha\,\S}$	8.4 (2.3)	8.4 (2.3)	8.9 (2.4)	8.5 (2.3)	8.3 (2.3)
FBG (mmol/l) $^{\alpha\,\ddagger}$	7.8 (2.1)	7.7 (2.0)	8.0 (2.1)	7.8(2.1)	7.5 (2.1)
HbA$_{1c}$ (%)$^{\alpha\,\S}$	8.4 (2.1)	8.4 (2.1)	8.7 (2.2)	8.5 (2.2)	8.3 (2.1)
HbA$_{1c}$ (%)$^{\alpha\,\ddagger}$	7.5 (1.4)	7.5 (1.3)	7.7 (1.3)	7.5(1.4)	7.4 (1.3)
Composite measure$^{\#\,\ddagger}$	283,419 (75)	208,787 (86)	177,689 (90)	272,875 (79)	314,574 (82)
Weight (kg) $^{\alpha\,\S}$	89.4(20.8)	90.3 (21.0)	91.1 (21.1)	89.6 (20.9)	89.3 (21.0)
Weight (kg) $^{\alpha\,\ddagger}$	85.0 (19.8)	86.6 (19.9)	87.6 (20.0)	85.5 (19.8)	86.1 (20.6)
BMI (kg/m^{2}) $^{\alpha\,\S}$	31.6 (6.7)	32.0 (6.7)	32.2 (6.7)	31.7 (6.7)	31.7 (6.8)
BMI (kg/m^{2}) $^{\alpha\,\ddagger}$	30.2 (6.1)	30.7 (6.1)	31.0 (6.2)	30.4(6.1)	30.7 (6.7)
Normal weight $^{\#}$	22311(12)	15,821 (11)	12,339 (11)	21,108 (12)	24,453 (13)
Overweight $^{\#}$	58,447 (32)	44,283 (32)	35,289 (31)	55,885 (32)	61,846 (32)
Grade 1 obese $^{\#}$	52,465 (29)	41,323 (30)	33,669 (30)	50,423 (29)	55,684 (29)
Grade 2 obese $^{\#}$	27,168 (15)	22,163 (16)	18,497 (16)	26,336 (15)	29,178 (15)
Any CVD$^{\#}$	106,523 (28)	67,011 (28)	51,905 (26)	96,147 (28)	93,703 (24)
CKD$^{\#}$	10,547 (3)	8,035 (3)	4,609 (2)	9,445 (3)	12,404 (3)
Cancer$^{\#}$	24,159 (6)	15,998 (7)	11,084 (6)	21,536 (6)	22,112 (6)
Hypertension$^{\#}$	149,752 (39)	104,916 (43)	79,193 (40)	137,440 (40)	140,341 (37)
Follow-up (years) *	11 (6,17)	10 (6,15)	11 (6,16)	11(6,17)	10 (5,16)

Legend*:* Selection algorithm 1: Read code only; Selection algorithm 2: Read code and lifestyle modification advice; Selection algorithm 3: Read code and medication and lifestyle modification advice; Selection algorithm 4: Read code and (medication or lifestyle modification advice); ML: Machine learned cohort; RBG: random blood glucose; FBG: fasting blood glucose; Composite measure: fasting blood glucose > 7mmol/l or random blood glucose >11.1mmol/l or HbA$_{1c}$ >6; BMI: Body Mass Index (kg/m^{2}); Normal : (18.5-24.99), Overweight: (25-29.99); Grade 1 obese: (30-34.99), Grade 2 obese (35-39.99); $^{\alpha}$: Mean(SD); *****: median (IQR); $^{\#}$: n (%); CKD: Chronic kidney disease ; Any CVD: any cardiovascular disease defined as occurrence of angina, MI, coronary heart disease (CHD), HF, stroke, and peripheral artery disease (PAD) on or before diagnosis of T2DM; §: measured at diagnosis and ‡ : an average over of all available measurements.

Table 1.4: Baseline characteristics and distribution of glycaemic markers among patients identified by ML.

	Machine Learned T2DM cohort (n=383,330)	
	With Read code	**Without Read code**
Patients $^{\#}$	319,979 (83)	63,351 (17)
Age at diagnosis (years) $^{\alpha}$	60 (14)	54 (24)
Age at diagnosis (years) *	60 (50, 70)	56 (33, 73)
≤ 40	25,645 (8)	46,107 (73)
41-50	56,583 (18)	2,230 (4)
51-60	81,262 (25)	3,015 (5)
61+	156,489 (49)	11,999 (19)
Male $^{\#}$	176,568 (55)	23,879 (38)
At least one prescription $^{\#}$	273,272 (85)	13,823 (22)
Prescription duration ≥ 6 months $^{\#}$	241,517 (76)	12,738 (20)
RBG >11.1 mmol/l $^{\#}$**,**	101,135 (32)	1,471 (2)
FBG > 7 mmol/l$^{\#}$	50,446 (16)	1,695 (3)
HbA$_{1c}$ > 6 %$^{\#}$	274,565 (86)	29,793 (47)
Composite measure$^{\#}$	274,565 (86)	29,793 (47)

Legend: RBG: random blood glucose; FBG: fasting blood glucose; Composite measure: fasting blood glucose > 7mmol/l or random blood glucose >11.1mmol/l or HbA$_{1c}$ > 6; ********:** *median (IQR),* ***#****: n (%),* ***α****: mean (SD)*

3.5 DISCUSSION

In this study, a number of problems encountered by computer-based methods in the complex tasks of identifying a disease cohort from large EMR databases are addressed. Specifically, (1) the common technical challenges in differentiating diabetes subtypes were defined and discussed, (2) combining clinical, medication and morbidity information with database patterns, a set of best predictors as feeds to ML algorithms that can be used to identify patients with T2DM in the absence of any disease code were selected, and (3) a comparison of T2DM cohorts identified by clinically guided selection algorithm and ML algorithm was made. The results of this study are of particular interest to researchers who work with the THIN database. However, methods explored in this study are generalizable for any EMR with different disease coding systems.

Although there was no difference in distributions of basic characteristics among cohorts obtained by deterministic and probabilistic approaches, ML algorithms were found to be superior. With the use of selected features, we could confirm that 83% of the patients identified by the ML algorithm had a Read code for T2DM (Table 4.3). Those without Read code had a comparable high risk as identified by the significant predictors. While 25 / 21% of patients with Read code / Read code + (medication or lifestyle advice) for T2DM did not have at least two elevated measures of blood glucose within one year, only 18% of ML identified cohort did not have such measures. Among Read code / ML defined patients without elevated composite glucose measure, 69 / 41 % did not receive ADD for at least 6 months. It is important to note that the patients without a Read code for diabetes are highly less likely to have a 2 elevated blood glucose measures within one year unless they were known to be diabetic or pre-diabetic.

Five of the six ML algorithms demonstrated similar performances in the training-testing data sets. Logistic regression approach was chosen as the best classifier for THIN database, however different feature patterns within other EMRs could potentially lead to better performance of other ML techniques to predict T2DM cohort. Tapak and colleagues [170] reported SVM as the better classifier, while Mani and colleagues [171] reported decision trees to outperform other ML algorithms. In this context it is important to mention that, ML algorithms cannot operate without meaningful data fed-in ("Garbage in, garbage out" principle). Although the use of different datasets makes it difficult for direct comparisons, a critical part of ML steps is the feature engineering or selection. Some recent studies have used large sets of variables associated with diabetes with the aim of enhancing the predictive accuracy [172,173]. However, this may be limited by the inclusion of irrelevant and redundant variables, and model overfitting in cases where the number of observations are less than the number of variables. While earlier studies were primarily based on clinically guided feature selection, a more holistic approach was adopted in the current study, initially to identify the data-driven candidates as potential predictors of T2DM from the whole database. Combining clinical knowledge and data-driven candidate predictors, the selection of the most robust set of 7 predictors, was ensured. Although selected features were not surprising, it was observed that BMI, lifestyle modification advice, and hypertension did not contribute to the models, while microvascular complications did.

The performances of six classification algorithms on a set of 150,000 instances were evaluated and reconfirmed to be large enough by assessing the performance curves of several incremental classifiers. Nevertheless, training dataset was small compared to the whole database; therefore, to ensure that the results were not prone to selection bias, the same analyses were performed on two other randomly selected training datasets and almost identical results obtained.

Unlike most ML applications that focus on training to ensure best fit for future predictions, in this study, various techniques to correct available labelling with the ultimate goal to improve quality of diseased cohort (Type 2 Diabetes) was used. It would be of great interest to compare ML error, Rule-based error, and human error in terms of predicting disease from available data. For this task, a "gold standard" dataset would consist of random patients whose true disease state was reconfirmed approaching both clinician and patient. The current study was not able to conduct this task, as the THIN database contains de-identified patient-level data, which is true for all large EMR databases that are used for research purposes. The THIN database also does not have data on surrogate markers that could improve the quality of the cohort identification algorithms. Miscoding between type 1 and type 2 diabetes in the primary care database is not uncommon [106,174]. It is important to mention that ML techniques may poorly distinguish between disease subtypes without incorporating additional classification rules. We have excluded patients with other diabetes Read codes from the dataset on which our ML algorithm was applied. Furthermore, for patients identified as T2DM without Read codes, the ML techniques are not able to provide an exact diagnosis date, therefore requiring incorporation of additional techniques.

3.6 CONCLUSION

Careful investigation of diagnostic codes patterns within the databases is essential before conducting analyses on the disease cohort. Direct extraction of a disease cohort using diagnostic codes may lead to the inclusion of falsely diagnosed patients and omitting patients with a true disease state. Rule-based techniques represent a conservative approach, which results in minimizing only false positive cases. ML techniques that minimize both false positives and false negatives cases represent a more robust approach. However, ML techniques heavily rely on meaningful input and use diagnostic codes for training purposes. Combining human expertise and machine power represents the best strategy that allows to test hypotheses on potential disease predictors, lower human interventions, and to reduce the burden of selection bias.

Chapter 4: The design of a comparative longitudinal case-control study and the imputation of missing longitudinal covariate data [1]

This chapter covers two methodological goals of this thesis and follows the identification of a T2DM cohort from a relational database in Chapter 3. A comparative longitudinal study design from which causal inferences can be drawn was developed. This is essential because significant differences between characteristics of cases and controls may be influenced by selection bias [175], which impacts on the ability to draw correct inferences about the internal and external validity of the study in question. With the study design of choice being a case-control study, the first goal of this chapter was to develop matched control subjects for the T2DM patients identified in Chapter 3. In Section 4.1, I demonstrate via algorithms, how to generate age- and sex-matched controls subjects for T2DM patients identified from a relational database. The matching method developed here has general applicability as it has been used on other projects unrelated to this thesis.

Given that missing data observed in the outcome of interest or other covariate data is common in longitudinal studies, the second goal of this chapter was to address the issue of missing longitudinal covariate data extracted for the cohort of T2DM patients (from Chapter 3) used in this thesis. In Section 4.2, exploratory data analysis was used to evaluate missing data patterns, and multiple imputations of missing covariate data was conducted. This section discusses and justifies the use of advanced multiple imputations techniques in this thesis.

[1] This chapter contains an exploratory analysis of missing data patterns and mechanisms within the cohort of T2DM patients used for this thesis. As a methodological chapter, results from the exploratory analyses were not published in any journal but illustrates how statistical methodologies can be generalised to EMR databases for robust inferences.

4.1 DEVELOPMENT OF CASE-CONTROL MATCHES WITHIN EMR DATABASES

This section describes the development and implementation of an algorithm for obtaining matched controls for patients with T2DM (cases) identified in Chapter 3.

4.1.1 Introduction

Matching as a means of establishing a similarity or comparability between groups (usually cases and control subjects) in observational studies can be performed either at the design stage or during the analysis stage. For this thesis, matching of controls to cases during the design stage is desired. For example, a 67-year old male Caucasian with T2DM should have a 67-year old male Caucasian without T2DM as his matched control, and so on. Given that there are half a million patients with diabetes in the THIN database (Chapter 3, Figure 3.1), the set of potential non-diabetic controls to be matched to each T2DM case are close to ~10 million patients. This large set of potential non-diabetic control subjects may guarantee at least one matched control for each case. Therefore, it was hypothesised that in a large EMR database like the THIN database, four control subjects could be exactly matched to a case on age, sex, and ethnicity. My objectives were to (1) develop an algorithm that will allow for 1: N case-control matching within EMRs and (2) apply the algorithm to obtain a matched case-control dataset for use in this thesis.

4.1.2 Methods

Algorithm for case-control matching in EMRs

Under the basic principle that once a match is made, it is never broken, 1: N matches can be obtained by building on the initial theory provided by Iacus and colleagues [176] as described mathematically below:

Let n be sample units that are subsets of a population of N units, where $n \leq N$.

Let N_1 and C_1 represent the total number of T2DM cases and potential controls respectively, with N_1 + C_1 = N.

Let X denote the set of dynamic matching covariates, where $\mathbf{X} = (\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3, \mathrm{X}_4, \cdots, \mathrm{X}_\mathrm{k})$, and $\mathbf{X}_\mathbf{j}$ is the subject-covariate dimension of observed values for variable j of n observations

$$\text{i.e., } \mathbf{X} = [\mathrm{X}_{ij}, i = 1,\ldots,\boldsymbol{n}, \boldsymbol{j} = 1,\ldots,\boldsymbol{k}].$$

Given a T2DM case, $i \in \mathbf{N_1}$ with its vector of covariates $\boldsymbol{X}_i$, the aim of matching is to discover a control unit $l \in \mathbf{C_1}$ with covariates $\boldsymbol{X}_l$ such that, the dissimilarity between $\boldsymbol{X}_i$ and $\boldsymbol{X}_l$ is very small in some metric (distance measure, d), that is d $(\boldsymbol{X}_i, \boldsymbol{X}_l) \approx 0$.

Now, under the assumption that "once a match is made, it is never broken", the following modification is made to allow for selection of more than one control for a case (Figure 4.1).

(1) $\boldsymbol{d_1}(\boldsymbol{X}_{i1}, \boldsymbol{X}_{l1}) \approx 0$ is the metric used to generate first the set of matches, $\mathbf{MC}_1$;

(2) Remove from $\mathbf{C}_1$**,** the matched controls generated by $\boldsymbol{d_1}$ to obtain a new set of potential controls $\boldsymbol{C_2}$ such that $\boldsymbol{C_2} < \boldsymbol{C_1}$;

(3) Generate a new set of matches $\mathbf{MC}_2$ from $\boldsymbol{C_2}$using the metric**,** $\boldsymbol{d_2}(\boldsymbol{X}_{i2}, \boldsymbol{X}_{l2})$;

(4) Repeat step (2) this time to obtain, $\boldsymbol{C_3}$ such that $\boldsymbol{C_3} < C_2$ and repeat step (3) to generate match set $\mathbf{MC}_3$ using $\boldsymbol{d_3}(\boldsymbol{X}_{i3}, \boldsymbol{X}_{l3})$;

(5) 1: N matches can be achieved by following the logic in steps [(2), (3), and (4)] until the desired matching ratio is obtained. Finally, the matched sets $\mathbf{MC}_1, \ldots, \mathbf{MC}_n$ can be pooled together.

Implementation

Cases (patients with T2DM) were defined as previously described in Chapter 3. A set of non-diabetic control patients (control pool) were obtained by selecting patients who had no diagnosis of T2DM or any other type of diabetes. Matching was then done on sex, year of birth, and ethnicity. The index date for controls was defined as the date of the diabetes diagnosis for their matched cases. Differences between patients with T2DM and their non-diabetic controls were evaluated using the rank sum test for continuous variables and the chi-squared /McNemar's test for binary data (presented in Table 4.1). Basic demographic characteristics of each ethnic group were summarized using the median and interquartile range for continuous variables and frequencies and percentages for categorical data. Scheffe's multiple comparison post hoc ANOVA test, a non-parametric Kruskal Wallis test, and the chi-square test were used to identify significant differences in different study parameters across the ethnic groups.

4.1.3 Results

The basic clinical characteristics of patients with T2DM with matched and unmatched patients are presented in Table 4.1. Before matching, the distribution of age at registration, sex, ethnicity and ex-smokers among patients with T2DM was significantly different from that of the non-diabetic controls. However, the significant difference between the characteristics of the patient with T2DM and their matched non-diabetic controls was no longer observed after matching. This shows the utility of matching techniques in improving the balance between the patients with T2DM and non-diabetic controls from the THIN database.

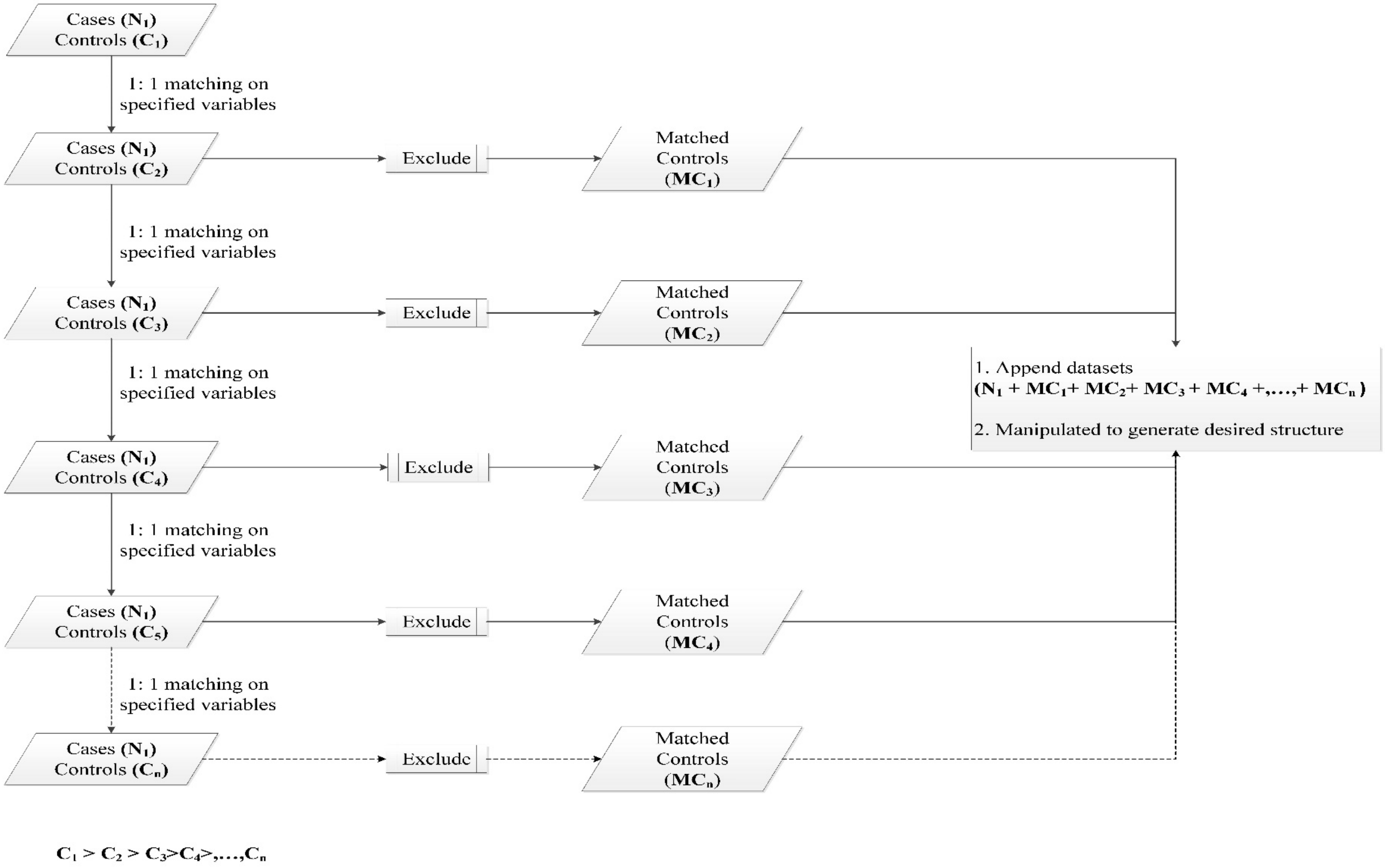

Cases (N_1)
Controls (C_1)
1: 1 matching on specified variables
Cases (N_1)
Controls (C_2)
Exclude
Matched Controls (MC_1)
1: 1 matching on specified variables
Cases (N_1)
Controls (C_3)
Exclude
Matched Controls (MC_2)
1. Append datasets
($N_1 + MC_1 + MC_2 + MC_3 + MC_4 +, \ldots, + MC_n$)
2. Manipulated to generate desired structure
1: 1 matching on specified variables
Cases (N_1)
Controls (C_4)
Exclude
Matched Controls (MC_3)
1: 1 matching on specified variables
Cases (N_1)
Controls (C_5)
Exclude
Matched Controls (MC_4)
1: 1 matching on specified variables
Cases (N_1)
Controls (C_n)
Exclude
Matched Controls (MC_n)
$C_1 > C_2 > C_3 > C_4 >, \ldots, C_n$

Figure 4.1: Illustration of the algorithm for obtaining matched controls for cases within a large EMR database

Table 4.1: Comparison of matching variables between patients with T2DM, matched and unmatched control patients

	T2DM (Cases)	**Unmatched Controls**	**Matched Controls**
Total patients, n	338,089	10,487,077	1,320,804
Age,			
at registration	49 (20)	27(22)	49 (20)
at diagnosis	60(14)	-	60 (14)
at end of data collection	72 (16)	44 (25)	72 (16)
Male,	185,745 (55)	5,020,437 (48)	742,980 (55)
Ethnicity,			
White	83,371 (25)	1,947,680 (19)	323,776(25)
Black	4,164 (1)	122,618 (1)	11,580 (1)
South Asians	7,337 (2)	165,085 (2)	15,880 (1)

Legend: $^{\alpha}$: Mean (SD); $^{\ddagger:}$ n (%);
Unmatched Controls: the original pool of non-diabetic control subjects;
Matched Controls: control subjects obtained from 1:4 matching using the modified algorithm.

4.1.4 Post matching processing

Demographic, anthropometric, clinical, and laboratory measurements along with the relevant dates of measurements were also extracted for the matched pairs (patients with T2DM and their matched controls). Notably, dates of exit of a patient from the database, generally through transfer to another practice/country or less commonly through death, previous medical history including any episode of cardiovascular events (myocardial infarction, stroke, coronary revascularisation, carotid or peripheral arterial revascularisation, or angina of cardiac origin), along with times of event were extracted. The longitudinal measures of anthropometric, clinical and laboratory parameters were arranged in 6-month windows (from 36 months pre-onset date to 84 months post onset date). The matched dataset generated from this process was used for the primary results included in Chapters 5 and 6.

4.2 MULTIPLE IMPUTATION OF MISSING LONGITUDINAL RISK FACTOR DATA

4.2.1 Introduction

Longitudinal studies (retrospective or prospective) are usually prone to missing or incomplete information which can impact the generalisability of study findings. Despite the availability of standard statistical techniques for addressing the problem of missing covariate data in longitudinal studies, few studies have provided sufficient information on the imputation methods used [106].

Several strategies for handling missing data are available and include methods such as complete case (CC) analyses, single imputation, and multiple imputation [109,177,178]. While CC analyses offer a straightforward approach, it ignores observations with missing data and in the process, loses information contained in the incomplete cases. Inferences from such analyses may not be generalizable to the population, particularly when the complete cases are small. Single imputation methods allow for substitution of missing data with a value. With information from complete cases, each missing value can be imputed with (1) the mean of the complete cases, (2) the mean conditional on observed values of other variables, (3) last observation carried forward, among others [109]. However, as pointed out by Rubin [109], single imputation does not account for variability of the predicted missing values, leading to bias in the resulting estimated variance of the parameter estimates.

Multiple imputation replaces each missing value with a set of plausible values that represent the uncertainty about the right value to impute. Multiple imputation does not attempt to estimate each missing value through simulated values. Instead, it draws a random sample of the missing values from its distribution. This process leads to valid statistical inferences that properly reflect the uncertainty due to missing values [109]. There is a large body of literature on the theoretical and methodological applications of various multiple imputation techniques. The choice of method depends on the assumption of missing data mechanism and underlying missing data pattern. This information is usually obtained via exploratory analysis to investigate missing data mechanism and patterns. Subsequently, under the assumptions that the missing data (1) mechanism is ignorable [i.e., missing at random (MAR) or missing completely at random (MCAR)], (2) are from a continuous multivariate distribution, and (3) can occur for any of the variables, missing values can be multiply imputed via regression methods or Bayesian-based Markov Chain Monte Carlo methods [109,177,178].

However, given the underlying nature of longitudinal studies, most multiple imputation methods are not adequately suited to account for the temporal order in which measurements are recorded [106]. This inherent problem is due to the assumption that variables in the imputation model follow a multivariate normal distribution. The full conditional specification (FCS) approach to multiple imputation has flexible properties that make it a good candidate for imputing longitudinal data as it does not depend on the assumption of multivariate normality [106,179]. This method fits a model for each variable with missing data (dependent) using all other variables as predictors, then iteratively imputes the missing values for the variable being fit [106,179]. With regards to longitudinally measured risk factor data collected on patients within a real-world primary care setting, there is an implementation of the FCS approach that is theoretically and pragmatically appealing. The FCS via predictive mean matching (PMM) imputes a value randomly from a set of observed values whose predicted values are closest to the predicted value from a specified regression model. Specifically, consider a series of variables $X_1, \dots, X_n$ where X_1 is missing and let $Z = (Z_1,\cdots, Zr)$ denote a set of variables with no missing data (fully observed). The variable X_1 is imputed by fitting a linear regression model of X_1 on Z. If the random draw of posterior predictive distribution of coefficients produced by the regression of X_1 on Z is denoted by $b*$, then a set of predicted values for X_1can be generated for both subjects with or without missing values on X_1. Using these predicted values, a set of subjects with predicted values close to the predicted value of subject with missing data are identified. Missing value is now substituted by assigning the predicted value of one of the close subjects that is randomly chosen. The whole process is repeated for the desired number of imputations.

For this thesis, longitudinal data on weight, BMI, SBP, and HbA_{1c} were extracted for the cohort of patients with T2DM identified in Chapter 3 and arranged in 6-monthly non-overlapping windows. However, as with all EMRs, a non-trivial amount of missing data exists for the longitudinally collected covariate data. Therefore, my objectives were to (1) compare uncertainty around inferences obtained from data imputed by PMM with complete data and (2) evaluate if the inference related to the association of BMI, SBP, and HbA_{1c} with the risk of death differed between complete and imputed data sets.

4.2.2 Methods

Data structure

To demonstrate the feasibility of multiple imputation approaches with this thesis, longitudinal data on weight, BMI, SBP, and HbA_{1c} obtained at diagnosis, 6, 12, 18, and 24 months post diagnosis were used. A generalised data structure that represents all possible scenarios under which follow-up data was recorded for patients with T2DM was postulated (Figure 4.2). As with common data structures, the first row represents variable names where “Dx” captures measurements recorded at diagnosis.

Follow-up data captured at 6, 12, 18, and 24 months post-diagnosis are represented by *“Dx+6”*, *“Dx+12”*, *“Dx+18”*, and *“Dx+24”* respectively. A value of 1 is assigned to the event of a recorded measurement for a patient and 0 for a missing measurement. Given that there are five variables representing measurements taken in 6 monthly windows over 2 years, patients can have missing values for:

i. All five variables [0/5 recorded],

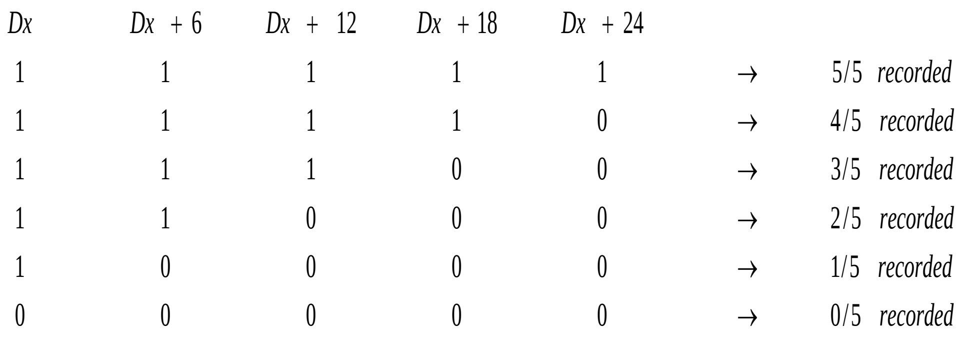
ii. Four variables [1/5 recorded]

iii. Three variables [2/5 recorded],

iv. Two variables [3/5 recorded],

v. One variable [4/5 recorded], and

vi. None of the five variables [5/5 recorded].



Dx	*Dx* + 6	*Dx* + 12	*Dx* + 18	*Dx* + 24		
1	1	1	1	1	→	5/5 *recorded*
1	1	1	1	0	→	4/5 *recorded*
1	1	1	0	0	→	3/5 *recorded*
1	1	0	0	0	→	2/5 *recorded*
1	0	0	0	0	→	1/5 *recorded*
0	0	0	0	0	→	0/5 *recorded*

Figure 4.2: Data structure representing all scenarios of recording longitudinal follow-up data. [*Dx:* At diagnosis; *Dx+6:* at 6 months post diagnosis*; Dx+12:* at 12 months post diagnosis; *Dx+18:* at 18 months post diagnosis*; Dx+24:* at 12 months post diagnosis]

Multiple imputation and data analyses

Among patients with a minimum of 2 years of follow-up post-diagnosis, the proportion of patients who had at least two 6-monthly longitudinal measures of weight, BMI, SBP, and HbA$_{1c}$ were calculated. The missing 6-monthly longitudinal measures of these risk factors were imputed separately, only if at least two measures of longitudinal body weight were available for an individual patient. The multiple imputation technique used was PMM, conditioning on the age at diagnosis, sex, smoking status, deprivation status, the usages of ADDs and other relevant drugs as appropriate. Twenty imputations were conducted and the results were pooled together according to Rubin’s rules [109]. For each time point during the 24 months, mean (95% CI) of imputed and complete data were estimated and plotted separately for each of the risk factors considered. Furthermore, a multivariate

Cox regression model was fitted to imputed and complete data to assess the relationship between risk factors measured at diagnosis and all-cause mortality. This model adjusted for age at diagnosis, sex, ethnicity, smoking status, deprivation status, use of ADDs, and use of cardioprotective medications. The hazard ratios and their 95% CI obtained were used to check the consistency of clinical inference.

4.2.3 Results

The proportion of patients with missing 6-monthly longitudinal measurements of weight, BMI, HbA_{1c} and SBP within two years of T2DM diagnosis are presented in Table 4.2. Among patients with a minimum 2 years of follow-up, the proportions patients who had at least 2 of the 5 measures missing was 16 % for weight, BMI, and SBP and 11% for HbA_{1c}.

A comparison of the mean (95% CI) of weight, BMI, SBP, and HbA_{1c} for the complete and imputed datasets are presented in Figure 4.4. The average (95% CI) weight, BMI, SBP and HbA_{1c} at diagnosis was 88.7 (88.6, 88.8), 31.4 (31.4, 31.5), 141.3 (141.2, 141.4), and 8.3 (8.3, 8.3) respectively. The distributions of imputed weight, BMI, and SBP over 24 months post diagnosis of T2DM were similar longitudinally compared to the complete data. However, imputed values of these three risk factors at 6 months post diagnosis was lower than to the respective values in the complete data. Also, only imputed values of HbA_{1c} at diagnosis and 6 months post diagnosis had similar distribution compared to their respective values in the complete data. Estimates of imputed HbA_{1c} were marginally lower from 12 to 24 months post diagnosis compared to the complete data. This clearly suggests that multiple imputation by PMM captured the true longitudinal distributions of the weight, BMI, and SBP but not HbA_{1c} (Figure 4.3).

When risk factors were considered as continuous measures in the assessment of mortality risk for the two imputation methods (Table 4.3), the confidence intervals of the risk estimates overlapped suggesting no statistically significant difference between the two methods. Also, when the continuous measures were converted to categorical variables [i.e. BMI categories (normal weight, overweight, and obese), SBP categories (< 140mmHg and ≥140mmHg), and HbA_{1c} groups (≤ 7%, 7.1-8.0%, 8.1-9.0%, and ≥ 9%)], there was agreement between the imputed data and the complete data.

Table 4.2: Proportion of missing 6-monthly longitudinal measurements of weight, BMI, HbA_{1c} and SBP within two years of T2DM diagnosis

Proportion of 6-monthly longitudinal measurements missing	n (%)			
	Weight, kg	BMI, kg/m^2	HbA_{1c}, %	SBP, mmHg
At least 2 of 5 measurements missing	62,126 (16)	62,126 (16)	41,653 (11)	58,977 (16)
At least 3 of 5 measurements missing	60,507 (16)	60,507 (16)	39,674 (10)	45,079 (12)
At least 4 of 5 measurements missing	53,123 (14)	53,123 (14)	40,868 (11)	44,283 (12)

Table 4.3: Hazard ratios (HR) and 95% CI for all-cause mortality for from complete and imputed data

	HR (95% CI) for ACM	
	Complete data	PMM
BMI at diagnosis	1.03 (1.02,1.03)	1.01 (1.00,1.03)
Normal weight	1.62(1.43, 1.84)	1.26 (1.11,1.44)
Overweight	1.13 (1.03,1.23)	0.98 (0.90,1.06)
Obese	Reference	Reference
SBP at diagnosis	1.00 (1.00,1.01)	1.00 (1.00,1.01)
≥140 mmHg	1.12(1.05,1.19)	1.12 (1.05,1.19)
<140 mmHg	Reference	Reference
HbA_{1c} at diagnosis	1.06 (1.04,1.08)	1.05 (1.05,1.07)
≤ 7.0 %	Reference	Reference
7.1-8.0 %	1.14 (1.01,1.29)	1.16 (1.06,1.29)
8.1-9.0 %	1.35 (1.18,1.55)	1.36 (1.23,1.50)
≥ 9.1 %	1.44(1.29,1.61)	1.36 (1.23,1,51)

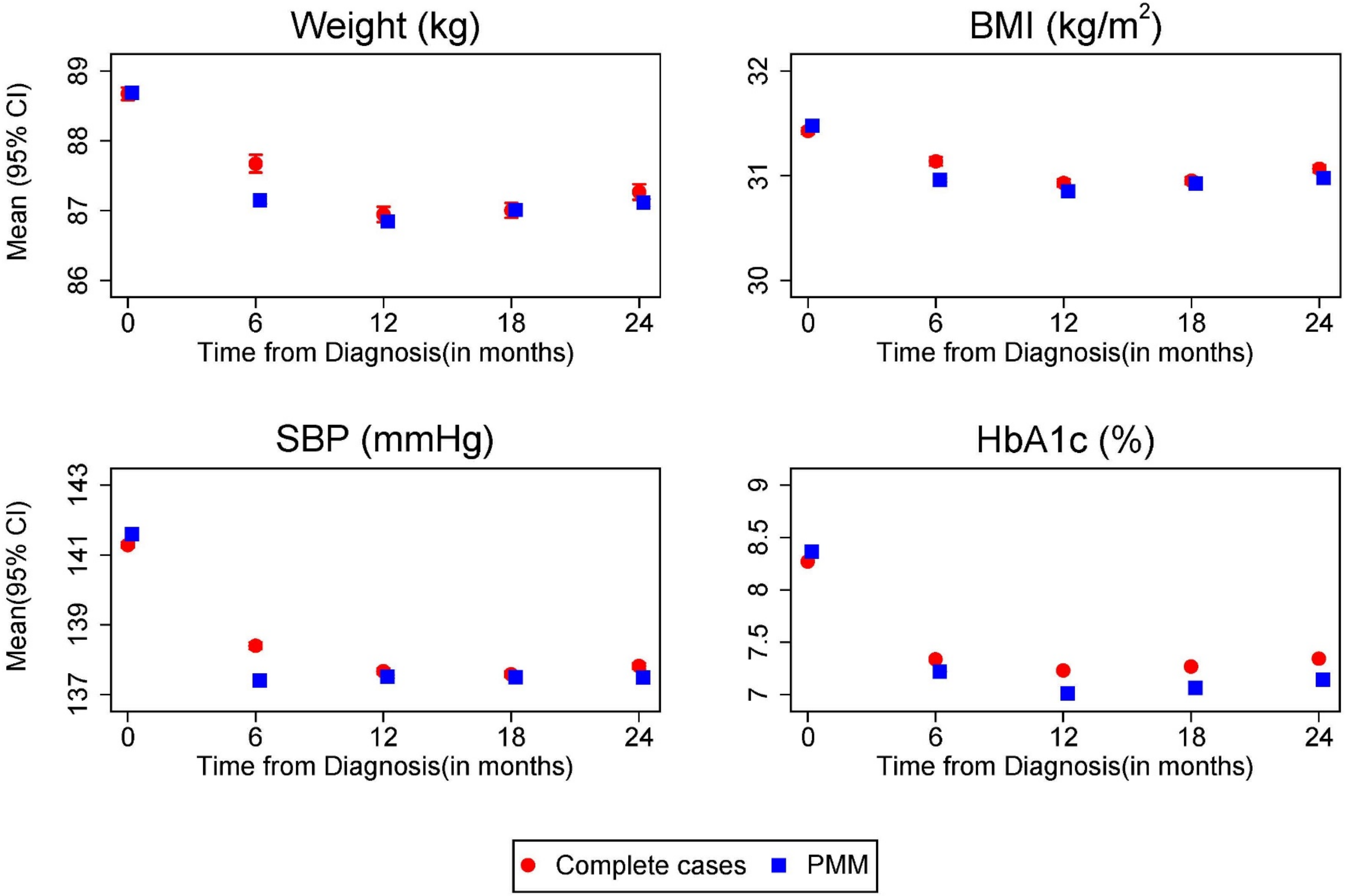


Figure 4.3: Comparison of original (complete) with imputed (PMM) weight, BMI, SBP, and HbA_{1c} from diagnosis of T2DM to 24 months post-diagnosis.

4.2.4 Discussion

This exploratory section evaluated the frequency and patterns of missingness in longitudinal clinical data collected on patients with T2DM from the time of diagnosis over 2 years. Multiple imputation of missing 6-monthly longitudinal clinical data (weight, BMI, SBP, and HbA_{1c}) was done using PMM. The uncertainty around imputed and complete (unimputed) data was compared and differences in the association of these risk factors with the risk of death between imputed and complete datasets were evaluated. It was observed that multiple imputation via PMM (1) captured the true longitudinal distribution of weight, BMI, and SBP over 24 months post diagnosis, (2) estimated similar HbA_{1c} values at diagnosis and 6 months post-diagnosis and marginally lower HbA_{1c} values from 12 months to 24 months post diagnosis when compared to complete data, and (3) leads to similar clinical inferences between complete data (CC) and imputed data based on analyses drawn on these risk factors.

Many patients had at least 2 missing 6-monthly longitudinal measurements over 24 months post-diagnosis. Primary care based EMR databases present a formidable challenge because “missing data” have an intermittent pattern of missingness over time (non-monotone) and are NMAR, so approaches such as CC analyses produces biased and statistically inefficient results[180]. Also, the prediction of unknown missing values from a set of known values can be biased as seen in the case of single imputation methods [65]. Multiple imputation approaches replace each missing value with a set of plausible values that represent the uncertainty about the right value to impute [109]. The PMM technique uses this underlying principle and imputes a value randomly from a set of observed values whose predicted values are closest to the predicted value from a specified regression model. The observation of similar longitudinal distribution of risk factors from the imputed data and complete data shows the robustness of using multiple imputation for making clinical inferences in this thesis.

The clinical contexts of evaluating the association of BMI at diagnosis with long-term cardiovascular and mortality risk, using continuous measures of risk factors or clinical categorisation of these risk factors, were well supported with confidence in making robust inferences using PMM method of imputation in the current thesis. More importantly, the observation that inferences about the association of the risk factors under consideration with mortality risk were similar between complete and imputed data ensures that the use of the imputed data in the current thesis will provide reliable results.

Chapter 5: Comparison of body mass index at diagnosis of diabetes in a multi-ethnic population: A case-control study with matched non-diabetic controls

This body of this chapter contains one published paper that discusses the distribution of BMI at diagnosis of diabetes in comparison to non-diabetic controls. The citation of the published paper is as follows:

Paul SK*, **Owusu Adjah ES***, Samanta M, Patel K, Bellary S, Hanif W, Khunti K. Comparison of body mass index at diagnosis of diabetes in a multi-ethnic population: A case-control study with matched non-diabetic controls. *Diabetes, Obesity and Metabolism* 2017;**19**(7):1014-1023**.** * Joint first authors

All the listed have agreed to the inclusion of this published scholarly work in this thesis and the statement of my contribution to the authorship of this published scholarly work is included below:

Contributor	Statement of contribution
Paul Sanjoy K.	Conceived the idea and was responsible for the primary design of the study. Contributed to the statistical analyses. Edited the first draft and contributed towards finalisation of the manuscript.
Owusu Adjah Ebenezer S. (Candidate)	Conceived the idea and was responsible for the primary design of the study. Responsible for extracting data from THIN database. Developed and applied exact case-control matching algorithm for matching of diabetes patients with non-diabetic controls within the THIN database. Responsible for data manipulation, aggregation, transformation in SAS and contributed towards the statistical analyses in STATA. Contributed towards the interpretation of results. Developed first draft and contributed towards finalisation of the manuscript.
Samanta Mauykh	Contributed to the interpretation of the results and manuscript finalisation.
Patel Kiran	Contributed to the interpretation of the results and manuscript finalisation.
Bellary Srikanth	Contributed to the interpretation of the results and manuscript finalisation.
Hanif Wasim	Contributed to the interpretation of the results and manuscript finalisation.
Khunti Khunti	Contributed to the interpretation of the results and manuscript finalisation.

5.1 ABSTRACT

Aims: To investigate the probability of developing type 2 diabetes mellitus (T2DM) at different body mass index compared to matched non-diabetic controls in a multi-ethnic population.

Materials and Methods: Case-control study of 90,367 patients with incident diabetes and 362,548 age-sex-ethnicity matched controls from UK primary care. The probability of developing T2DM was estimated.

Results: Case and control patients were 56 years old at index and 56% were male. Patients with T2DM had significantly higher mean BMI level by about 5 kg/m^2 at diagnosis (32.2 kg/m^2), compared to the matched controls (27.4 kg/m^2). White European (n=79,270), African-Caribbean (n=4,115) and South Asians (n=7,252) were 58, 48, and 46 years old with mean BMI of 32.5, 31.1, 29.2 kg/m^2 respectively at diagnosis. More South Asians developed T2DM at BMI below 30 kg/m^2 (38%) than White Europeans (26%) and African-Caribbeans (29%), (all $p<0.01$). Within the 18-70-year age range, South Asian males and females had significantly higher probability of developing diabetes in the continuously measured BMI range of 18-30 kg/m^2, compared to White Europeans and African-Caribbeans. Across all age groups < 70 years, South Asians and African-Caribbeans had significantly higher probability of developing T2DM in the normal weight and overweight categories, compared to White Europeans. However, this risk patterns of developing diabetes was reversed amongst the obese at all age groups.

Conclusion: Risk patterns of developing diabetes at different levels of obesity varies between ethnic groups across all age groups, while South Asians and African-Caribbeans carry the highest risk at younger age and at lower adiposity burden.

5.2 INTRODUCTION

Obesity [body mass index (BMI) ≥ 30 kg/m^2] is a worldwide epidemic affecting people of all ages and is a major risk factor for type 2 diabetes mellitus (T2DM) and cardiovascular diseases (CVD) [69]. Current epidemiological indices of obesity have doubled since 1980, with about 13% of the world's population being obese as of 2014 [69,181]. Some population-based studies have been conducted to assess the impact of BMI classification, including overweight and various grades of obesity, on the risk of T2DM [30,182,183]. The BMI cut-points of 25 kg/m^2 and 30 kg/m^2 were defined as the basis for identifying overweight and obesity based on epidemiological studies investigating the association with mortality and morbidity, primarily in White population. However, there is increasing evidence that levels of risk associated with the classification of overweight and obesity vary across ethnic groups [28,104,184-187].

The propensity to develop T2DM varies considerably between ethnic groups and identifying the BMI cut points within specific ethnic groups at which the risk of T2DM increases is useful to inform public health policy. Some studies have evaluated the ethnicity-specific diabetes incidence rates in association with prior BMI levels using population level and primary care data [28,29,185,188]. These studies were limited by the number of subjects in the non-white ethnic groups. The pooled analysis of survey data from various countries conducted by the DECODE-DECODA study group in 2003 evaluated the association of ethnicity, BMI and prevalence of T2DM [188]. However, the BMI measurements were not consistently taken at the time of diagnosis of diabetes. Only one previous study has compared the distribution of BMI at diagnosis of T2DM with non-diabetic controls [30]. Ganz and colleagues defined BMI at diagnosis as the last measurement of BMI taken within one year prior to diagnosis of T2DM, and randomly matched controls to cases [30]. While this study reported an increased risk of developing T2DM with higher BMI levels, the differential aspects of ethnicity in the relationship between BMI and risk were not addressed.

No study has compared the distribution of BMI at diagnosis of T2DM by ethnicity with non-diabetic controls. In addition, we are not aware of any population-based study evaluating differences in the risk of developing T2DM in men and women in different age levels between different ethnic groups over the whole spectrum of BMI distribution at diagnosis. Using a large cohort of incident T2DM patients, and an age-sex-ethnicity matched non-diabetic control cohort from United Kingdom primary care, the aims of this study were to evaluate for each ethnic group (1) the distribution of BMI, glycaemic, and vascular risk factors at diagnosis of T2DM, and (2) the probability of developing T2DM over the entire spectrum of BMI and age.

5.3 MATERIALS AND METHODS

5.3.1 Data source

Data for this study were obtained from The Health Improvement Network (THIN) database, a large anonymised longitudinal dataset derived from a network of more than 600 primary care providers across the United Kingdom. With longitudinal data on approximately 11 million individuals registered with the primary care system, the THIN database has been extensively used for academic research in various disciplines [122]. The accuracy and completeness of this database has been previously described [124,125]. Notably, the database has a similar distribution of major chronic diseases including diabetes, heart failure and obesity when compared to UK national statistics [87,124]. Clinically diagnosed diseases are recorded using Read codes [128] and with each diagnosis, an event date is entered. THIN database provides comprehensive patient-level longitudinal information on demographic, anthropometric, clinical and laboratory measures, clinical diagnosis of diseases/events, along with complete information on prescriptions for medications with dates and doses. Formal access to the database has been obtained and the study protocol approved by the Scientific Review Committee of the THIN database, UK (reference number: 15THIN030).

5.3.2 Identification of T2DM cases

Patients with T2DM were identified through various steps of clinically guided iterative processes. Specifically, the T2DM cases were selected if:

(i) Patient had a record of Read code related to T2DM,

(ii) Patient from step (i) above had received at least one prescription for an antidiabetic drug in addition to the clinical diagnosis, or

(iii) Patient in step (i) above had received a lifestyle modification intervention.

A set of 345,013 patients with newly diagnosed T2DM (from January 1990 to September 2014) was identified, who had complete information on age at diagnosis (≥18 years) and sex. Of these patients, only 90,754 patients had their ethnicity identified as White European, African-Caribbean or South Asian, (Figure 6.1). South Asians were defined as patients with Indian, Pakistan, Sri Lanka, and Bangladesh origin while African-Caribbeans were defined as patients with Black-African and/or Caribbean origin. White Europeans were patients with self-reported ethnicity as White, European, Caucasian, and/or New Zealand European.

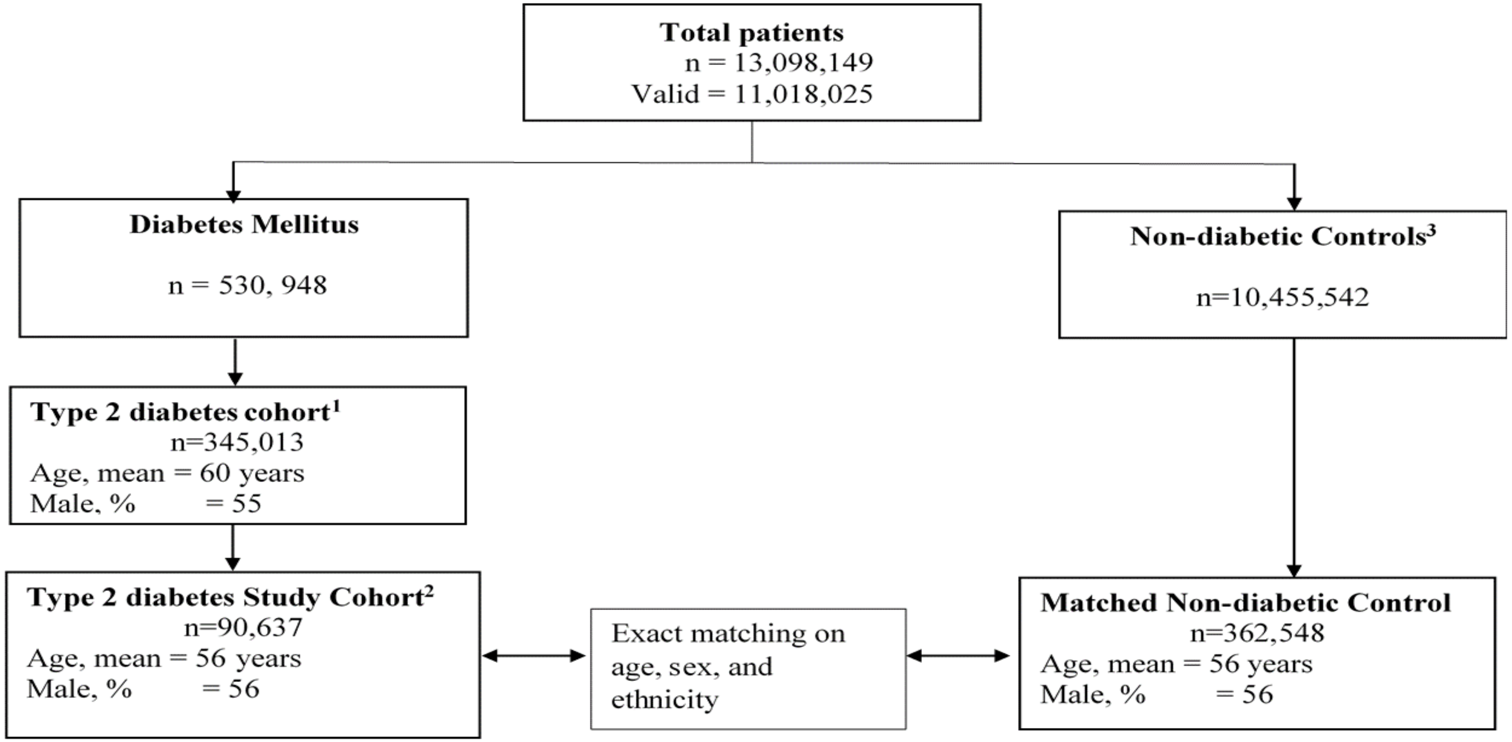


Figure 5.1: The identification of T2DM study cohort and their matched controls from THIN database.

[[1]: T2DM Read code + (antidiabetic medication or lifestyle modification intervention, excludes patients with any other type of diabetes (e.g. Type 1 diabetes, Gestational diabetes); [2]: Age at index date greater or equal to 18, complete information on gender and ethnicity (White European, African-Caribbean and South Asian only]; [3] Excludes patients who have ever received anti-hyperglycaemic drugs.

5.3.3 Development of control subjects

A control pool of patients without T2DM was obtained by selecting individuals who had no diagnosis of any type of diabetes and had never received an antidiabetic prescription. Exact matches based on ethnicity, age, and sex were obtained from this pool of potential controls without replacement. To the 90,637 eligible T2DM cases with identified three ethnic groups, 362,548 controls were successfully matched in a 1:4 ratio. The index date for controls was defined as the date of the diabetes diagnosis for their matched cases.

The following information on index date was extracted for all patients where available: smoking status, deprivation score, weight, BMI, glycated haemoglobin (HbA_{1c}), systolic blood pressure (SBP), diastolic blood pressure (DBP), low density lipoproteins (LDL-C), high density lipoproteins (HDL-C), and triglycerides. All available measures on or within 3 months prior to the index date were considered as the baseline measures. Anti-glycaemic agents, anti-hypertensive agents, cardio-protective medications (CPM), weight lowering drugs and anti-depressants were also obtained along with dates of prescription. The CPMs were defined as the use of statins or angiotensin-converting enzyme inhibitors or angiotensin II receptor blockers or beta blockers on or before diagnosis. BMI categories were defined following WHO established criteria [5] as follows: normal weight (18.5-24.99 kg/m^2), overweight (25-29.99 kg/m^2), Grade 1 obese (30-34.99 kg/m^2), Grade 2 obese (35-39.99 kg/m^2) and Grade 3 obese (≥ 40 kg/m^2). In addition, records of cardiovascular diseases (CVDs), renal diseases and cancer on or before the index date were also obtained. A composite variable for CVD (any CVD) was defined as the occurrence of angina or myocardial infarction or coronary artery disease (including bypass surgery and angioplasty) or heart failure or stroke before diagnosis.

5.3.4 Statistical analysis

Basic characteristics of incident T2DM patients and their matched non-diabetic control population, separately for ethnic group, were summarized using number (%), means ± SD or median (first quartile, third quartile) as appropriate. Differences between patients with T2DM and their matched controls were evaluated using the rank sum test for continuous variables and the chi-squared for binary data. The distributions of BMI at diagnosis of T2DM in three different ethnic groups were compared using analysis of variance models, separately for different age groups at diagnosis.

To explore the association of BMI at diagnosis with the risk of developing T2DM, in interaction with different ethnic groups, multivariate logistic regression models were fitted. The covariates for adjustments were age, sex, smoking status, deprivation score, and the history of CVD, cancer, and chronic kidney disease (CKD) on or prior to the index date. The probability of developing T2DM over the whole distribution of BMI in different ethnic groups was evaluated, using both continuous measures of BMI and the World Health Organisation defined categories of BMI. To explore the possible differences in the patterns of association of BMI with the risk of developing T2DM in different ethnic groups for male and female, and also over different age groups at index date, separate adjusted models were fitted. The differences in predicted probabilities between ethnic groups were calculated using the methodology described by King and colleagues (2000) [189] and Zelner (2009) [190]. The estimated probabilities and the 95% confidence intervals were presented as appropriate. Sensitivity analyses to support the above analyses include (1) an extended model incorporating measures of SBP, LDL-C, HDL-C and triglyceride at index, the use of CPMs, weight lowering drugs, anti-hypertensives, and anti-depressants before diagnosis, and (2) comparison of distribution of BMI and HbA1c at diagnosis for patients diagnosed after 01 Jan 2006.

5.4 RESULTS

The basic demographic and clinical profiles of 90,367 patients with T2DM and 362,548 age-sex-ethnicity matched controls, separately for ethnic groups, are shown in Table 5.1. Case and control patients were 56 years old at index and 56% were male. Patients with T2DM had significantly higher mean BMI level by about 5 kg/m^2 at diagnosis (32.2 kg/m^2), compared to the matched controls (27.4 kg/m^2). However, this difference was smaller in the African-Caribbean (2.4 kg/m^2) and South Asians (2.8 kg/m^2). Furthermore, cases were more likely to receive anti-hypertensives, cardio-protective medications, and anti-depressants and were more likely to have any CVD before diagnoses than controls. South Asians were more likely to develop diabetes at a significantly lower age (mean age 46 years, 31% below 40 years) compared to the White Europeans (mean age 58 years, 9% below 40 years) and African-Caribbeans (mean age 48 years, 23% below 40 years). Compared to male patients, females developed T2DM at a significantly higher BMI level across all ethnic groups. More South Asians developed T2DM at BMI below 30 kg/m^2 (38%) than White Europeans (26%) and African-Caribbeans (29%), (all $p<0.01$). Those not included in the study because of non-availability of ethnicity data were older (mean age 61 years compared to 56 years in the study cohort) but had a similar distribution of sex, BMI, current smokers, ex-smokers, and never smokers.

The average SBP in South Asians at the time of diagnosis of T2DM (132 mmHg) was significantly lower with only 20% having SBP ≥ 140 mmHg, compared to the two other ethnic groups ($p<0.01$). Among those who developed T2DM, only 28% of South Asians were current or ex-smokers at index date, compared to 60% and 33% in the White and African-Caribbean ethnic groups respectively. African-Caribbean and South Asians had significantly higher LDL-C levels at diagnosis of T2DM (LDL-C ≥ 100 mg/dl: 34% and 28% respectively) compared to the White Europeans (LDL-C ≥ 100 mg/dl: 23%, Table 5.1). African-Caribbean patients had significantly higher mean HbA$_{1c}$ level at diagnosis [9.1% (76 mmol/mol), 30% with HbA$_{1c}$ ≥ 7.5 % (58 mmol/mol)] compared to South Asians [8.5% (69mmol/mol), 28% with HbA$_{1c}$ ≥ 7.5% (58 mmol/mol)] and White Europeans [8.2% (66 mmol/mol), 22% with HbA$_{1c}$ ≥ 7.5% (58 mmol/mol)].

The distributions of BMI at diagnosis of T2DM in different ethnic groups, by age groups at diagnosis, are presented in Table 5.2. South Asians and African-Caribbeans aged 18-70 years at diagnosis developed T2DM at significantly lower BMI than White Europeans. Over the whole distribution of BMI level, the probability of developing T2DM in South Asians compared with other ethnic groups, separately for male and female and by different age groups, are presented in Figure 1 and Figure 2 respectively. When analysed with a continuous measure of BMI, compared to both White Europeans and African-Caribbeans, South Asians had a significantly higher probability of developing T2DM within the range of BMI from 18 kg/m^2 to about 30 kg/m^2, for both males and females (Figure 7.1 A-D). The adjusted probability (95% CI) of developing T2DM at different BMI category levels compared between the three ethnic groups, separately for different age groups, are presented in Figure 7.2. Across all age groups within the age of 70 years, South Asians and African-Caribbeans had a significantly higher probability of developing T2DM in the normal weight and overweight categories, compared to White Europeans. Sensitivity analysis with an extended list of covariates revealed similar results.

Table 5.1: Distribution of basic characteristics of T2DM patients and their matched controls, stratified by ethnicity

	White European (N=396,350)		**African-Caribbean (N=20,575)**		**South Asian (N=36,260)**		**ALL (N=453,770)**	
	T2DM	Control	T2DM	Control	T2DM	Control	T2DM	Control
Patients*	79,270	317,080	4,115	16,460	7,252	29,008	90,637	362,548
Age at diagnosis (years) †	58 ± 12	58 ± 12	48 ± 12	48 ±12	46 ± 12	46 ± 12	56 ± 13	56 ± 13
Age at diagnosis (years) ‡	58 (49, 67)	58 (49, 67)	48 (40, 56)	48 (40, 56)	45 (38, 54)	45 (38, 54)	57 (47, 66)	57 (47, 66)
Age group *								
≤40	6,724 (9)	26,896 (9)	949 (23)	3,796 (23)	2,204 (30)	8,816 (30)	9,877 (11)	39,508 (11)
41-50	15,614 (20)	62,456 (20)	1,511 (37)	6,044 (37)	2,598 (36)	10,392 (36)	19,723 (22)	78,892 (22)
51-60	22,425 (28)	89,700 (28)	1,007 (25)	4,028 (25)	1,529 (21)	6,116 (21)	24,961 (28)	99,844 (28)
61-70	21,751 (27)	87,004 (27)	491 (12)	1,964 (12)	688 (10)	2,752 (10)	22,930 (25)	91,720 (25)
71+	12,756 (16)	51,024 (16)	157 (4)	628 (4)	233 (3)	932 (3)	13,146 (15)	52,584 (15)
Male *	44,651 (56)	178,604 (56)	2,102 (51)	8,408 (51)	4,005 (55)	16,020 (55)	50,758 (56)	203,032 (56)
Current smokers *	15,581 (20)	59,060 (19)	527 (13)	2,307 (14)	967 (13)	3,460 (12)	17,075 (19)	64,827 (18)
Ex-smokers *	31,966 (40)	114,099 (36)	808 (20)	2,612 (16)	1,058 (15)	3,497 (12)	33,832 (37)	120,208 (33)
Never smokers *	31,427 (40)	138,903 (44)	2,769 (67)	11,130 (68)	5,195 (72)	21,162 (73)	39,391 (44)	171,195 (47)
Highest affluence *	3,391 (4)	15,054 (5)	564 (14)	1,954 (12)	665 (9)	2,525 (9)	4,620 (5)	19,533 (5)
Lowest affluence *	16,923 (21)	56,124 (18)	994 (24)	4,290 (26)	1,852 (26)	7,073 (24)	19,769 (22)	67,487 (19)
HbA$_{1c}$ (%) ,[mmol/mol] §	8.2 ± 2.1 [66 ± 23.0]		9.1 ± 2.7 [76 ± 29.5]		8.5 ± 2.1 [69 ± 23.0]		8.3 ± 2.1 [67 ± 23.0]	
HbA$_{1c}$ ≥ 7.5%*§	17,730 (22)		1,234 (30)		2,021 (28)		20,985 (23)	
Weight (kg) †	92.2 ± 21.0	77.8 ± 16.8	88.3 ± 18.7	81.0 ± 16.2	78.8 ± 17.1	71.3 ± 14.6	91.0 ± 21.0)	77.5 ± 16.8
Weight (kg) ‡	90.0 (78, 104)	76.2 (66, 88)	86.0 (75, 99)	79.7 (70, 90)	76.0 (67, 88)	70.0 (61, 80)	88.9 (76, 103)	76.0 (66, 87)
BMI (kg/m^2) † -- All	32.5 ± 6.8	27.4 ± 5.2	31.1 ± 6.2	28.7 ± 5.6	29.2 ± 5.7	26.5 ± 4.8	32.2 ± 6.8	27.4 ± 5.2
BMI (kg/m^2) † -- Male	31.7 ± 6.0	27.5 ± 4.7	29.3 ± 5.3	27.3 ± 4.6	28.4 ± 5.4	26.2 ± 4.6	31.4 ± 6.0	27.4 ± 4.7
BMI (kg/m^2) † -- Female	33.4 ± 8.0	27.4 ± 5.8	33.0 ± 6.6	30.0 ± 6.2	30.2 ± 5.8	26.6 ± 5.1	31.5 ± 6.8	27.0 ± 5.3
BMI (kg/m^2) ‡	31.4 (28, 36)	26.8 (24, 30)	30.2 (27, 35)	28.0 (25, 32)	28.3 (25, 32)	26.0 (23, 29)	31.1 (28, 36)	26.8 (24, 30)
Normal weight*	4,958 (6)	24,052 (8)	365 (9)	834 (5)	921 (13)	2,374 (8)	6,244 (7)	27,260 (8)
Overweight*	15,439 (20)	30,135 (10)	820 (20)	1,283 (8)	1,830 (25)	2,376 (8)	18,089 (20)	33,794 (9)
Grade 1 obese*	15,592 (20)	13,654 (4)	719 (18)	764 (5)	1,050 (15)	896 (3)	17,361 (19)	15,314 (4)
Grade 2 obese*	8,973 (11)	4,073 (1)	383 (9)	288 (2)	386 (5)	251 (1)	9,742 (11)	4,612 (1)
SBP (mmHg) †	141 ± 19	136 ± 19	137 ± 19	133 ± 19	132 ± 18	128 ± 17	140 ± 19	135 ± 18
SBP ≥ 140 mmHg *	28,971 (37)	52,256 (17)	1,104 (27)	1,742 (11)	1,461 (20)	2,092 (7)	31,536 (35)	56,090 (16)
DBP (mmHg) †	82 ± 11	80 ± 10	83 ± 11	81 ± 11	82 ± 11	79 ± 10	82 ± 11	80 ± 10

LDL-C(mg/dl) [†]	117 ± 42	120 ± 40	126 ± 39	124 ± 35	119 ± 39	121 ± 35	117 ± 42	120 ± 39
LDL-C ≥ 100 mg/dl*	17,944 (23)	25,328 (8)	1,388 (34)	1,531 (9)	2,003 (28)	2,684 (9)	21,335 (24)	29,543 (8)
HDL-C (mg/dl)[†]	46 ± 14	56 ±17	48 ± 13	58 ± 17	44 ± 11	51 ± 14	46 ± 13	56 ± 17
HDL-C ≤ 45 mg/dl*	19,024 (24)	11,941 (4)	924 (23)	501 (3)	2,041 (28)	1,452 (5)	21,989 (24)	13,894 (4)
Triglycerides (mg/dl) [‡]	159 (122, 213)	115 (87, 159)	115 (81, 159)	82 (62, 115)	151 (115, 204)	115 (89, 168)	159 (115, 213)	115 (84, 159)
Triglyceride ≥ 150 mg/dl*	18,654 (24)	13,323 (4)	601 (15)	260 (2)	1,681 (23)	1,316 (5)	20,936 (23)	14,899 (4)
Complications*								
CKD (≥ stage 3)	1,752 (2)	4,493 (1)	52 (1)	146 (1)	34 (1)	143 (1)	1,838 (2)	4,782 (1)
Cancer	4,746 (5)	18,793 (6)	100 (2)	278 (2)	68 (1)	339 (1)	4,914 (5)	19,410 (5)
Myocardial Infarction	4,802 (6)	9,120 (3)	34 (1)	85 (1)	181 (3)	355 (1)	5,017 (6)	9,560 (3)
Heart Failure	1,743 (2)	2,635 (1)	29 (1)	38 (0)	41 (1)	76 (<0.1)	1,813 (2)	2,749 (1)
Angina	6,529 (8)	12,975 (4)	48 (1)	93 (1)	196 (3)	440 (2)	6,773 (8)	13,508 (4)
Stroke	4,014 (5)	9,521 (3)	105 (3)	198 (1)	111 (2)	275 (1)	4,230 (5)	9,994 (3)
Any CVD	15,769 (20)	33,155 (11)	225 (6)	443 (3)	519 (7)	1,101 (4)	16,513 (18)	34,699 (10)
Hypertension	33,234 (42)	62,749 (20)	1,419 (35)	2,852 (17)	1,803 (25)	2,941 (10)	36,456 (40)	68,542 (19)
Anti-hyperglycaemic drugs (Ever prescribed)*								
None	10,767 (14)	317,080 (100)	294 (7)	16,460 (100)	536 (7)	29,008 (100)	11,597 (13)	362,548(100)
Insulin	17,693 (22)		917 (22)		1,381 (19)		19,991 (22)	
Biguanides	62,598 (79)		3,530 (86)		6,324 (87)		72,452 (80)	
Sulphonylureas	38,281 (48)		2,090 (51)		3,739 (52)		44,110 (49)	
Thiazolidinedione	14,481 (18)		622 (15)		1,411 (20)		16,514 (18)	
GLP1-RA	3,769 (5)		117 (3)		225 (3)		4,111 (5)	
DPP-4	11,078 (14)		633 (15)		1,231 (17)		12,942 (14)	
Alpha glucosidase	1,545 (2)		58 (1)		117 (2)		1,720 (2)	
SGLT2	615 (1)		21 (1)		59 (1)		695 (1)	
Metglinides	902 (1)		58 (1)		100 (1)		1,060 (1)	
Other medications (Ever prescribed) *								
Antihypertensive	3,577 (5)	7,496 (2)	173 (4)	390 (2)	145 (2)	320 (1)	3,895 (4)	8,206 (2)
Diuretics	20,688 (26)	41,226 (13)	585 (14)	1,421 (9)	684 (9)	1,508 (5)	21,957 (24)	44,155 (12)
Beta blockers	18,938 (24)	42,580 (13)	427 (10)	1,103 (7)	758 (11)	1,913 (7)	20,123 (22)	45,596 (13)
Calcium blockers	15,070 (19)	30,059 (10)	724 (18)	1,588 (10)	703 (10)	1,386 (5)	16,497 (18)	33,033 (9)
Statins	19,376 (24)	35,419 (11)	627 (15)	811 (5)	1,219 (17)	1,679 (6)	21,222 (23)	37,909 (11)
Ace inhibitors	17,145 (22)	30,152 (10)	558 (14)	889 (5)	823 (11)	1,370 (5)	18,526 (20)	32,411 (9)
CPMs	34,595 (44)	73,919 (23)	1,272 (31)	2,473 (15)	1,997 (28)	3,671 (13)	37,864 (42)	80,063 (22)
Anti-depressants	17,419 (22)	54,807 (17)	393 (10)	1,413 (9)	926 (13)	3,025 (10)	18,738 (21)	59,245 (16)

Anti-obesity	3,213 (4)	2,940 (1)	121 (3)	171 (1)	200 (3)	229 (1)	3,534 (4)	3,340 (1)

*: *n (%); †: mean ± SD; ‡: median (Q1, Q3);§: Not presented for control subjects; Any CVD: Defined as the occurrence of angina or myocardial infarction or coronary artery disease (including bypass surgery and angioplasty) or heart failure or stroke before diagnosis*

Table 5.2: Mean ± SD of BMI in three ethnic groups at the time of diagnosis of T2DM, by different age groups at diagnosis among patients diagnosed from 01 Jan 2006.

Age group:	n	BMI†	Absolute difference in mean (p-value)			BMI categories*		
			WE Vs SA	WE Vs AC	AC Vs SA	Grade 1 Obese	Grade 2 Obese	Grade 3 Obese
≤ 40 years								
White European	2051	36.3 ± 8.6	6.6 (<0.001)	5.2(<0.001)	1.4 (<0.022)	509 (24)	468 (23)	615 (30)
African-Caribbean	311	31.1 ± 6.5				89 (29)	43 (14)	28 (9)
South Asian	892	29.8 ± 6.2				212 (24)	90 (10)	64 (7)
41-50 years								
White European	5717	35.1 ± 7.4	5.3 (<0.001)	3.6 (<0.001)	2.1 (<0.001)	1635 (29)	1360 (24)	1263 (22)
African-Caribbean	716	31.5 ± 6.3				215 (30)	128 (18)	57 (8)
South Asian	1092	29.4 ± 5.8				278 (26)	109 (10)	44 (4)
51-60 years								
White European	7907	33.4 ± 6.7	3.8 (<0.001)	2.2 (<0.001)	1.6 (<0.001)	2537 (32)	1644 (21)	1139 (14)
African-Caribbean	497	31.2 ± 6.2				132 (27)	82 (17)	43 (9)
South Asian	679	29.6 ± 5.7				167 (25)	73 (11)	27 (4)
61-70 years								
White European	8487	32.1 ± 6.2	3.5 (<0.001	1.1 (<0.038)	2.3 (<0.001)	2849 (34)	1446 (17)	824 (10)
African-Caribbean	188	30.9 ± 5.4				65 (35)	27 (14)	12 (6)
South Asian	264	28.6 ± 5.4				56 (21)	21 (8)	9 (3)
70+ years								
White European	5198	30.2 ± 5.4	2.7 (<0.001)	0.44 (1.000)	2.3 (0.015)	1554 (30)	637 (12)	255 (5)
African-Caribbean	80	29.7 ± 6.6				19 (24)	9 (11)	8 (10)
South Asian	100	27.4 ± 5.5				18 (18)	8 (8)	3 (3)
Female								
White European	12,292	33.9 ± 7.7	3.4 (<0.001)	0.6 (0.051)	2.8 (<0.001)	3414 (28)	2534 (21)	2328 (19)
African-Caribbean	823	33.3 ± 6.4				265 (32)	194 (24)	107 (13)
South Asian	1303	30.5 ± 6.0				344 (26)	193 (15)	86 (7)
Male								
White European	17068	32.3 ± 6.2	3.7 (<0.001)	2.8 (<0.001)	0.9 (0.002)	5,670 (33)	3,012 (16)	1768 (10)

African-Caribbean	969	29.5 ± 5.4	255 (26)	95 (10)	41 (4)
South Asian	1724	28.6 ± 5.6	387 (22)	108 (6)	61 (4)

**: n (%); †: mean ± SD;*
The p values are estimated to present the significance of differences in the distribution of BMI between each combination of two ethnic groups. The proportions of patients under different obesity grades are presented by number (%). WE: White European; SA: South Asian; AC: African-Caribbean

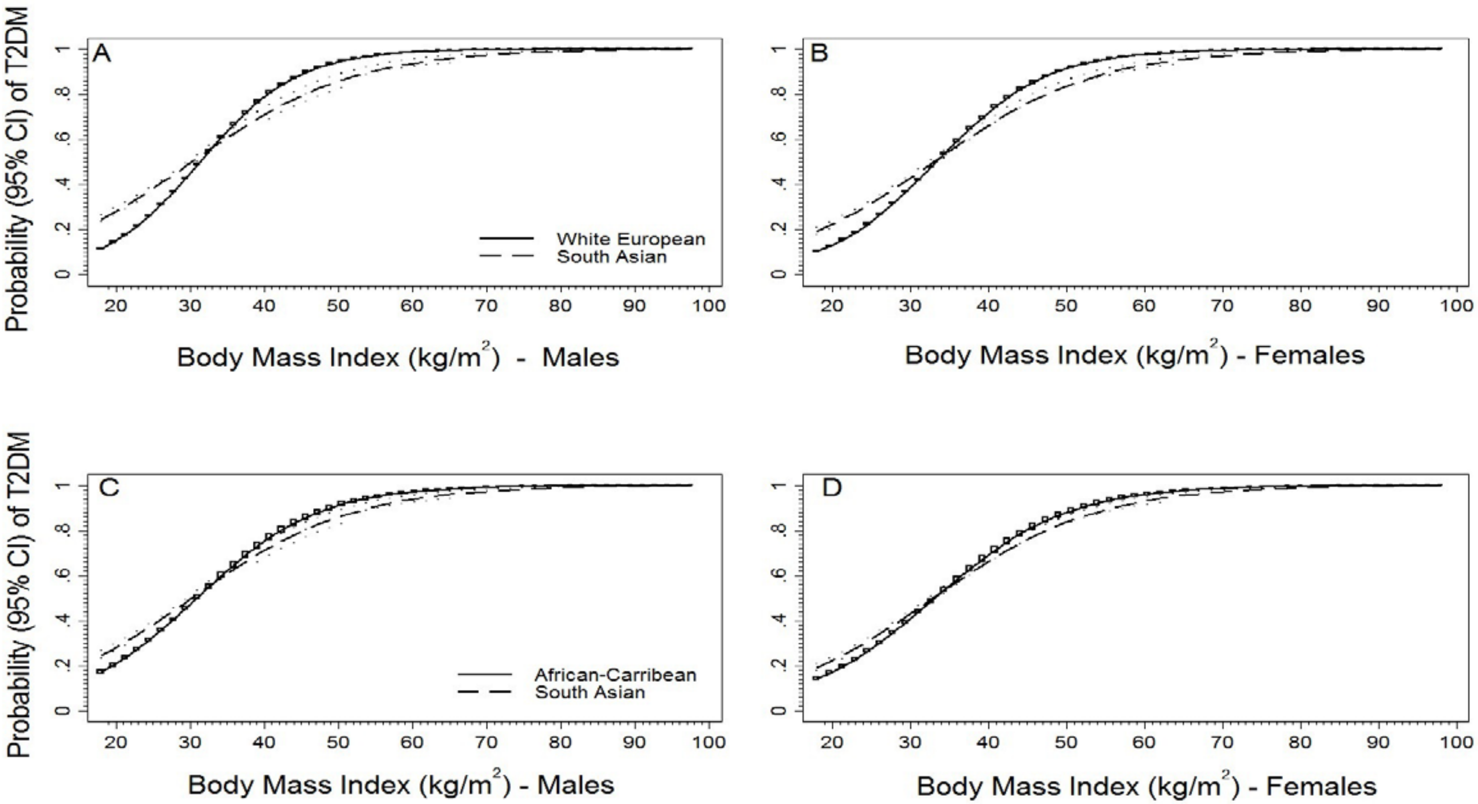


Figure 5.2: The association between BMI at diagnosis and risk of T2DM (95% CI) compared between three ethnic groups.

[***A****: Adjusted probability of developing T2DM (95% CI) at different levels of BMI for male South Asians, compared to male White Europeans;* ***B****: Adjusted probability of developing T2DM (95% CI) at different levels of BMI for female South Asians compared to female White Europeans ;* ***C****: Adjusted probability of developing T2DM at different BMI levels for male South Asians compared to male African-Caribbeans;* ***D****:Adjusted probability of developing T2DM at different BMI levels for female South Asians compared to female White Europeans ; Probability estimated in Figures 1A, 1B, 1C, & 1D are adjusted for age, smoking status, deprivation score, and history of CVD, cancer and CKD on or prior to the index date.*].

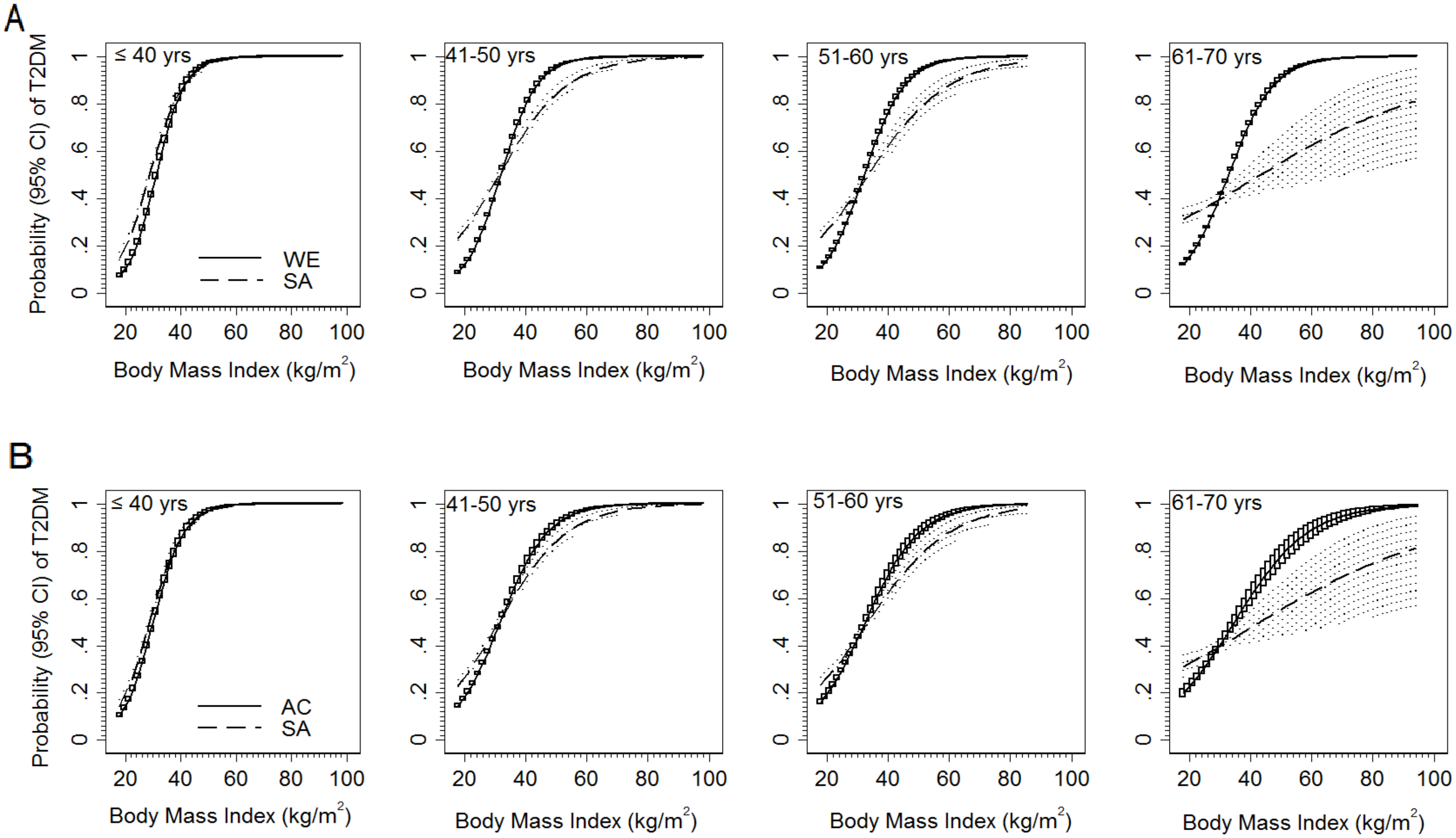


Figure 5.3: The adjusted probability of developing T2DM (95% CI) across levels of BMI for different age groups.

[***A***: *South Asians compared to White Europeans;* ***B***: *South Asians compared to African-Caribbeans; Estimates are adjusted for sex, smoking status, deprivation score, and history of CVD, cancer and CKD on or prior to the index date; SA: South Asian; WE: White European; AC: African-Caribbean*].

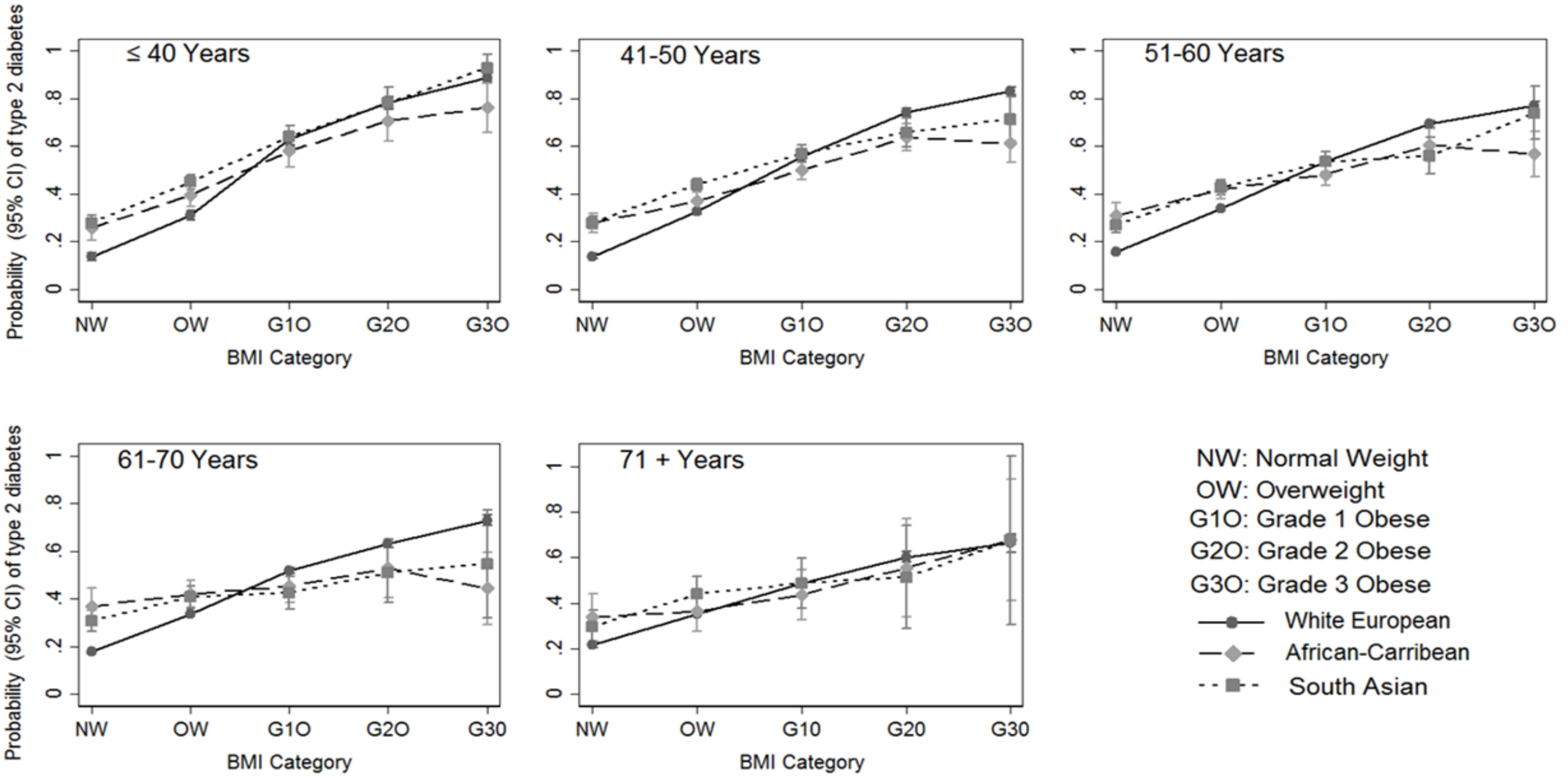


Figure 5.4: The adjusted probability of developing T2DM (95% CI) across levels of BMI categories for different age groups.

[*Adjusted for sex, smoking status, deprivation score, and history of CVD, cancer, and CKD on or prior to the index date; NW: Normal weight; OW: Overweight; G1O: Grade 1 Obese; G2O: Grade 2 Obese; G3O: Grade 3 Obese*].

Given the different patterns of risk of developing T2DM among obese patients between ethnic groups across different age groups (Figure 5.3 and 5.4), we evaluated the odds of developing T2DM in African-Caribbeans and South Asians, compared to White Europeans, separately for each age group. Within each age group, the probability of developing T2DM was greater amongst South Asians at lower BMI. This relationship was reversed at higher BMI levels and compared to White Europeans, the South Asians had 22% (95% CI of OR: 0.65,0.95), 30% (95% CI of OR: 0.62,0.81), 24% (95% CI of OR: 0.65,0.88) and 39% (95% CI of OR: 0.48,0.77) lower odds (adjusted) of developing T2DM in the age groups ≤ 40, 41-50, 51-60 and 61-70 years respectively. African-Caribbean patients had 43% (95% CI of OR: 0.45, 0.73), 43% % (95% CI of OR: 0.50, 0.66), 33% (95% CI of OR: 0.57, 0.78) and 35% (95% CI of OR: 0.52, 0.82) lower adjusted odds of developing T2DM in the respective age groups. However, these odds were not statistically significantly different between African-Caribbeans and South Asians (Table 5.3).

Table 5.3: Odds ratio (95% CI) for development of T2DM among obese African-Caribbean and South Asian compared to obese White European, separately for each age group.

	White European	**African-Caribbean**	**South Asian**
	OR (95% CI)	**OR (95% CI)**	**OR (95% CI)**
Age group: ≤ 40 years	Reference	0.57 (0.45,0.73)	0.78 (0.65,0.95)
Age group: 41-50 years	Reference	0.57 (0.50,0.66)	0.67 (0.57,0.78)
Age group: 51-60 years	Reference	0.67 (0.57,0.78)	0.76 (0.65,0.88)
Age group: 61-70 years	Reference	0.65 (0.52,0.82)	0.61 (0.48,0.77)
Age group: 70+ years	Reference	0.84 (0.58,1.23)	0.91 (0.61,1.34)

5.5 DISCUSSION

This case-control study with a large number of White European, South Asian and African-Caribbean individuals from a nationally representative primary care database reveals significantly different (1) distributions of body weight, BMI and other cardiovascular risk factors at the time of diagnosis of T2DM and (2) probability of developing T2DM over the whole spectrum of BMI, in interaction with age and sex. This study also reveals that the risk patterns of developing diabetes at different levels of obesity varies between ethnic groups across all age groups. To the best of our knowledge, this is the first study exploring the variations in T2DM risk over the whole distribution of BMI at the time of diagnosis across the South Asian, African-Caribbean and White European populations.

Our findings confirm the association of increased risk of T2DM with increasing BMI. More importantly, it adds to the evidence that for any given age, South Asians have a greater risk of T2DM at lower BMI. Typically, African-Caribbeans and South Asians in our study were significantly younger and had a distinct metabolic risk profile compared to White Europeans characterised by lower body weight, systolic blood pressure, and lower rates of smoking but significantly higher HbA_{1c} levels. The observed higher HbA_{1c} level in South Asians and African-Caribbeans is in line with earlier findings[191]. While different possible reasons, including ethnic differences in pre- and post-prandial glycaemia and glycation rate of haemoglobin have been postulated, no confirmatory mechanistic study has yet been reported on this aspect. At the population level, evaluation of longitudinal patterns of pre- and post-prandial glucose changes along with the measures of insulin deficiency from the pre-diabetes state may reflect some light on this issue. Based on a US population with 12,179 T2DM patients and 25,177 controls, Ganz and colleagues (2014) reported a mean age and BMI of 55 years and 35 kg/m^2 respectively at diagnosis [30]. With similar age at diagnosis of T2DM, our UK study cohort had a significantly lower BMI level (32 kg/m^2). However, this distribution of BMI at diagnosis is consistent with other studies reporting BMI at diagnosis of T2DM in the UK population [12,192,193]. Similarly, our finding that women developed T2DM at significantly higher BMI level compared to men across ethnicity is consistent with earlier reports [194,195]. Earlier studies have shown that the onset of T2DM occurs up to a decade early amongst South Asians. A Canadian cohort based diabetes incidence study reported that the median age at diagnosis was lowest among South Asians (49 years), followed by African-Caribbeans (57 years), and Whites (58 years) [185]. Our data is consistent with these observations and on average South Asians were 12 years younger at diagnosis compared to their White European counterparts. Factors that influence the predisposition of South Asians to develop T2DM at younger are largely unknown and may be related to a combination of genetic and environmental factors that have not yet been fully characterised [194 195].

Ganz and colleagues (2014) reported significantly increased odds of developing T2DM with increasing BMI [30]. While this study also evaluated the risk of developing T2DM in various age groups using additive models, the interaction of age and body weight in the risk of developing T2DM was not explored. Given that age and obesity are two major risk factors for T2DM, we explored the interaction of age and BMI levels (separately for male and female) in evaluating the risk of developing T2DM across the three ethnic groups. One of the novelties of this study is a comparative exploration of the probability of developing T2DM over the whole continuous distribution as well as categories of BMI by ethnic groups. When BMI was analysed as a continuous variable, we found that South Asians aged 40 years and above had a significantly greater probability of T2DM at lower BMI levels (18-30kg/m^2) compared to the other two ethnic groups. When analysed with BMI as a categorical variable, the higher probability of T2DM for South Asians with lower BMI extended from those younger than 40 years to those less than 70 years of age. This difference in observed probability of developing T2DM using continuous BMI versus BMI categories is reflective of the fact that there is a loss of statistical information when converting a continuous variable to a categorical variable. Interestingly, in both the analyses, we identified a distinct pattern of risk between South Asians and White Europeans with the probability of T2DM being greater at lower BMI for South Asians and at higher BMI for White Europeans.

Earlier ethnicity-specific studies evaluating the association of prior BMI with the incident rates of T2DM reported higher risk in South Asians at a lower BMI level [28,185,196]. During a median follow-up of 6 years, Chui and colleagues reported higher T2DM incidence rates T2DM in South Asians at lower ages and BMI compared to the White Europeans [185]. A similar observation was made in another follow-up study reported by Tillin and colleagues [28]. Our study elaborates on the significantly higher likelihood of developing T2DM at lower BMI levels among South Asians, compared to White Europeans and African-Caribbeans. We have also identified a significant change in the risk pattern at higher BMI levels while compared between ethnic groups at different age levels (Figure 5.3 and 5.4). Additionally, our study provides detailed information on the contrasting probability of developing T2DM at different age groups and ethnicity across the entire spectrum of BMI. There are several possible explanations for this contrasting effects of obesity on the probability of diabetes between ethnic groups. While BMI is an accepted measure of obesity, it does not differentiate between patterns of obesity (visceral vs. subcutaneous). It is well known that South Asians have excess visceral adiposity even at lower BMI and that may explain the higher propensity of South Asians to develop T2DM at lower BMI. However, that pattern would be expected to persist even at higher levels of BMI and therefore South Asians would be expected to have a greater risk in comparison with White Europeans for any level of BMI. Our observation that this difference is only evident at lower BMI but not at higher BMI range would suggest that the excess visceral adiposity alone does not explain this variance. An alternative explanation could be that relative contribution of obesity to the risk of T2DM may be greater amongst White Europeans compared to that in South Asians and that the risk of T2DM in South Asians may additionally be determined by underlying beta cell dysfunction. Clearly, this needs to be addressed in future studies.

The strength of this study is that it includes a large number of T2DM patients from a primary care system with a large number of South Asians; a representative age and sex-matched non-diabetic control cohort; use of anthropometric, clinical risk factor measures at index date; and a robust analysis approach to explore the potential interactions between age, sex and BMI in different ethnic groups. Patient-level data from electronic health records present challenges in terms of accuracy and completeness of the study variables of interest. The limitations of this study include (1) availability of ethnicity data on a limited number of patients, (2) missing risk factor data, (3) potential for residual confounding, and (4) inability to draw a causal link between BMI and T2DM, as with all observational studies. However, ethnicity recording for South Asians and African-Caribbeans in the electronic database used for this study is comparable to the general population of UK [197]. We also attempted to minimize bias introduced by confounders through the use of multivariate models with a detailed list of possible confounders. However, unavailability of information on education, physical activity, diet, and other risk factors may have introduced bias into the risk estimates.

5.6 CONCLUSION

The South Asian and African-Caribbean populations have an increased burden of T2DM with its complications. In this large case control data analysis, we have demonstrated that T2DM occurs in South Asians at least 12 years earlier with mean age 46 years, when compared to White Europeans, with about a third developing under the age of 40 years. We believe the early presentation of diabetes in this ethnic population contributes to the glycaemic load and the burden of complications. Hence, the early diagnosis of diabetes, recognising the lower age of presentation may help to ameliorate the glycaemic burden and to do this a lower age cut-off for screening in national programmes for South Asians is required. This is the first large cohorts that we are aware of in which it has been demonstrated that South Asians develop diabetes at a mean lower BMI by 5 kg/m^2 when compared to White Europeans. This has implications both in terms of diagnosing obesity in South Asians along with appropriate management interventions.

Chapter 6: Prevalence and incidence of complications at diagnosis of T2DM and during follow-up by BMI and ethnicity: a matched case-control analysis

This body of this chapter contains one published paper that discusses the prevalence of complications at diagnosis of diabetes, and incidence of complications during follow-up in comparison to non-diabetic controls. The citation of the published paper is as follows:

Owusu Adjah ES, Bellary S, Hanif W, Patel K, Khunti K, Paul SK. Prevalence and incidence of complications at diagnosis of T2DM and during follow-up by BMI and ethnicity: a matched case-control analysis. *Cardiovascular Diabetology* 2018;**17**(1):70.

All the listed have agreed to the inclusion of this published scholarly work in this thesis and the statement of my contribution to the authorship of this published scholarly work is included below:

Contributors	Statement of contribution
Owusu Adjah Ebenezer S. (Candidate)	Conceived the idea and was responsible for the primary design of the study. Responsible for the data extraction from THIN database. Responsible for data manipulation, aggregation, transformation in SAS. Conducted the statistical analyses in STATA and interpretation of results. Developed first draft and contributed towards finalisation of the manuscript.
Samanta Mayukh	Contributed to the interpretation of the results and manuscript finalisation.
Patel Kiram	Contributed to the interpretation of the results and manuscript finalisation.
Bellary Srinath	Contributed to the interpretation of the results and manuscript finalisation.
Hanif Wasim	Contributed to the interpretation of the results and manuscript finalisation.
Khunti Khunti	Contributed to the interpretation of the results and manuscript finalisation.
Paul Sanjoy K.	Conceived the idea and was responsible for the primary design of the study. Contributed to the statistical analyses. Developed first draft and contributed towards finalisation of the manuscript.

6.1 Abstract

Aims: To estimate the risk of developing long-term major cardiovascular and renal complications in relation to levels of body mass index (BMI) in a population of White European (WE), African-Caribbean (AC), and South Asian (SA) patients with type 2 diabetes mellitus (T2DM).

Materials and methods: Patients with a new diagnosis of T2DM, aged ≥18 years from January 2000 (n=69,436) and their age-sex-ethnicity matched non-diabetic controls (n =272,190) were identified from UK primary care database. Incidence rates ratios (IRRs) for non-fatal major cardiovascular events (MACE) and chronic kidney disease (CKD) in patients with T2DM compared to controls were estimated using multivariate Mantel-Cox model.

Results: Among normal weight patients with T2DM, WEs had a similar prevalence of cardiovascular multi-morbidity (95% CI: 9.5, 11.3) compared to SAs (95% CI: 4.8, 9.5). African-Caribbean (AC) and SA overweight and obese patients had similar prevalence, while obese WEs had a significantly higher prevalence. During a median 7 years of follow-up, the risk of MACE was significantly higher for overweight (95% CI of IRR: 1.50, 2.46) and obese (95% CI of IRR: 1.49, 2.43) SAs compared to their WE counterparts. However, similar risk levels were observed for normal weight WEs and SAs respectively. Risk of CKD was higher and uniform for BMI ≥ 25 kg/m^2 amongst WEs and ACs, whereas only overweight patients had a significantly higher risk of CKD amongst SA [IRR: 2.08 (95 % CI: 1.49, 2.93)].

Conclusion: Risk of MACE / CKD varies over levels of BMI within each ethnic group, with overweight SAs having a disproportionate risk of CKD.

Keywords

Type 2 diabetes, Body Mass Index, Complications, Ethnicity, Prevalence, and Incidence

6.2 Introduction

Ethnicity remains one of the key risk factors for type 2 diabetes mellitus (T2DM) and the predisposition of certain ethnic groups to develop T2DM is now well known [198]. Not only does diabetes occur early in some ethnic groups [41,199], but there is also a greater predisposition to develop diabetes-related complications [36]. This disproportionate predisposition of certain ethnic groups to T2DM and its complications is commonly attributed to the complex interaction of genetic and environmental factors [200,201]. Several studies have compared the prevalence and severity of diabetes complications between South Asians and White Europeans [202-207]. Although some studies have generally reported a higher prevalence of some complications (particularly nephropathy and retinopathy) [206,208], other studies have shown these differences are not as significant as thought [205,209].

The UK Prospective Diabetes Study Group (UKPDS) evaluated the incidence of myocardial infarction (MI) by ethnicity and found no additional risk of MI among South Asian (SA) and African-Caribbean (AC) participants respectively compared to White European (WE) participants [205]. While this study accounted for some cardiovascular risk factors in their risk assessment model, body mass index (BMI) which is an important cardiovascular risk factor in patients with T2DM was not included. Furthermore, while other studies have evaluated the ethnicity-related differences in the incidence of cardiovascular events in patients with T2DM [204,210-212], no separate assessment of the potential differences in the risk paradigm by adiposity levels were evaluated for each ethnic group.

Given that BMI and ethnicity play important roles in cardiovascular risk profiles of patients with T2DM, we are not aware of any study that has evaluated ethnicity-specific long-term cardiovascular and non-cardiovascular complications in T2DM by BMI categories at the population level. Such evaluations are of immense public health importance given the increased burden of complications associated with T2DM [187,213,214], and will address the knowledge gap in terms of the interplay between ethnicity, BMI, cardiovascular, and non-cardiovascular complications in patients with T2DM [3]. Therefore, the aims of this primary care data based retrospective longitudinal case-control study were to evaluate (1) comorbidities and cardiovascular risk factors at diagnosis of T2DM in different ethnic groups, and (2) the likelihood of developing long term complications by BMI categories in different ethnic groups compared to non-diabetic controls.

6.3 Methods

6.3.1 Data source

Data from the primary care database of UK [The Health Improvement Network (THIN)] was used. Patients are registered with one general practitioner (GP) even though secondary care treatment can be provided elsewhere, and under terms specified by the UK's National Health Service (NHS), GPs contribute data to THIN. Thus, daily electronic medical records (EMRs) of patients in participating practices are regularly submitted to THIN using the INPS ViSion software [122]. The database is linked to other sources of hospital and national statistics data and is demographically representative of the UK. Currently, data from over 600 general practices involved with THIN from 1990 to 2014 is available. The source population includes over 13 million patients, 85% of whom have records that are considered valid and acceptable for research. The accuracy and completeness of this database have been previously described elsewhere [124,125]. This database provides comprehensive patient-level longitudinal information on demographic, anthropometric, clinical and laboratory measures, clinical diagnosis of diseases and events, along with complete information on prescriptions for medications with dates and doses. Clinically diagnosed diseases are recorded using Read codes [128], and with each diagnosis, an event date is entered. Similarly, prescriptions are recorded with both British National Formulary (BNF) codes and Anatomical Therapeutic Chemical (ATC) codes along with their prescription dates.

6.3.2 Study population

The primary design and results have already been published [41]. Briefly, from THIN database 69,436 patients with newly diagnosed T2DM from January 2000 were identified using a robust machine-learning algorithm, which uses the disease Read codes [128], antidiabetic medications, and lifestyle modification interventions as feeds. Patients were included if they had (1) complete information on age at diagnosis (≥ 18 years) and sex, and (2) self-identified ethnicity as WE, AC or SA. South Asians (SAs) were defined as patients with Indian, Pakistani, Sinhalese, and Bangladeshi origin, while ACs were defined as patients with Black-African and/or Caribbean origin. White Europeans (WEs) were patients with self-reported ethnicity as White, European, European, and/or New Zealand European. Those with Read codes for type 1 diabetes mellitus (T1DM) and gestational diabetes were excluded. Non-diabetic patients were patients in the THIN database with no diagnosis of any type of diabetes and had never received a prescription of an anti-diabetes therapy. Up to four non-diabetic control patients (n=272,190) were matched to each identified T2DM patient based on age, sex and ethnicity using an exact matching algorithm. The index date for controls was defined as the date of the diabetes diagnosis for their matched cases.

6.3.3 Study variables and outcome measurements

Clinical and demographic variables including smoking status, deprivation score (measure of socioeconomic status based on residential address), weight, BMI, glycated haemoglobin (HbA1c), systolic blood pressure (SBP), diastolic blood pressure (DBP), low density lipoproteins (LDL), high density lipoproteins (HDL), and triglycerides were extracted for each patient where appropriate. All available measures on or within 3 months prior to the index date were considered as baseline measures. For all clinical parameters, longitudinal data 12 months prior to index date and 2 years post index date were extracted on a 6-monthly window. Categories for BMI were defined following WHO established criteria as follows: normal weight (18.5-24.9 kg/m^2), overweight (25-29.9 kg/m^2), and obese (≥ 30 kg/m^2). For South Asians, BMI in the ranges 18.5-22.9, 23-27.4, ≥27.5kg/m^2 were used to define normal weight, overweight and obese patients respectively [104]. Prescription information on anti-diabetes therapies, antihypertensive agents, cardio-protective medications (CPM), weight-lowering drugs and anti-depressants were also obtained, where appropriate.

Patients with a recorded diagnosis of stroke, heart failure (HF), angina, MI, coronary artery disease (including bypass surgery and angioplasty), cancer, or renal diseases (including chronic kidney disease (CKD)) before diagnosis were considered to have relevant comorbidities at diagnosis. Subsequently, cardiovascular multi-morbidity was defined as ≥ 2 episodes of a major cardiovascular condition at diagnosis. A composite variable for major cardiovascular events (MACE) was defined as the occurrence of non-fatal MI, HF or stroke during follow-up. Time to a specific disease event was calculated as the time from diagnosis date to the first occurrence of the disease event and patients were censored on the end date (September 2014) or on drop out date.

6.3.4 Statistical analysis

Baseline characteristics of patients with incident T2DM and their matched non-diabetic controls were summarized using number (%), means (95% CI) or median (first quartile, third quartile) as appropriate. Age-sex standardised proportions of existing comorbidities at diagnosis were calculated with indirect standardisation to the internal data structure. Age groups (18-40, 41-50, 51-60, 61-70, and 71+ years) and sex (male vs. female) were used to achieve stratum-specific proportions for indirect standardisation.

Major cardiovascular event (MACE) and CKD (stage $\geq$ 3) incident rates (rates per 1000 person-years) were estimated by BMI categories for T2DM cases and controls separately for each ethnic group. To estimate MACE and CKD (stage $\geq$ 3) incidence rate ratio (IRR) for T2DM cases compared to controls, a multivariate Mantel-Cox model was fitted: adjusting for age, sex, baseline SBP, smoking status (current, ex, and never smokers), and deprivation score by stratification. Robust estimates of IRRs (95% CI) were obtained, and Bayesian information criteria (BIC) was used to compare the model fits.

6.4 Results

6.4.1 Demographic and clinical characteristics

The demographic and clinical profiles of T2DM patients (n=69,436) and matched non-diabetic controls (n =272,190) are presented in Table 6.1. Overall, the mean age at diagnosis was 57 years, 57% were male, and median follow-up time was similar across T2DM cases and their non-diabetic controls (7 years). Within subgroups defined by ethnicity, T2DM patients and their non-diabetic controls were well matched on age and sex distributions. The distribution of current or ex-smokers in T2DM patients and controls were 55% and 50% respectively, and the proportions of patients with SBP $\geq$ 140 mmHg were 39% and 18% respectively.

Compared to WEs and ACs, SAs developed diabetes significantly earlier by (~10 and 2 years) and at lower BMI (3 and 2 kg/m^2, Table 6.1). More SAs (66%) developed T2DM within the age of 50 years, while 27% and 59% of WEs and ACs developed the disease within the same age limit respectively. Significantly higher proportions of WE cases and controls had SBP above 140 mmHg (41 and 21%), compared to ACs (30 and 12%) and SAs (23 and 9%) respectively.

Table 6.1: Baseline clinical characteristics of patients with T2DM and their matched non-diabetic controls separately for each ethnic group

	White European (296,288)		**African-Caribbean** (16,958)		**South Asian** (28,380)		**Overall** (341,626)	
	T2DM	**Control**	**T2DM**	**Control**	**T2DM**	**Control**	**T2DM**	**Control**
Patients†	60,233 (20)	236,055 (80)	3,425 (20)	13,533 (80)	5,778 (20)	22,602 (80)	69,436 (20)	272,190 (80)
Age at index (years) ‡	58 (58.2,58.4)	58 (58.3,58.4)	49 (48.2,49.0)	49 (48.4,48.8)	47 (46.3,46.9)	47 (46.4,46.7)	57 (56.8,57.0)	57 (56.8,56.9)
Age groups †								
18-40	4,530 (8)	17,912 (8)	744 (22)	2,932 (22)	1,707 (30)	6,685 (30)	6,981 (10)	27,529 (10)
41-50	11,297 (19)	44,306 (19)	1,278 (37)	5,063 (37)	2,097 (36)	8,196 (36)	14,672 (21)	57,565 (21)
51-60	16,552 (28)	65,033 (28)	843 (25)	3,326 (25)	1,217 (21)	4,767 (21)	18,612 (27)	73,126 (27)
61-70	17,010 (28)	66,485 (28)	415 (12)	1,641 (12)	559 (10)	2,185 (10)	17,984 (26)	70,311 (26)
71+	10,844 (18)	42,319 (18)	145 (4)	571 (4)	198 (3)	769 (3)	11,187 (16)	43,659 (16)
Male †	34,342 (57)	134,630 (57)	1,778 (52)	7,040 (52)	3,232 (56)	1,2631 (56)	39,352 (57)	154,301 (57)
Current smokers †	12,830 (21)	46,926 (20)	449 (13)	1,984 (15)	820 (14)	2,839 (13)	14,099 (20)	51,749 (19)
Ex-smokers †	23,196 (39)	80,860 (34)	614 (18)	2,066 (15)	756 (13)	2,605 (12)	24,566 (35)	85,531 (31)
Deprivation Status								
Highest affluence †	12,856 (21)	41,726 (18)	833 (24)	3,548 (26)	1,483 (26)	5,586 (25)	15,172 (22)	50,860 (19)
Lowest affluence †	2,500 (4)	11,462 (5)	457 (13)	1,595 (12)	518 (9)	1,989 (9)	3,475 (5)	15,046 (6)
HbA$_{1c}$ (%), ‡ ¶	8.2 (8.2,8.2)		9.1 (9.0,9.2)		8.5 (8.4,8.6)		8.3 (8.3,8.3)	
Weight (kg) ‡	92.9 (92.8,93.1)	78.7 (78.5,78.8)	88.7 (87.9,89.4)	81.6 (81.0,82.1)	79.2 (78.7,79.8)	72.2 (71.8,72.5)	91.7 (91.5,91.9)	78.3 (78.2,78.4)
BMI (kg/m^2) ‡	32.6 (32.6,32.7)	27.8 (27.8,27.8)	31.5 (31.3,31.7)	28.2 (28.1,28.2)	29.6 (29.5,29.8)	26.3 (26.3,26.4)	32.3 (32.3,32.4)	27.7 (27.7,27.7)
Normal weight †	4,242 (7)	29,128 (12)	359 (11)	1,174 (9)	360 (6)	1,517 (7)	4,961 (7)	31,819 (12)
Overweight †	14,446 (24)	178,671 (76)	842 (25)	10,611 (78)	1,505 (26)	16,750(74)	16,793 (24)	206,032 (76)
Obese †	41545 (69)	28256 (12)	2224 (65)	1748 (13)	3,913(68)	4,335 (19)	47,682 (69)	34,339 (13)
SBP (mmHg) ‡	140 (139.7,140)	136 (135.6,135.8)	136 (135.7,137.2)	133 (132.3,133.3)	132 (131.0,132.1)	128 (127.7,128.5)	139 (138.9,139.2)	135 (135,135.2)

SBP ≥ 140 mmHg †	24571 (41)	46081 (20)	1029 (30)	1624 (12)	1302 (23)	1926 (9)	26902 (39)	49631 (18)
LDL(mg/dl) ‡	119 (118.6,119)	122.3 (122.2,122.3)	127 (126.4,128.3)	128.7 (128.4,128.9)	121 (120.3,121.8)	123.9 (123.7,124.1)	119 (119.2,119.6)	122.7 (122.7,122.8)
HDL (mg/dl) ‡	46 (45.6,45.8)	55 (55.3,55.4)	48 (47.4,48.2)	57 (56.4,56.6)	43 (43.1,43.6)	51 (51.3,51.5)	46 (45.5,45.7)	55 (55.0,55.1)
Triglycerides (mg/dl) §	159 (121, 213)	115 (88, 159)	115 (81, 159)	82 (62, 115)	151 (115, 204)	115 (89, 168)	159 (115, 213)	115 (84, 159)
Comorbidities	18,014 (30)	46,449 (20)	382 (11)	963 (7)	647(11)	1,741(8)	19,043 (27)	49,153 (18)
Cardio-protective drugs †								
Beta blockers	17042 (28)	38032 (16)	393 (12)	1041 (8)	710 (12)	1783 (8)	18145 (26)	40856 (15)
Calcium blockers	13736 (23)	27188 (12)	694 (20)	1535 (11)	673 (12)	1283 (6)	15103 (22)	30006 (11)
Statins	18971 (32)	33604 (14)	620 (18)	794 (6)	1205 (21)	1618 (7)	20796 (30)	36016 (13)
ACE inhibitors	16165 (27)	27922 (12)	532 (16)	857 (6)	793 (14)	1290 (6)	17490 (25)	30069 (11)
Follow-up §	7.0 (4, 11)	8.0 (4, 11)	7.0 (4, 10)	7.0 (4, 10)	6.0 (3, 10)	7.0 (4, 10)	7.0 (4, 11)	7.0 (4, 11)

†: n (%);
‡: mean (95% CI);
§: Median (Q1, Q3);
¶: Not presented for non-diabetic controls.

Abbreviations and definitions:
ACE, angiotensin-converting enzyme;
SBP, systolic blood pressure;
DBP, diastolic blood pressure,
LDL-C, low-density lipoprotein cholesterol;
HDL, high-density lipoprotein cholesterol;
Comorbidities: Pre-existing cardiovascular (myocardial infarction, stroke, heart failure, angina, or coronary heart disease) or non-cardiovascular disease (renal diseases including chronic kidney disease, cancer, or depression) at the time of diagnosis.

Table 6.2: Age-sex-adjusted prevalence (95% CI) of cardiovascular complications at diagnosis by BMI categories among patients with T2DM, separately for each ethnic group

		Prevalence (95% CI)			
		MACE	MI	HF	STROKE
Normal weight	White European	10.4(9.5,11.3)	4.6(4.0,5.2)	4.6(4.1,5.2)	5.0(4.4,5.6)
	African-Caribbean	6.5(4.0, 10.4)	2.0(0.9,4.9)	2.1(0.9,4.7)	4.2(2.3,7.7)
	South Asian	6.8(4.84, 9.5)	4.0 (2.6,6.2)	4.0(2.6,6.2)	3.1(1.8,5.1)
Overweight	White European	11.7(11.3,12.2)	6.1(5.7,6.5)	6.1(5.7,6.5)	5.2(4.9,5.6)
	African-Caribbean	7.2(5.1,9.9)	2.1(1.0,4.1)	2.1(1.0,4.1)	4.9(3.4,7.2)
	South Asian	9.0(7.4,10.9)	5.3(4.1,6.9)	5.3(4.1,6.9)	3.7(2.7,5.1)
Obese	White European	12.6(12.3,12.9)	6.5(6.3,6.7)	6.5(6.3,6.7)	5.4(5.2,5.6)
	African-Caribbean	5.5(4.1,7.4)	1.1(0.5,2.5)	1.1(0.5,2.5)	4.4(3.2,6.2)
	South Asian	8.5(6.2,11.7)	4.7(2.9,7.4)	4.7(2.9,7.4)	2.5(1.5,4.1)

6.4.2 Prevalence of comorbidities at diagnosis

T2DM cases had a significantly higher proportion of existing comorbidities at diagnosis compared to controls (27% vs. 18%, Table 6.1). The prevalence (95% CI) of cardiovascular complications at diagnosis by BMI categories among patients with T2DM, separately for each ethnic group are presented in Table 6.2. Among normal weight patients with T2DM, WEs had similar prevalence of cardiovascular multi-morbidity (prevalence: 10.4%; 95% CI: 9.5, 11.3), compared to SAs (prevalence: 6.8%; 95% CI: 4.8, 9.5), and ACs (prevalence; 95% CI: 4.0, 10.4). African-Caribbean and SA overweight and obese patients had a similar prevalence of cardiovascular multi-morbidity across all adiposity levels, while obese WEs had significantly higher prevalence compared to their normal weight population and also compared to other ethnic groups (Table 6.2).

The prevalence of cardiovascular and non-cardiovascular diseases at diagnosis between T2DM cases and their non-diabetic controls, separately for each ethnic group are presented in Figure 6.1 and 6.2 respectively. White Europeans with or without diabetes had a significantly higher prevalence of cancer, compared to SA cases and controls (Figure 6.2A). The prevalence of depression among WE cases and controls were significantly higher (95% CI of proportion - cases: 21.8 - 22.5%; controls: 17.3 - 17.5%) compared to other ethnic groups, while SA and AC cases and controls had similar prevalence (range of 95% CI of prevalence: 6.6 - 9.7%). The prevalence of CKD at diagnosis was similar across all ethnic groups and did not differ significantly between T2DM cases and their non-diabetic controls (Figure 6.2).

6.4.3 The incidence of major cardiovascular diseases during follow-up

In individuals without any history of comorbidities at index date, the rates per 1000 person-years and incidence rate ratios for non-fatal major cardiovascular events and chronic kidney disease during follow-up in patients with T2DM, compared to non-diabetic controls, are presented in Tables 6.3 and 6.4, and Figure 6.3 separately for ethnic groups and BMI categories at index date.

Overall, the risk of developing MACE in patients with T2DM, compared to non-diabetic controls, were similar for WEs (95% CI of IRR: 1.29, 1.38) and ACs (95% CI of IRR: 1.34, 2.25), but significantly higher for SAs (95% CI of IRR: 1.56, 2.22) compared to WEs (Table 6.3).

The risk of developing MACE was significantly higher for overweight (95% CI of IRR: 1.50, 2.46) and obese (95% CI of IRR: 1.49, 2.43) SAs compared to their WE counterparts (95% CI of IRR: 1.29, 1.42 in overweight; 1.29, 1.43 in obese). However, similar risk levels were observed for WEs and SAs who were normal weight (Figure 6.3A, Table 6.3).

White European patients with T2DM had similar rates of MACE (range of 95% CI of rate/1000 person-years: 10.55, 14.66, Table 6.3) across all BMI level, and these rate estimates were almost two-fold higher compared to that across all adiposity levels in ACs (range of 95% CI of rate/1000 person-years: 2.96, 8.78) and SAs (range of 95% CI of rate/1000 person-years: 4.69, 12.91, Table 6.3).

6.4.4 The incidence of chronic kidney disease (Stage 3 and above) during follow-up

Across all BMI categories, the rates of CKD were consistently higher among WE cases (range of 95% CI of IR: 12.89, 19.73) and controls (range of 95% CI of IR: 6.31, 8.48), compared to AC cases (range of 95% CI of IR: 3.04, 10.89) and controls (range of 95% CI of IR: 2.52, 7.20), and SA cases (range of 95% CI of IR: 2.66, 9.21) and controls (range of 95% CI of IR: 1.11, 3.54, Table 6.4). While obese WEs with T2DM had significantly lower CKD incidence rate compared patients with BMI < 30 kg/m^2, the observed CKD incidence rates were similar across all BMI groups in WEs without diabetes. The incidence rates for CKD were similar across all BMI categories among AC and SA cases. Obese SAs with diabetes had almost half the incidence rate for CKD (IR: 3.9) compared to ACs (IR: 7.3) and about one fourth compared to WEs (IR: 13.4).

The risk of developing CKD in normal weight and obese patients with T2DM, compared to non-diabetic controls, was significantly higher among WEs only (Figure 6.3B). However, overweight individuals with T2DM had significantly higher and similar risk of developing CKD (range of 95% CI of IRR: 1.5, 3.4), across ethnic groups (Figure 6.3B, Table 6.4).

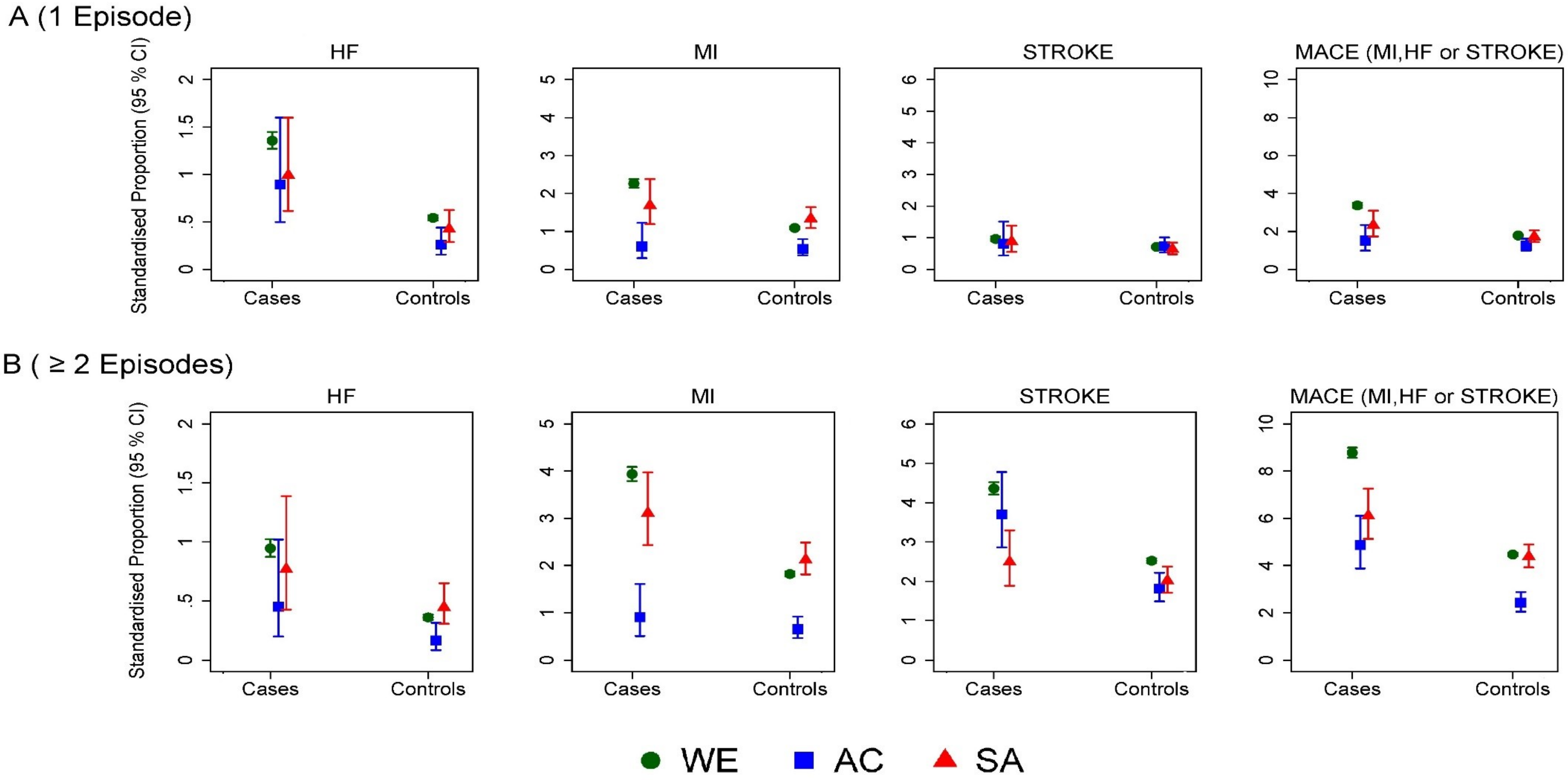


Figure 6.1: Age-sex standardised proportions [% (95 CI)] of macrovascular diseases at diagnosis for patients with T2DM and their matched controls, separately for each ethnic group.

*[**(A)** The proportion of patients with at least one episode of a macrovascular event at diagnosis; **(B)** The proportion of patients with two or more episodes of macrovascular disease events at diagnosis. [HF: Heart failure; MACE: Three (3) point major cardiovascular event defined as the occurrence of myocardial infarction, heart failure or stroke before diagnosis]. WE: White European; AC: African-Caribbean; SA: South Asian.]*

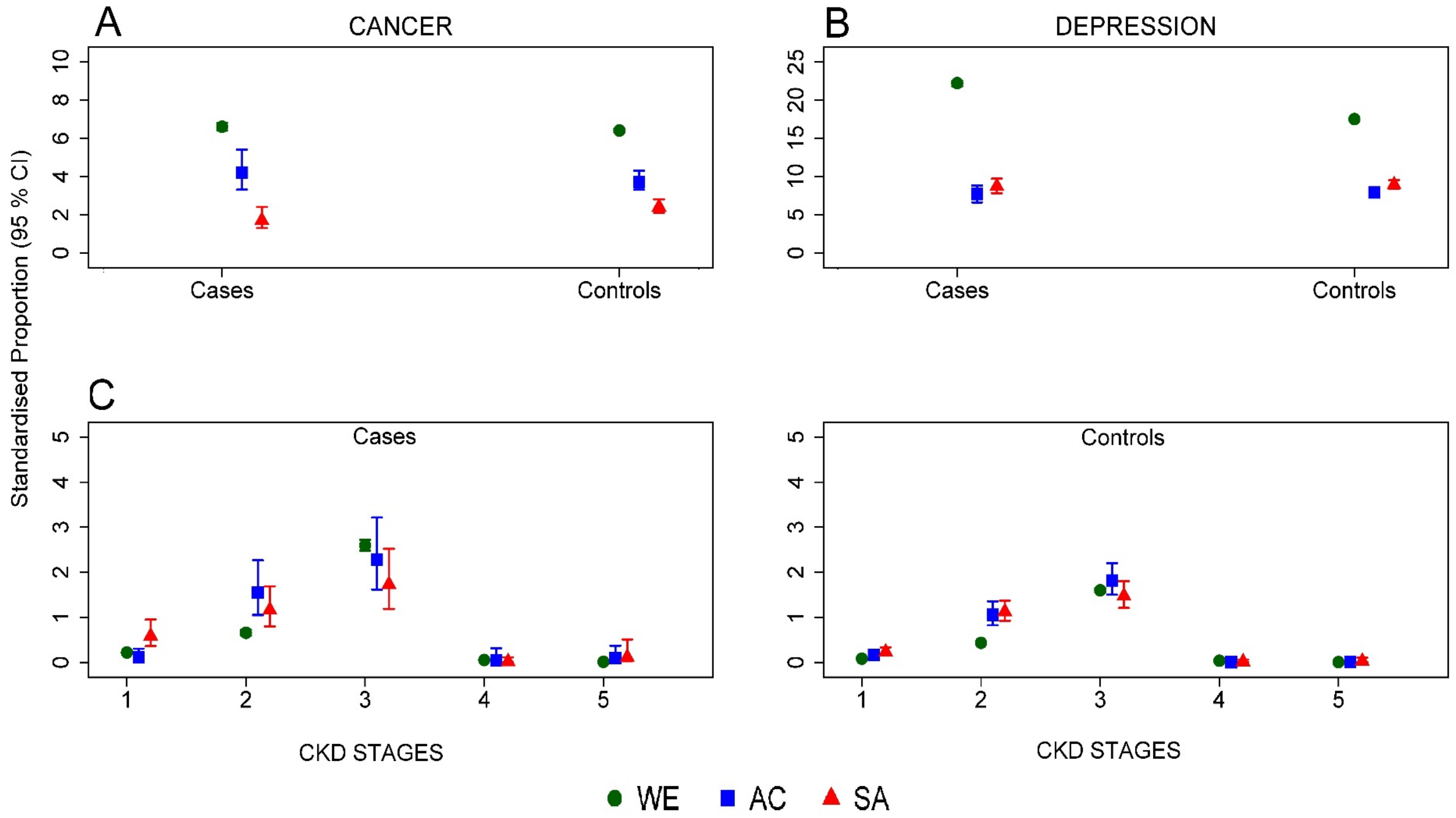


Figure 6.2: Age-sex standardised proportions [% (95 CI)] of selected non-cardiovascular diseases at diagnosis for patients with T2DM and their matched controls, separately for each ethnic group.

[(A) Proportion of patients with at cancer at diagnosis; **(B)** Proportion of patients with depression at diagnosis; **(C)** Proportion of patients with CKD (stage 1 to 5) at diagnosis. CKD: Chronic kidney disease; WE: White European; AC: African-Caribbean; SA: South Asian]

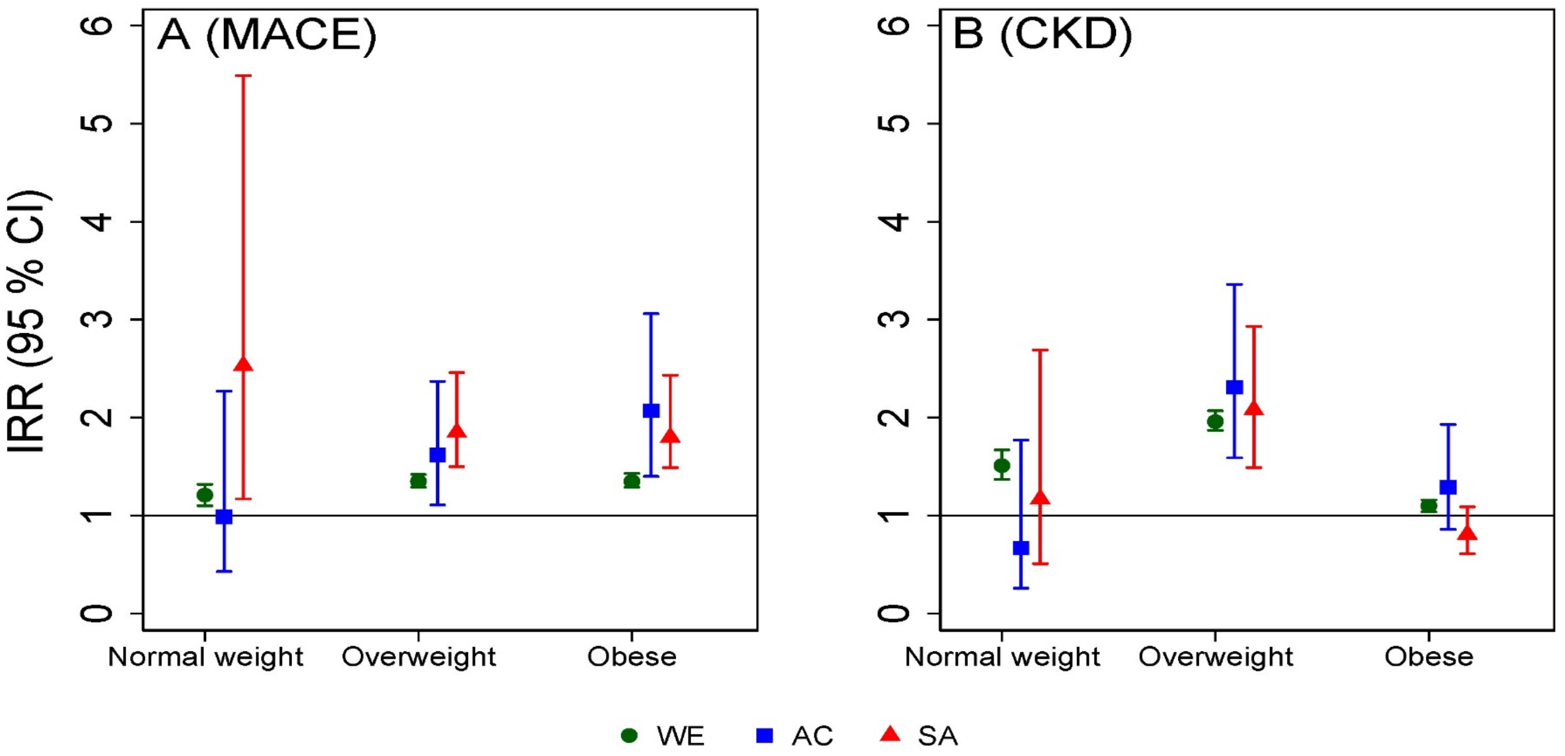


Figure 6.3: Adjusted incidence rate ratios [IRR (95% CI)] for MACE, and CKD in T2DM cases vs. matched non-diabetic controls without established comorbidities at index date.

[Data are presented separately by ethnicity for each BMI category at index date. WE: White European; AC: African-Caribbean; SA: South Asian]

Table 6.3: Incidence rates, and adjusted incidence rate ratios (95% CI) for major cardiovascular events (myocardial infarction, heart failure or stroke) in T2DM cases and matched non-diabetic controls without established comorbidities at index date.

	T2DM			**Non-diabetic controls**			
	Follow-up §	Events (%)	IR (95% CI)	Follow-up §	Events (%)	IR (95% CI)	IRR (95% CI) ¶
White European (WE)		**n=42,219**			**n=189,606**		
All WE	8 (4,11)	3378(8)	11.24(10.87,11.63)	8(5,11)	10854(6)	7.88 (7.73, 8.03)	1.33 (1.29,1.38)
Normal weight	7 (4,11)	252(1)	12.96 (11.46,14.66)	7 (4, 10)	1145(1)	8.18 (7.72, 8.67)	1.20 (1.10, 1.32)
Overweight	8 (4,11)	762(2)	11.58 (10.79,12.43)	8(5,11)	8809(5)	8.01 (7.85, 8.18)	1.35 (1.29,1.42)
Obese	8 (4,11)	2364(6)	11.00 (10.55,11.43)	7 (4,10)	900(<1)	6.51 (6.10, 6.95)	1.35 (1.29,1.43)
African-Caribbean(AC)		**n=3,043**			**n=12,570**		
All AC	7 (4,10)	107(4)	5.23 (4.33, 6.32)	7 (4,10)	238(2)	2.82 (2.48, 3.20)	1.74 (1.34,2.25)
Normal weight	7 (3, 9)	9(<1)	4.57 (2.38, 8.78)	6 (4, 9)	24(<1)	3.79 (2.54, 5.65)	0.99 (0.43,2.27)
Overweight	6 (3,9)	20(1)	4.58 (2.96, 7.10)	7 (4,10)	189(2)	2.76 (2.39, 3.18)	1.62 (1.11, 2.37)
Obese	7 (4,11)	78(3)	5.53 (4.43, 6.90)	6 (3, 9)	25(<1)	2.62 (1.77, 3.9)	2.07 (1.40, 3.06)
South Asian (SA)		**n=5,131**			**n=20,861**		
All SA	7 (3,10)	213(4)	6.38 (5.58, 7.30)	7 (4,10)	410(2)	2.98 (2.71, 3.28)	1.86 (1.56, 2.22)
Normal weight	6 (3, 9)	15(<1)	7.78 (4.69,12.91)	6(4,10)	30(<1)	3.76 (2.63,5.37)	2.53 (1.17, 5.49)
Overweight	6 (3, 10)	53(1)	6.45 (4.93,8.44)	7 (4,10)	316(2)	3.00 (2.69, 3.35)	1.85 (1.50, 2.46)
Obese	7 (4,10)	145(3)	6.24 (5.30,7.34)	5 (3, 8)	64(<1)	2.63 (2.06, 3.36)	1.80 (1.49, 2.43)

§: Median (Q1, Q3), ¶: Incident rate ratios (IRRs) were adjusted for age, sex, smoking status (never, current, or ex-smoker), deprivation score (i.e lowest affluence to highest affluence), baseline systolic blood pressure.

The follow-up period was from 2000 to 2014.

Three (3) point major cardiovascular event defined as the occurrence of myocardial infarction or heart failure or stroke during follow-up.

IR: Incidence rates per 1000 person-years.

IRR: Incidence rate ratio

Data are presented for all subjects, and separately by BMI categories at index date.

Table 6.4: Incidence rates, and adjusted incidence rate ratios (95% CI) for chronic kidney disease (stage ≥ 3) in T2DM cases and matched non-diabetic controls without established comorbidities at index date.

		T2DM			**Non-diabetic controls**		
	Follow-up §	Events (%)	IR (95% CI)	Follow-up §	Events (%)	IR (95% CI)	IRR (95% CI) ¶
White European (WE)		**n=42,219**			**n=189,606**		
All WE	8 (4,11)	4574(11)	14.39 (13.98,14.81)	8(5,11)	9571(5)	6.68 (6.55, 6.82)	1.47 (1.42,1.52)
Normal weight	7 (4,11)	370(1)	17.82(16.10, 19.73)	7 (4, 10)	1045(1)	7.19 (6.77, 7.64)	1.51 (1.37,1.67)
Overweight	8 (4,11)	1168(3)	16.72(15.79, 17.71)	8(5,11)	7389(4)	6.46 (6.31, 6.61)	1.96 (1.87,2.07)
Obese	8 (4,11)	2119(5)	13.36(12.89, 13.84)	7 (4,10)	1137(1)	8.00 (7.55, 8.48)	1.10 (1.04,1.16)
African-Caribbean(AC)		**n=3,043**			**n=12,570**		
All AC	7 (4,10)	152(5)	7.25 (6.18, 8.50)	7 (4,10)	270(2)	3.16 (2.80,3.56)	1.56 (1.20,2.03)
Normal weight	7 (3, 9)	11(<1)	5.49 (3.04, 9.91)	6 (4, 9)	33(<1)	5.12 (3.64, 7.20)	0.67 (0.26,1.77)
Overweight	6 (3, 9)	35(1)	7.82 (5.61,10.89)	7 (4,10)	201(2)	2.89 (2.52, 3.32)	2.31 (1.59,3.36)
Obese	7 (4,11)	106(3)	7.31 (6.05, 8.85)	6 (3, 9)	36(<1)	3.74 (2.70, 5.19)	1.29 (0.86,1.93)
South Asian (SA)		**n=5,131**			**n=20,861**		
All SA	7 (3,10)	146(3)	4.23 (3.60,4.98)	7 (4,10)	317(2)	2.27 (2.04, 2.54)	1.17 (0.95,1.44)
Normal weight	6 (3, 9)	10 (<1)	4.95(2.66,9.21)	6(3,10)	15 (<1)	1.85(1.11,3.06)	1.17 (0.51,2.69)
Overweight	6 (3, 10)	42 (<1)	4.95(3.66,6.69)	7 (3,10)	233 (1)	2.18(1.92,2.48)	2.08(1.49, 2.93)
Obese	7 (4,10)	94 (<1)	3.92(3.20,4.80)	5 (3, 8)	69 (<1)	2.80(2.21,3.54)	0.81(0.61,1.09)

§: Median (Q1, Q3), ¶: Multivariate incident rate ratios (IRRs) were adjusted for age, sex, smoking status (never, current, or ex-smoker), deprivation score (i.e lowest affluence to highest affluence), baseline systolic blood pressure.
The follow-up period was from 2000 to 2014.
IR: Incidence rates per 1000 person-years.
IRR: Incidence rate ratio;
Data are presented for all subjects, and separately by BMI categories at index date.

6.5 Discussion

This longitudinal case-control study of patients with newly diagnosed T2DM and their matched non-diabetic controls evaluated the prevalence of comorbidities at diagnosis of T2DM and the risk of developing long-term major cardiovascular and renal complications by BMI categories in different ethnic groups. There are several important findings from our study. Firstly, the relationship between obesity and risk of MACE /CKD does not appear to be linear. Secondly, at all levels of BMI, diabetes is associated with a significantly greater risk of MACE. Thirdly, there are important distinctions between the ethnic groups, with South Asians showing greater susceptibility to MACE and CKD even at lower BMI levels.

Obesity is a major risk factor for T2DM and is an independent risk factor for cardiovascular disease (CVD) as well as CKD [215,216]. Few studies, however, have explored the relationship between levels of adiposity and CVD in patients with T2DM and any underlying differences between ethnic groups given their differential susceptibility to T2DM. The large size of our cohort matched with a non-diabetic control population has allowed us to not only compare the effects of obesity on people with and without diabetes within each ethnic group but also to examine the differences between ethnic groups.

The independent effect of BMI on CVD risk has been confirmed in several population studies. Moreover, the linearity of this relationship has been shown in both Caucasian and Asian populations. In a study involving the Asian population, the risk of CVD increased significantly with each 2 kg/m^2 increase in BMI [217]. In patients with diabetes, however, this relationship is less clear and existing data suggest that the relationship may not be linear [218]. In our study, we did not find a linear relationship between BMI and CVD or between BMI and CKD. On the contrary, our data show that patients with diabetes have the same or even greater degree (in the case of SAs) of risk even when they are of normal weight. The absence of this linear relationship between BMI and CVD may be due to the fact that the mechanisms by which BMI and diabetes influence CVD risk are different. Alternatively, the higher burden of other known risk factors for CVD (i.e., hypertension, dyslipidaemia and insulin resistance) seen in patients with diabetes could have a greater impact on the overall CVD risk thus mitigating the effects of obesity. In this context, it is worth noting that interventions in patients with diabetes targeting weight loss have been less successful in lowering cardiovascular (CV) risk [219].

Across all ethnic groups, diabetes was associated with greater risk of MACE. This relationship did not change with levels of adiposity, except in ACs, suggesting that in some ethnic groups diabetes confers excess risk of MACE. These findings are not surprising given that patients with diabetes have a significantly greater burden of CV risk factors and are likely to be exposed to these risk factors for a much longer time. Similar trends were observed in relation to CKD, except in SAs, where the overall risk of CKD amongst diabetic and non-diabetic controls was similar in the overweight group, diabetes was associated with increased risk. Our data show that in addition to the elevated HbA1c, a greater proportion of patients with diabetes had poorly controlled blood pressure, elevated triglycerides and more likely to be obese or overweight than their non-diabetic counterparts. Despite the adverse risk profile, the use of cardio and reno-protective agents such as statins and ACE inhibitors was low suggesting there may have been opportunities for better control of risk factors. It must, however, be noted that these figures date back to the year 2000 and that management of these known risk factors has improved considerably since then [220].

Although there are many common features, our data has highlighted important differences between ethnic groups. As expected, SAs were significantly younger than WEs and ACs whereas, WEs were more likely to have a diagnosis of cancer or depression and had higher systolic blood pressure levels. The overall IR for MACE and CKD was significantly greater amongst WEs compared to ACs or SAs and this risk was evenly distributed amongst all levels of adiposity in WEs. On the other hand, the risk of MACE and CKD was greater for SAs who were either normal and/or overweight when compared to WEs. We have previously shown that SAs develop diabetes much earlier and at significantly lower BMI than other ethnic groups [41]. It is possible that exposure to diabetes at a much younger age may result in adverse vascular profile which in turn influences the risk of MACE and CKD. It is well known that SAs have excess visceral adiposity which may contribute to the overall metabolic risk in this ethnic group even at lower levels of BMI. It is also possible that BMI may not be an ideal measure of adiposity in SA and other measures such as waist/hip ratio could instead be more appropriate when assessing adiposity in this ethnic group [221]. While there is a need for better understanding of the effects of adiposity on MACE / CKD in different ethnic groups, the clear message from this study is to recognise that SAs have a disproportionate risk of cardiovascular disease even at normal BMI.

Although the large multi-ethnic cohort and the availability of longitudinal data for a population sharing the same health care system have been the strengths of this study, it has some limitations. First, there was a small number of events in BMI subgroups among African-Caribbean and South Asians. Second, we have in this study used BMI as a measure of obesity and it can be argued that BMI is not an ideal measure of obesity especially in certain ethnic groups such as SA. We are aware that this may have limited our ability to explore the relationship between adiposity and the risks of MACE/CKD. On the other hand, BMI is a commonly used measure of obesity and is well recorded than other measures such as waist/hip or waist/height ratios. Further, we have used ethnic-specific cut-offs for BMI [104] to provide as reliable an estimate of adiposity as possible.

Our understanding of the differences between ethnic groups towards susceptibility to diabetes has improved considerably in recent times. The findings of this study add to this knowledge and provide a greater understanding of the relationship between levels of adiposity and diabetes complications in different ethnic groups. The results of this study should enable clinicians to better diagnose and manage diabetes amongst people of different ethnicities.

6.6 List of abbreviations

BMI: Body mass index; WE: White European; AC: African-Caribbean; SA: South Asian; T2DM: Type 2 diabetes mellitus; UK: United Kingdom; IRR: Incident rate ratio; MACE: Major cardiovascular event; CKD: Chronic Kidney Disease; UKPDS: UK Prospective Diabetes Study; MI: Myocardial infarction; THIN: The Health Improvement Network; GP: General practice; EMR: Electronic medical records; NHS: National Health Service; BNF: British National Formulary; ATC: Anatomical Therapeutic Chemical; T1DM: type 1 diabetes mellitus; HbA1c: glycated haemoglobin; SBP: systolic blood pressure; DBP: diastolic blood pressure; LDL: low density lipoproteins; HDL high density lipoproteins; WHO: World Health organisation; CPM: Cardio-protective medications; HF: heart failure.

.

6.7 Declarations

6.7.1 Ethics approval and consent to participate

Formal access to the THIN database has been obtained from the Independent Scientific Review Committee for the THIN database (Protocol Number: 15THIN030) and the study was approved by the Institutional Review Board of QIMR Berghofer Medical Research Institute.

6.7.2 Consent for publication

Not applicable

6.7.3 Availability of data and material

The datasets generated and/or analysed during the current study are available from the corresponding author on reasonable request.

6.7.4 Competing interests

S. K. P. has acted as a consultant and/or speaker for Novartis, GI Dynamics, Roche, AstraZeneca, Guangzhou Zhongyi Pharmaceutical and Amylin Pharmaceuticals LLC. He has received grants in support of investigator and investigator-initiated clinical studies from Merck, Novo Nordisk, AstraZeneca, Hospira, Amylin Pharmaceuticals, Sanofi Aventis and Pfizer. K. K. has acted as a consultant, speaker and has received funds for research from Novartis, Novo Nordisk, Sanofi Aventis, Lilly, Merck Sharp & Dohme, Janssen, Astra Zeneca, Boehringer Ingelheim, Roche and Servere. W. H. has received travel grants, research grants and consultancy fees from Novo Nordisk, Eli Lilly, Sanofi, MSD, Jansen, Astra Zeneca and Boehringer Ingelheim. S. B. has received research grants from The Binding Site and Novo Nordisk (UK) and has received honoraria for speaking and for participation in advisory boards of Astra Zeneca, Boehringer Ingelheim, Janssen, Sanofi Aventis, Eli Lilly, MSD and Novo Nordisk. E. S. O. A. has no conflicts of interest to declare

6.7.5 Funding

E.S.O.A. was supported by QIMR Berghofer International Ph.D. Scholarship and The University of Queensland International Scholarship. No separate funding was obtained for this study. University of Melbourne and Melbourne EpiCentre gratefully acknowledges the support from the Australian Government's National Collaborative Research Infrastructure Strategy (NCRIS) initiative through Therapeutic Innovation Australia and the research project funding from the National Health and Medical Research Council of Australia (Project Number: GNT1063477).

6.7.6 Authors' contributions

S.K.P. and K.K. developed the concept of study and E.S.O.A and S.K.P. contributed to the study design. E.S.O.A. conducted the data extraction, data manipulation, statistical analyses and developed the first draft of the manuscript. E.S.O.A., S.B., W.H., K.P., K.K., and S.K.P. contributed to the finalization of the manuscript. S.K.P. contributed to the statistical analyses and had full access to all the data in the study and is the guarantor, taking responsibility for the integrity of the data and the accuracy of the data analysis.

6.7.7 Acknowledgements

Not applicable

Chapter 7: Weight loss and mortality risk in patients with different adiposity at diagnosis of type 2 diabetes: a longitudinal cohort study

This body of this chapter contains one published paper that evaluates whether the weight change pattern or weight loss (an indication of potential latent disease) prior to diagnosis of diabetes can explain the observed higher mortality risk in patients with normal body weight at diagnosis compared to those with obesity at diagnosis. The citation of the published paper is as follows:

Owusu Adjah ES, Samanta M, Shaw JE, Majeed A, Khunti K, Paul SK. Weight loss and mortality risk in patients with different adiposity at diagnosis of type 2 diabetes: a longitudinal cohort study. *Nutrition & diabetes* 2018;**8**(1):37

All the listed have agreed to the inclusion of this published scholarly work in this thesis and the statement of my contribution to the authorship of this published scholarly work is included below:

Contributor	Statement of contribution
Owusu Adjah Ebenezer S (Candidate)	Conceived the idea and was responsible for the primary design of the study. Conducted the data extraction from THIN database. Responsible for data manipulation, aggregation, and transformation in SAS. Performed the statistical analyses in STATA and interpretation of results. Developed first draft and contributed towards finalisation of the manuscript.
Samanta Mayukh	Contributed to the interpretation of the results and manuscript finalisation.
Shaw Jonathan E	Contributed significantly to the study design, interpretation of results and manuscript finalisation.
Majeed Azeem	Contributed to the interpretation of the results and manuscript finalisation.
Khunti Khunti	Contributed to the interpretation of the results and manuscript finalisation.
Paul Sanjoy K	Conceived the idea and was responsible for the primary design of the study. Contributed to the statistical analyses. Developed first draft and contributed towards finalisation of the manuscript.

7.1 ABSTRACT

Background: Undiagnosed comorbid diseases that independently lead to weight loss before type 2 diabetes mellitus (T2DM) diagnosis could explain the observed increased mortality risk in T2DM patients with normal weight.

Objectives: To evaluate the impact of weight change patterns before the diagnosis of T2DM on the association between body mass index (BMI) at diagnosis and mortality risk.

Methods: This was a longitudinal cohort study using 145 058 patients from UK primary care, with newly diagnosed T2DM from January 2000. Patients aged 18-70, without established disease history at diagnosis (defined as the presence of cardiovascular diseases, cancer, and renal diseases on or before diagnosis) were followed up to 2014. Longitudinal 6-monthly measures of body weight three years before (used to define groups of patients who lost body weight or not before diagnosis) and two years after diagnosis were obtained. The main outcome was all-cause mortality.

Results: At diagnosis, mean (SD) age was 52 (12) years, 56% were male, 52% were current or ex-smokers, mean BMI was 33 kg/m^{2,} and 66% were obese. Normal weight and overweight patients experienced a small but significant reduction in body weight six months before diagnosis. Among all categories of obese patients, consistently increasing body weight was observed within the same time window. Among patients who did not lose body weight pre-diagnosis (n=117 469), compared with the grade 1 obese, normal weight patients had 35% (95% CI of HR: 1.17, 1.55) significantly higher adjusted mortality risk. However, among patients experiencing weight loss before diagnosis (n=27 589), BMI at diagnosis was not associated with mortality risk (all p>0.05).

Conclusions: Weight loss before the diagnosis of T2DM was not associated with the observed increased mortality risk in normal weight patients with T2DM. This emphasises the importance of addressing risk factors post diagnosis for excess mortality in this group.

7.2 INTRODUCTION

Recent epidemiological studies have raised the controversy of the *obesity paradox* in type 2 diabetes mellitus (T2DM). While some studies reported significantly higher mortality risk in those with normal body weight at diagnosis of T2DM, compared to those with obesity [7,12,19,25,26,79,82,87], others could not find such evidence [22,24]. Latent diseases that independently lead to weight loss before T2DM diagnosis could explain the observed increased mortality risk in those with normal weight [222]. This is particularly important, because the undiagnosed conditions leading to weight loss may also increase the risk of developing or being diagnosed with diabetes, but may be clinically diagnosed after the diagnosis of diabetes, and falsely appear as a consequence of diabetes. In this context, evaluation of weight change before and after diagnosis of diabetes along with comorbidities is crucial. However, data on these aspects at pollution level is scarce.

Only a few epidemiological studies have evaluated body weight or BMI before and after diagnosis of diabetes [223-227]. However, these studies were limited by small sample sizes [224-226], measurement of weight at only two-time points usually many years apart [223,227], and they did not include evaluation of the mortality risk in association with weight change. To the best of our knowledge, none of the studies that have evaluated the obesity paradox in T2DM patients conducted a dedicated analysis of body weight changes pre- and post-diagnosis of T2DM. With a large cohort of patients with incident T2DM, the aims of this real-world primary care based longitudinal study were to evaluate: (1) body weight changes over 3 years pre-diagnosis of T2DM, (2) body weight changes over 24 months post diagnosis of T2DM, stratified by BMI category at time of T2DM diagnosis separately for those who have died and those who have not, and (3) the impact of weight change pattern before diagnosis on the association of BMI at time of T2DM diagnosis with mortality risk.

7.3 MATERIALS AND METHODS

7.3.1 Participants

Primary care patients with T2DM were identified from Read codes or the date of the first prescription for an anti-diabetes drug (ADD), through various steps of a clinically guided iterative machine learning algorithm based on regression methodologies [228] (see expanded procedure in Chapter 3). The algorithm identified a cohort of 406,098 patients with T2DM between January 1990 and September 2014. The cohort of T2DM patients for this study (1) were newly diagnosed with T2DM from January 2000 onwards, (2) had a minimum follow-up of 1 year, (3) had complete data on age, sex, and BMI ($\geq$ 15 kg/m^2), and (4) were without an established diagnosis of cardiovascular diseases (CVD), chronic kidney disease (CKD) or cancer at time of diagnosis of T2DM (Figure 7.1). Those with Read codes for type 1 diabetes mellitus (T1DM) or gestational diabetes, those who received insulin as the first ADD, and those who had undergone bariatric surgery before or after diagnosis were excluded.

7.3.2 Study variables

Patients with CVDs, CKD (any stage), and cancer with dates of diagnoses after the T2DM diagnosis date were identified using Read codes. A composite variable for CVD (any CVD) was defined as the occurrence of angina, myocardial infarction, coronary artery disease (including bypass surgery and angioplasty), heart failure or stroke. Complete records on the prescriptions of different classes of ADDs, antihypertensive drugs, weight lowering drugs, anti-depressant drugs, and lipid-modifying drugs were extracted along with the dates of prescriptions.

Information on deaths with dates and possible cause of death were also extracted. Time to a specific disease event or death was calculated as the time from the diagnosis of T2DM to the first occurrence of the disease event or date of death respectively. Patients who were still alive at the end of the study (September 2014) or had dropped out were censored on the end date or drop out date.

Demographic, clinical and laboratory data extracted at time of T2DM diagnosis included: smoking status, deprivation score (a socioeconomic status measure based on residential address[126]), ethnicity, body weight, BMI, glycated haemoglobin (HbA_{1c}), systolic blood pressure (SBP), diastolic blood pressure (DBP), low density lipoproteins (LDL-C), high density lipoproteins (HDL-C), and triglycerides. BMI categories at diagnosis of T2DM were defined as normal weight (18.5-24.99 kg/m^2), overweight (25-29.99 kg/m^2), grade 1 obese (30-34.99 kg/m^2), grade 2 obese (35-39.99 kg/m^2) and grade 3 obese ($\geq$ 40 kg/m^2)[104].

Longitudinal measures of body weight and BMI in the 36 months before and 24 months after the T2DM diagnosis date were extracted and arranged in six-monthly windows. All available measures on or within three months before the T2DM diagnosis date were considered as the baseline measures. If more than one measurement existed within this interval, the closest to the T2DM diagnosis date was taken.

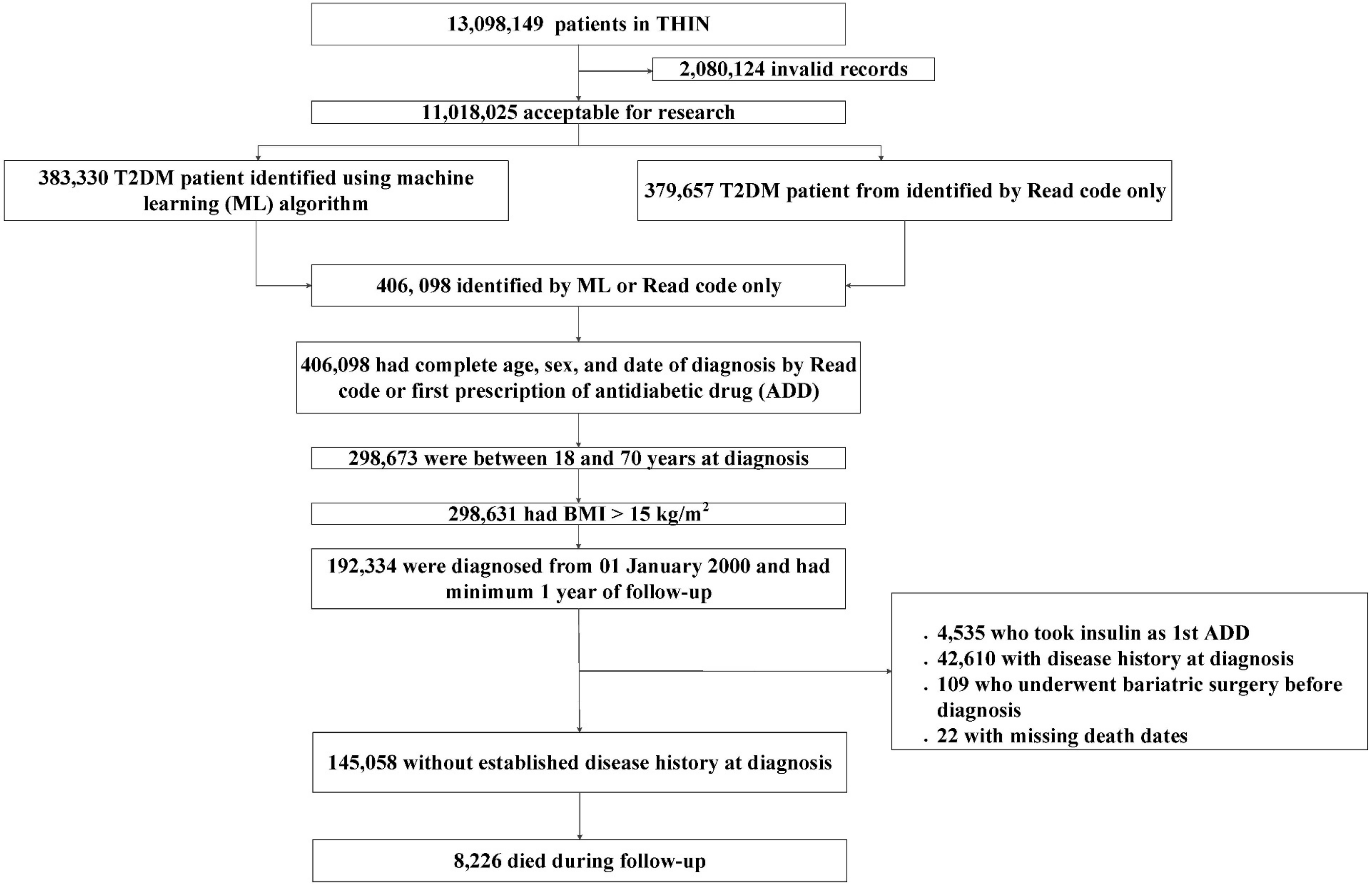

13,098,149 patients in THIN
2,080,124 invalid records
11,018,025 acceptable for research
383,330 T2DM patient identified using machine learning (ML) algorithm
379,657 T2DM patient from identified by Read code only
406, 098 identified by ML or Read code only
406,098 had complete age, sex, and date of diagnosis by Read code or first prescription of antidiabetic drug (ADD)
298,673 were between 18 and 70 years at diagnosis
298,631 had BMI > 15 kg/m2
192,334 were diagnosed from 01 January 2000 and had minimum 1 year of follow-up
. 4,535 who took insulin as 1st ADD
. 42,610 with disease history at diagnosis
. 109 who underwent bariatric surgery before diagnosis
. 22 with missing death dates
145,058 without established disease history at diagnosis
8,226 died during follow-up

Figure 7.1: Study cohort selection flowchart.

To identify patients who lost body weight (LBW) by at least 2 kg before the diagnosis of T2DM, two different approaches were used based on 6 possible longitudinal body weight measures over 36 months as follows:

i. Approach 1: the body weight measure in the 6 months prior to diabetes diagnosis was ≥ 2 kg less than the mean of the 5 possible prior measures;

ii. Approach 2: the body weight measure in the 6 months prior to diabetes diagnosis was ≥ 2 kg less than all of the other weight measures in the three years before diagnosis.

Those who did not lose body weight or increased body weight during 36 months before diagnosis were identified as “no weight loss” (NWL).

The study protocol was approved by the Independent Scientific Review Committee for the THIN database (Protocol Number: 15THIN030) and the Institutional Review Board of QIMR Berghofer Medical Research Institute.

7.3.3 Statistical Methods

The summary statistics were presented by number (percentage), mean (SD) or median (first quartile, third quartile), and by survival status (alive or dead) where appropriate. Age-weighted rates (per 1000 person-years) for CVD, CKD, cancer, hypertension during follow-up were estimated by BMI categories and mortality status. Age-weighted mortality rates were also computed for patients under each BMI category.

Weight trajectories before and after diagnosis were evaluated by fitting a generalised linear model under general estimating equations setup, with unstructured covariance. Separate analyses were conducted for each BMI category. Among patients who did not die within two years post diagnosis of T2DM or remained censored, the unadjusted and adjusted mean (95% confidence intervals, CI) of longitudinal 6-monthly measures of body weight before and post T2DM diagnosis were estimated respectively. Adjustment factors for post-diagnosis weight trajectory were age, sex, smoking status, the incidence of CVD, CKD or cancer, and the use of insulin, GLP-1 receptor agonists or sulfonylurea during two years of follow-up.

Under the hypothesis that the pattern of weight change before T2DM diagnosis could be a modifying factor on the association between BMI categories at the time of diagnosis and mortality risk, a multivariate stratified Cox regression model was fitted separately for patients under different weight loss pattern before the diagnosis of diabetes (i.e., LBW and NWL groups). Under the null hypothesis of no difference in risk patterns by BMI categories at diagnosis of T2DM, we aim to evaluate the alternative hypothesis of risk difference in patients with normal body weight compared to those with grade 1 obesity (BMI 30-34.9 kg/m^2) at 5% level of significance. The hazard ratio (HR) for all-cause mortality was calculated for each BMI category using individuals with grade 1 obesity as the reference group. The adjustment factors were - age, sex, deprivation score, and smoking status at diagnosis; use of insulin, oral anti-diabetes drugs, and cardio-protective medications during follow-up as fixed covariates. Age groups (defined as 18-40, 41-50, 51-60, 61-70 years) at T2DM diagnosis were used as the stratification factor. Robust estimates of hazard ratios (95% CI) were obtained, and Bayesian information criteria (BIC) was used to compare the model fits. The proportional hazards assumption was assessed using scaled Schoenfeld residuals, and variables that violated the proportional hazards assumptions (incidence of cancer, any CVD or CKD during follow-up) were included in the model as time-varying covariates. All primary analyses were conducted using the imputed body weight data, with additional analyses based on complete cases for sensitivity analyses.

In sensitivity analyses for mortality, an extended model was fitted incorporating measures of HbA$_{1c}$, SBP, LDL-C, HDL-C, and triglyceride at baseline. Other sensitivity analyses involved (1) excluding the time-varying covariates that violated the proportionality assumption (see Appendix D); (2) excluding current and ex-smokers; (2) including patients who never developed cancer, (3) possible interaction of age groups and BMI categories (stratified by weight loss patterns). Data extraction from the THIN database was conducted using SAS® 9.4 (SAS Institute), and statistical analyses were performed using STATA version 14 MP, at a 2-tailed α level of 0.05**.**

Data access

Data **were** made are available to the corresponding author (SKP) under a licensing agreement from IMS Health UK (now IQVIA). All data access enquiries should be forwarded to Professor Sanjoy K. Paul.

Code availability

The programming code is available from ESOA

7.4 RESULTS

7.4.1 Cohort characteristics at diagnosis

In this cohort of 145,058 patients with incident T2DM, the mean (SD) age at diagnosis was 52 (12) years, 56% were male, 52% were current or ex-smokers, and 66% were obese. Among patients who were censored at the end of study (still alive or moved out of practice), the mean (SD) age at diagnosis was 51 (12) years with 26% aged above 60 years, 56% were men, and the proportion of patients in the normal weight, overweight and obese categories were 7%, 27%, and 67% respectively. Over a median follow-up of 8 years, those who died were significantly older (mean age: 60 years vs 51 years), had a higher prevalence of current and ex-smokers (63% vs 51%), and had a higher SBP level (mean: 144 mmHg vs 139 mmHg) at diagnosis compared to those who were censored (Table 7.1). Across all BMI categories the incidence rates (per 1000 person-years) for any CVD, cancer, and CKD were significantly higher among those who died compared to those censored (Table 7.2, all $p<0.01$).

7.4.2 Weight changes before the diagnosis of T2DM

A small but significant drop in body weight during the six months before diagnosis of T2DM was observed in patients belonging to the normal and overweight categories at diagnosis, although a stable body weight trajectory was observed during 30 months before that time window (Figure 7.2A). Among all categories of obese patients, consistently increasing body weight was observed before the diagnosis of diabetes, followed by a sharp drop in body weight during the 12 months after the diagnosis of T2DM. The proportions of patients who lost body weight in the 36 months before diagnosis date in the normal weight, overweight, grade 1 obese, grade 2 obese, and grade 3 obese categories were 28%, 21%, 18%, 17% and 16% respectively (Table 7.3).

7.4.3 Weight change after diagnosis of T2DM

Among normal weight patients who died, there was no indication of any weight loss during 24 months before death, while a consistently increasing body weight trajectory was observed among those who did not die (Figure 7.3). With an initial significant decline in body weight within six months post diagnosis of diabetes, overweight, grade 1 and 2 obese patients slowly gained weight over the following 18 months, with no difference in the longitudinal patterns by mortality status. For grade 3 obesity, those who died had a higher weight throughout the post-diagnosis period than those who remained alive. The trajectories of body weight were similar for both imputed data and the complete case analyses.

Table 7.1: Baseline characteristics of patients with T2DM and without a history of CVD, CKD and cancer at the time of T2DM diagnosis, by mortality status.

	Mortality status at study end date		
	Alive	**Dead**	**All**
Patients, number (%) *	136,832 (94)	8,226 (6)	145,058 (100)
Age in years, mean (SD)†	51 (12)	60 (9)	52 (12)
Age group *			
≤40 years	25,693 (19)	304 (4)	25,997 (18)
41-50 years	33,313 (24)	824 (10)	34,137 (24)
51-60 years	43,069 (32)	2,420 (29)	45,489 (31)
61-70 years	34,757 (26)	4,678 (57)	39,435 (27)
Male *	76,054 (56)	4,890 (60)	80,944 (56)
Smoking status *			
Current smoker	28,875 (21)	2,385 (29)	31,260 (22)
Ex-smoker	40,821 (30)	2,805 (34)	43,626 (30)
Never smoked	66,182 (48)	2,823 (34)	69,005 (48)
Townsend deprivation *			
Least deprived	21,542 (16)	1,443 (18)	22,985 (16)
Most deprived	26,678 (20)	1,400 (17)	28,078 (19)
Weight in kg, mean (SD) †	93.4 (19.3)	90.4 (19.1)	93.2 (19.3)
BMI (kg/m^2),† mean (SD)	32.7 (6.3)	31.8 (6.4)	32.7 (6.3)
BMI categories *			
Underweight	208 (<0.1)	52 (1)	260 (<0.1)
Normal weight	9,770 (7)	764 (9)	10,534 (7)
Overweight	36,404 (27)	2,444 (30)	38,848 (27)
Grade 1 Obese	52,400 (39)	3,159 (38)	55,559 (39)
Grade 2 Obese	22,790 (17)	1,054 (13)	23,844 (16)
Grade 3 Obese	15,260 (11)	753 (9)	16,013 (11)
SBP (mm/Hg) †	139 (17)	144 (18)	140 (17)
SBP ≥ 140 *	64,881 (47)	4,973 (60)	69,854 (48)

HbA_{1c}, mmol/mol[†]	69 (18.6)	68 (17.5)	69 (18.6)
$HbA_{1c} \geq$ 58 mmol/mol [*]	96,567 (70)	5,956 (72)	102,523 (71)
LDL-C, mmol/L [†]	3.26 (0.75)	3.15 (0.67)	3.23(0.75)
HDL-C, mmol/L [†]	1.16 (0.28)	1.21 (0.31)	1.16(0.28)
Triglycerides, mmol/L[‡]	1.90 (1.50, 2.36)	1.87 (1.5, 2.29)	1.90 (1.50, 2.35)
Follow-up (years) [‡]	7 (4, 11)	11 (8, 13)	8 (4, 11)

[*]: n (%)

[†]: mean (SD)

[‡]: median (Q1, Q3)

Abbreviations: BMI: Body mass index; SPB: Systolic blood pressure; LDL-C: Low-density lipoprotein cholesterol; HDL-C: High-density lipoprotein cholesterol

Table 7.2: Proportion and event rates per 1000 person-years (95% CI) for patients without established disease history at diagnosis by baseline BMI categories.

	Normal weight (10,534)		**Overweight (38,848)**		**Grade 1 Obese (55,559)**		**Grade 2 Obese (23,844)**		**Grade 3 Obese (16,013)**	
	Alive (9,770)	**Dead (764)**	**Alive (36,404)**	**Dead (2,444)**	**Alive (52,400)**	**Dead (3,159)**	**Alive (22,790)**	**Dead (1,054)**	**Alive (15,260)**	**Dead (753)**
MI *	186(2)	30(4)	712(2)	151(6)	922(2)	184(6)	322(1)	56(5)	197(1)	23(3)
MI †	2.6	4.0	2.6	6.1	2.4	5.8	2.0	5.3	1.9	
	(2.2,3.0)	(2.8,5.8)	(2.4,2.9)	(5.23,7.20)	(2.2, 2.5)	(5.0, 6.7)	(1.8, 2.3)	(4.0,6.8)	(1.7, 2.2)	3.1 (2.0,4.6)
HF *	110(1)	34(4)	380(1)	179(7)	563(1)	204(6)	289(1)	92(9)	234(2)	80(11)
HF †	1.5	4.6	1.4	7.3	1.4	6.5	1.8	8.9	2.3	11.2
	(1.3, 1.8)	(3.3, 6.4)	(1.3,1.5)	(6.3,8.5)	(1.3, 1.6)	(5.6,7.4)	(1.6,2.0)	(7.2,10.9)	(2.0, 2.6)	(9.0, 14.0)
Stroke *	308(3)	58(8)	1103(3)	252(10)	1428(3)	258(8)	531(2)	93(9)	345(2)	43(6)
Stroke †	4.3	8.1	4.1	10.6	3.7	8.3	3.3	8.9	3.4	5.9
	(3.8, 4.8)	(6.2,10.4)	(3.9, 4.3)	(9.3,12.0)	(3.47,3.9)	(7.3, 9.3)	(3.0,3.6)	(7.3,10.9)	(3.0,3.7)	(4.4,7.9)
Any CVD *	1125(12)	184(24)	4319(12)	687(28)	5406(10)	800(25)	2115(9)	290(28)	1292(8)	171(23)
Any CVD †	16.7	29.5	17.1	34.2	14.7	29.8	13.8	32.8	13.2	26.7
	(15.8,17.7)	(25.5,34.1)	(16.6,17.6)	(31.8,36.9)	(14.3,15.1)	(27.8,32.0)	(13.2,14.4)	(29.2,36.8)	(12.5,13.9)	(23.0, 31.0)
CKD *	970(10)	126(16)	3748(10)	404(17)	4318(8)	436(14)	1978(9)	222(21)	1277(8)	155(21)
CKD †	14.5	19.2	14.8	18.3	11.7	14.9	13.1	24.5	13.3	24.5
	(13.6,15.4)	(16.1,22.9)	(14.4,15.3)	(16.6, 20.1)	(11.3,12.0)	(13.6, 16.3)	(12.5,13.7)	(21.5, 27.9)	(12.6,14.0)	(20.9, 28.7)
Cancer *	498(5)	271(35)	2043(6)	930(38)	2473(5)	1068(34)	962(4)	368(35)	623(4)	204(27)
Cancer †	7.0	46.0	7.7	46.5	6.4	40.1	6.0	41.2	6.1	31.1
	(6.4,7.7)	(40.8,51.8)	(7.3, 8.0)	(43.6, 49.6)	(6.2, 6.7)	(37.8, 42.6)	(5.7, 6.4)	(37.2, 45.7)	(5.7,6.6)	(27.1, 35.6)

n (%); †: rate per 1000 person-years (95%CI);
MI: Myocardial infarction; HF: Heart failure; Any CVD: any cardiovascular disease defined as the occurrence of angina, MI, coronary heart disease (CHD), stroke, and HF post diagnosis of T2DM.

Table 7.3: Mortality risk by Body Mass Index (BMI) category at the time of diabetes diagnosis stratified by weight trajectory patterns prior to diagnosis.

	BMI category				
	Normal weight (N=10,534)	Overweight (N=38,848)	Grade 1 Obese (N=55,559)	Grade 2 Obese (N=23,844)	Grade 3 Obese (N=16,013)
Lost Body weight (LBW, n=27,589)					
Patients *	2,983 (28)	8,266 (21)	9,721 (17)	3,977 (17)	2,544 (16)
Deaths *	205 (2)	543 (1)	520 (1)	171 (1)	120 (1)
Person-time in years †	1,093	2,957	2,327	1,008	667
Rate per 1000 person-years	11.0 (9.6,12.7)	11.1 (10.2,12.1)	10.6 (9.7,11.6)	8.6 (7.42, 10.1)	9.8 (8.2,11.9)
HR (95% CI) ‡	0.89 (0.65,1.22)	0.93 (0.77,1.12)	1.00 (reference)	1.01 (0.81,1.27)	1.30 (0.86,1.95)
No Weight Loss (NWL, n=117,469)					
Patients *	7,551 (72)	30,582 (79)	45,838 (83)	19,867 (83)	13,469 (84)
Deaths *	559 (5)	1,901 (5)	2,639 (5)	883 (4)	633 (4)
person-time in years †	3,069	10,981	14,443	5,411	3,826
Rate per 1000 person-years	12.4 (11.4, 13.5)	10.0 (9.5,10.4)	9.7 (9.3,10.1)	8.1 (7.6,8.7)	8.9 (8.2,9.7)
HR (95% CI) ‡	1.35 (1.17,1.55)	0.99 (0.91,1.07)	1.00 (reference)	0.95 (0.86,1.04)	1.06 (0.89,1.25)

*: Number (proportion)

†: Person-time (years) contributed by patients who died during follow-up. Follow-up period from 2000 to 2014

‡: Estimates of hazards ratios were adjusted for baseline BMI, sex, smoking status, deprivation score, insulin, oral antidiabetic drugs, cardio-protective medicine and time-varying incidence of cancer, chronic kidney disease and incidence of any cardiovascular disease using age group at baseline as a stratification factor.

Table 7.4: Mortality rates (per 1000 person-years) and risk with their 95% CIs, by BMI categories and age groups in patients without established disease history before diagnosis.

			BMI category		
	Normal weight	Overweight	Grade 1 Obese	Grade 2 Obese	Grade 3 Obese
18-40 years					
Patients *****	1644(16)	4427(11)	10390(19)	5409(23)	4071(25)
Rate/1000 person-years	2.12 (1.42, 3.16)	1.37 (1.02,1.84)	1.52 (1.27,1.81)	1.24 (0.94,1.63)	2.17 (1.70,2.79)
HR (95% CI) §	1.76 (0.96, 3.21)	1.00 (0.66,1.52)	1.00 (reference)	0.79 (0.54,1.16)	1.15 (0.59,2.25)
41-50 years					
Patients *****	2000(19)	7813(20)	12985(23)	6297(26)	5,097 (31)
Rate/1000 person-years	3.77 (2.89,4.91)	2.92 (2.51,3.40)	3.17 (2.84,3.55)	3.28 (2.79,3.86)	4.15 (3.52,4.90)
HR (95% CI) §	1.65 (1.18, 2.28)	1.12 (0.91,1.37)	1.00 (reference)	0.97 (0.78,1.21)	0.98 (0.71,1.34)
51-60 years					
Patients *****	3070(29)	12983(33)	17612(32)	7246(30)	4,584 (28)
Rate /1000 person-years	8.59 (7.47,9.88)	6.63 (6.14,7.16)	7.16 (6.72,7.63)	6.61(5.95,7.35)	9.01 (8.00,10.15)
HR (95% CI) §	1.49 (1.21,1.83)	1.05 (0.93,1.19)	1.00 (reference)	0.89 (0.77,1.03)	1.04 (0.81,1.32)
61-70 years					
Patients *****	3820(36)	13625(35)	14572(26)	4892(21)	2,470 (15)
Rate/1000 person-years	17.13 (15.68,18.72)	15.47 (14.72,16.25)	17.06 (16.29,17.87)	15.29 (14.02,16.67)	17.31 (15.39,19.47)
HR (95% CI) §	1.02 (0.85,1.22)	0.91 (0.82,1.00)	1.00(reference)	1.04 (0.91,1.18)	1.21 (0.96,1.53)

§: Estimates of hazards ratios (HR) were adjusted for baseline BMI, sex, smoking status, deprivation score, insulin, oral antidiabetic drugs, cardio-protective medicine and time-varying incidence of cancer, chronic kidney disease and incidence of any cardiovascular disease using weight loss pattern prior to diagnosis as a stratification factor.

7.4.4 Mortality rate and risk by BMI categories

Overall, 6% of the patients died during a median 8 years of follow-up (n=8,226, Table 7.1). The median follow-up time was similar among all BMI categories, and separately for patients in the NWL and LBW groups. The number, proportions and person-time (in years) of patients who died during follow-up under different BMI categories, separately for each weight change pattern before diagnosis, are presented in Table 7.4. Overall, patients with normal weight had significantly increased the adjusted risk of mortality compared to those with grade 1 obesity (Figure 7.2B).

Among patients with NWL before diagnosis (n=117,469), the age-weighted mortality rate per 1000 person–years in normal weight patients at diagnosis was significantly higher (rate=12.4; 95% CI: 11.4, 13.5) than for those who were grade 1 obese (rate= 9.7; 95% CI: 9.3, 10.1), grade 2 obese (rate 8.1; 95% CI: 7.6, 8.7) and grade 3 obese (rate 8.9; 95% CI: 8.2, 9.7) (Table 7.3). With grade 1 obese patients as reference, normal weight patients in the NWL group had 35% increased risk of mortality (Adjusted HR = 1.35; 95% CI: 1.17, 1.55; p<0.01).

For patients in the LBW group, mortality rate 1000 person-years in normal weight patients at diagnosis was not significantly higher (rate=11.0; 95% CI: 9.6, 12.7) than for those who were grade 1 obese (rate= 10.6; 95% CI: 9.7, 11.6), grade 2 obese (rate= 8.6; 95% CI: 7.4, 10.1), grade 3 obese (rate 9.8; 95% CI: 8.2, 11.9). Subsequently, there was no significant association between BMI categories and mortality risk in the LBW group (all p>0.05) (Table 7.3).

7.4.5 Sensitivity analyses

The mortality risk estimates were similar in subgroups of patients who did not develop cancer and those who never smoked. The extended risk analyses by incorporating HbA1c, blood pressure and lipids at diagnosis as covariates, also revealed similar mortality risk estimates, separately for groups of patients with and without weight loss prior to diagnosis. Sensitivity analysis with identification of weight loss by Approach 2 also provided similar results.

Compared to grade 1 obese patients, normal weight patients in the age groups 41-50 years and 51-60 years had significantly higher mortality risk by 65% (95% CI of HR: 1.18, 2.28), 49% (95% CI of HR: 1.21, 1.83) respectively. Across all age groups, grade 2 or grade 3 obese patients did not have higher mortality risk compared to grade 1 obese patients (Table 7.4).

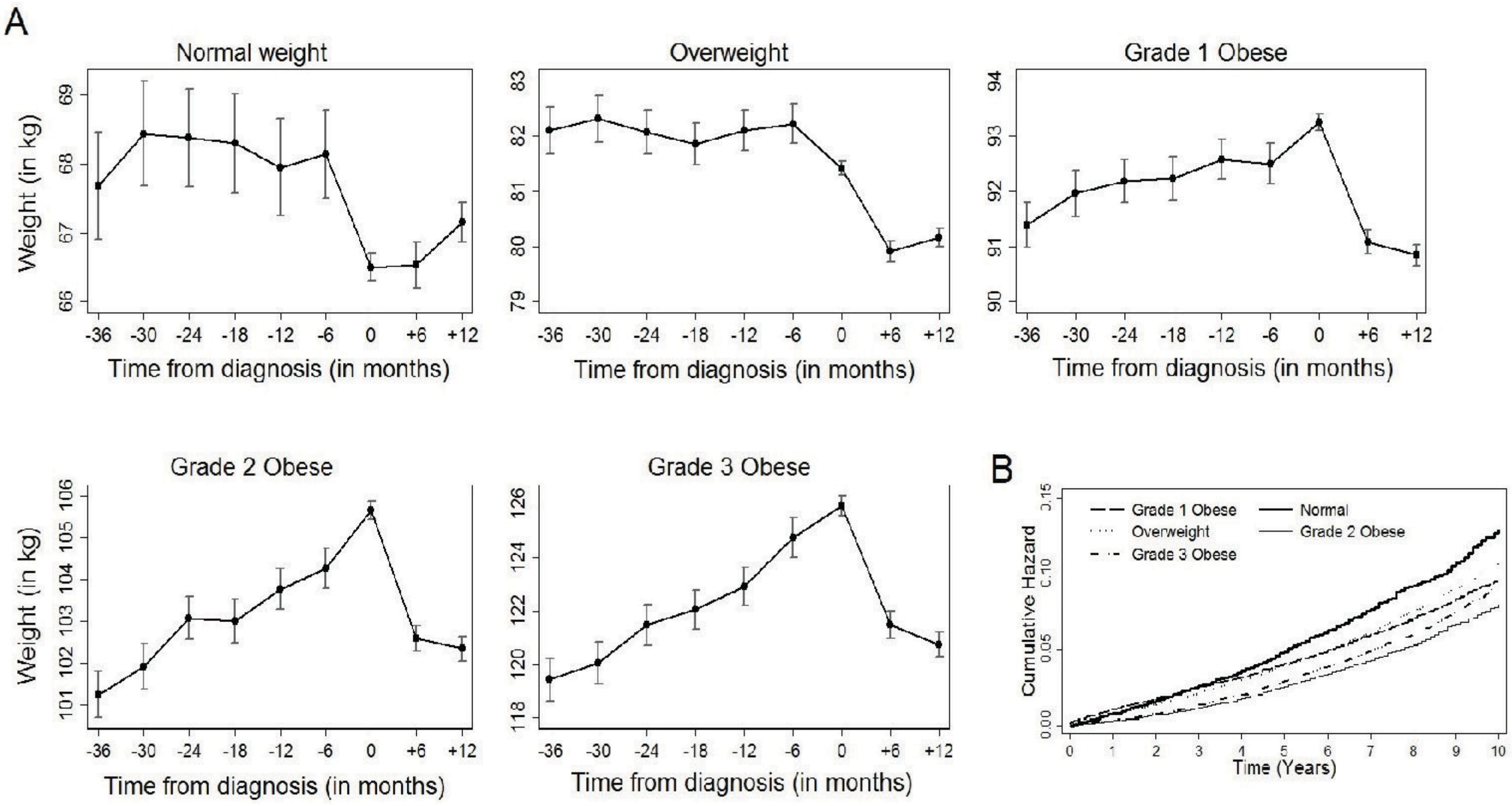


Figure 7.2: (A) Six-monthly trajectory [mean (95% CI)] of body weight (kg) over 3 years prior to diagnosis of T2DM, at diagnosis and one-year post diagnosis separately for different BMI categories at diagnosis of T2DM, for patients without a history of diseases at diagnosis. (B) The cumulative hazard function for all-cause mortality in patients without disease history, by BMI categories at diagnosis.

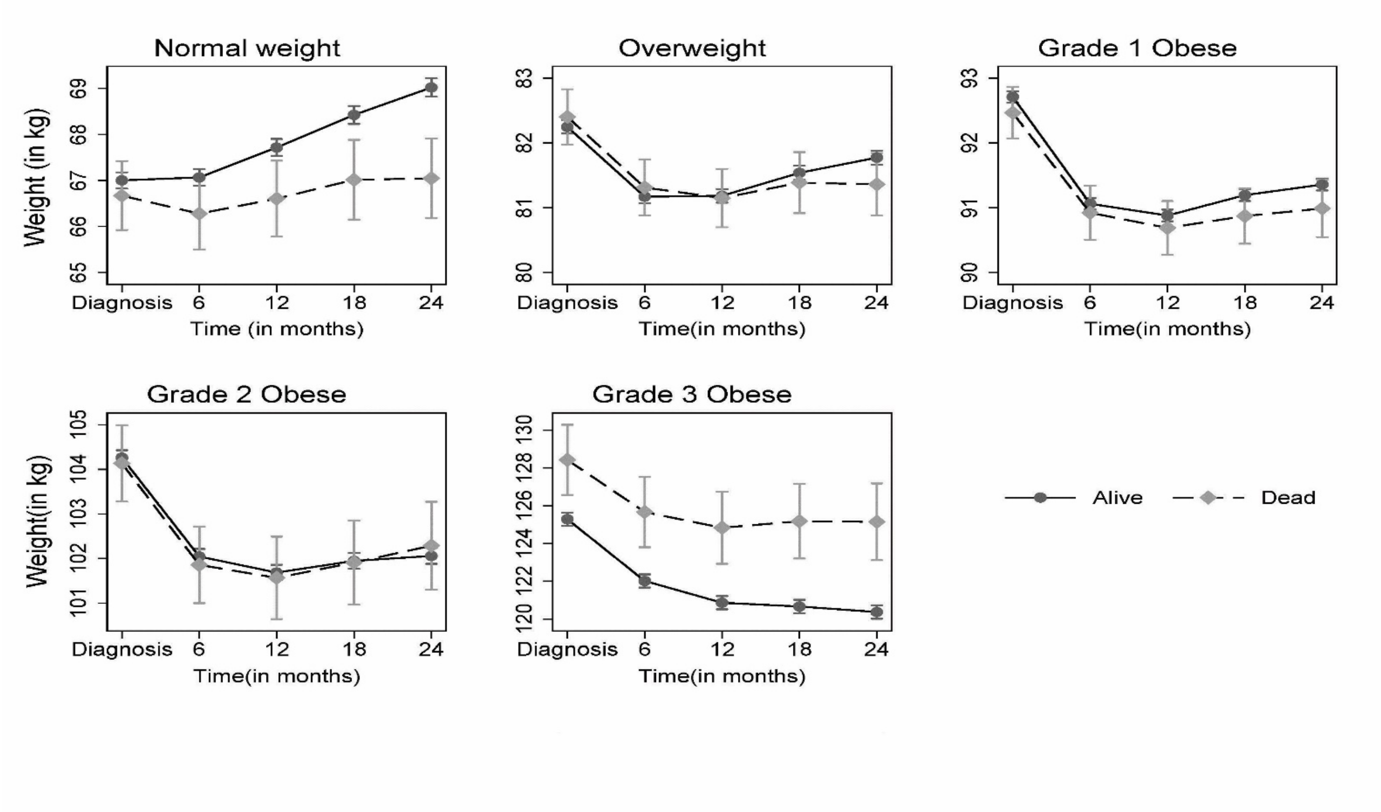


Figure 7.3: Weight (in kg) trajectory by mortality status post diagnosis of T2DM for patients without disease history.

Weight trajectory estimates were adjusted for age at diagnosis, sex, smoking status, the incidence of chronic kidney disease or cancer or any CVD, and the use of insulin or sulphonylureas or GLP1RA within 2 years of diagnosis.

7.5 DISCUSSION

In this longitudinal study of a large number of incident T2DM patients from the UK, we observed: (1) a significant drop in body weight over the 6 months before diagnosis of T2DM in normal weight and overweight patients, followed by a marginal increases in body weight post-diagnosis; (2) no significant weight change over 24 months post diagnosis among normal weight patients who died; (3) patients with normal body weight at time of T2DM diagnosis had a significantly higher adjusted rate and risk of all-cause mortality compared to grade 1 obese patients, and this was not explained by weight loss before diagnosis, (4) a significant age and BMI interaction, with elevated mortality risk for normal weight patients aged 41 – 60 years; and (5) patients with BMI ≥ 35 kg/m^2 at diagnosis did not have significantly higher mortality risk compared to grade 1 obese patients across all age groups.

One novel aspect of this study was the evaluation of 6-monthly longitudinal changes in body weight over 24 months post diagnosis of T2DM by mortality status and BMI at diagnosis. In the normal weight category, patients had an increasing weight trajectory over 24 months irrespective of mortality status, suggesting no sudden weight loss in these patients post-diagnosis. This observation coupled with the fact that underlying comorbidities/latent diseases were not over-represented in the normal weight group contradicts the assertion of possible weight loss due to underlying diseases [222]. While this study was not designed to assess the impact of lifestyle modifications on weight in patients with T2DM, the observed weight changes in overweight and obese groups post diagnosis of T2DM are consistent with previous studies that studied this effect [229,230]. We observed a marginal decrease in body weight during the six months post diagnosis of diabetes in overweight and obese patients, followed by a plateau, similar to that observed in other studies [229,230]. In the study by Aucott and colleagues, using approximately 30 000 obese or overweight Scottish adults with incident diabetes, weight change was not associated with mortality risk, while 36% reduced body weight at two years post-diagnosis [229]. Furthermore, given the adjusted trajectories of body weight by mortality status over two years across BMI categories in our study (Figure 3), weight change that occurs post diagnosis rather than pre-diagnosis is likely to be associated with long-term mortality risk.

The obesity paradox in T2DM is the phenomenon whereby significantly higher mortality risk is observed among those with normal body weight at diagnosis of T2DM, compared to those with obesity. Our finding of significantly higher mortality risk in normal weight T2DM patients at the time of diagnosis corroborates other findings and contributes to the current debate on the *obesity paradox* in T2DM [12,25,78,79,231]. We report an obesity paradox regardless of disease history at diagnosis, an observation previously reported by Thomas and colleagues [12]. Some researchers have suggested that the obesity paradox could be explained by unmeasured confounders (e.g., unrecognised underlying comorbidity/latent diseases) that lead to weight loss and are therefore over-represented in the normal weight group [110,112,232]. By considering the weight loss pattern before the diagnosis of T2DM as a potential confounder in the relationship between adiposity status at diagnosis and mortality risk, and by undertaking separate analyses for patients with and without co-morbid disease at diagnosis, our observation is unlikely to be biased by underlying diseases. Furthermore, a detailed exploration of the patterns of weight change over 24 months post diagnosis of diabetes establishes the robustness of our finding.

Our study reveals that patients who were obese at the time of T2DM diagnosis experienced a steady rise in body weight before diagnosis, an observation which is consistent with a previous study in Pima American Indians [225]. While only two studies either statistically modelled the trajectory of body weight or evaluated one-point observed weight 10 years prior to diagnosis of T2DM, our study explored the 6-monthly trajectory of observed body weight during the 36 months prior to diagnosis, accounted for prevalence of diseases, and assessed weight change over 24 months post diagnosis of diabetes [226,233]. We also note that normal weight and overweight patients experienced significant weight loss during six months before diagnosis of diabetes –a rather common, yet unexplained clinical manifestation. Our study identifies patients who consistently lost body weight and patients who remained weight neutral over three years before clinical diagnosis of diabetes. We found that, though a significantly larger proportion of the normal weight patients lost body weight before the diagnosis of T2DM, compared to overweight, grade 1 and grade 2 obese patients, weight loss before diagnosis was not associated with increased mortality in normal weight patients.

The strengths of this longitudinal study include a large number of incident T2DM patients with 8 years of median follow-up, a nationally representative cohort, a thorough assessment of the longitudinal trajectory of body weight before and post diagnosis of T2DM, and identification of weight loss patterns and comorbid conditions before diagnosis. Clinically diagnosed T2DM patients with a diagnosis from January 2000 were selected to ensure the quality of diagnosis. Age at diagnosis was also restricted to a maximum of 70 years to avoid including older patients who were already at significantly increase mortality risk. We attempted to minimise possible confounding by adjusting for several possible confounders including antidiabetic treatment, cardio-protective medications, and smoking status while evaluating the cohort with no history of major diseases at diagnosis. However, electronic health records present challenges in terms of the accuracy and completeness of the required data. The limitations of this study include non-availability of complete and reliable data on ethnicity and smoking cessation during follow-up, missing body weight data during the 36 months before diagnosis of diabetes, information on diet, exercise or weight lowering medications, and the potential residual confounders as is common in observational studies. Also, there is the potential for misdiagnosis, misclassification, and miscoding of diagnostic codes in electronic medical records [130-133]. We utilised other clinical data to minimize potential misclassification of T2DM (see Chapter 3). Although we excluded all T2DM patients, who received insulin as their first anti-diabetes drug from our study cohort, some patients with T1DM might still be misclassified as having T2DM.

7.6 CONCLUSION

In conclusion, weight loss before the diagnosis of T2DM occurred independently of established severe disease conditions and was not associated with the observed increased mortality risk in normal weight patients with T2DM. While the cause of this excess mortality in T2DM who were normal weight at diagnosis remains unclear, it may reflect differences in the aetiology of diabetes in normal weight people and emphasises the importance of addressing risk factors for excess mortality in this group.

Chapter 8: Ethnicity-Specific association of BMI levels at diagnosis with cardiovascular disease and all-cause mortality risk

This body of this chapter contains one published paper that evaluates the potential role of ethnicity in the observed increased mortality in individuals with normal weight (BMI: 18.5-25 kg/m^2). The citation of the published paper is as follows:

Owusu Adjah ES, Ray KK, Paul SK. Ethnicity-specific association of BMI levels at diagnosis of type 2 diabetes with cardiovascular disease and all-cause mortality risk. *Acta Diabetologica* 2018;**56**(1):87-96.

All the listed have agreed to the inclusion of this published scholarly work in this thesis and the statement of my contribution to the authorship of this published scholarly work is included below:

Contributor	Statement of contribution
Owusu Adjah Ebenezer S (Candidate)	Conceived the idea and was responsible for the primary design of the study. Conducted the data extraction from THIN. Performed data manipulation, aggregation, and transformation in SAS. Conducted statistical analyses in STATA and interpreted the results. Developed first draft and contributed towards finalisation of the manuscript.
Ray K	Contributed to the interpretation of the results and manuscript finalisation.
Paul SK	Responsible for the primary design of the study. Contributed to the statistical analyses. Developed first draft and contributed towards finalisation of the manuscript.

8.1 ABSTRACT

Aim: To evaluate the risk of CVD and all-cause mortality at different BMI levels in conjunction with weight change prior to the diagnosis of T2DM in a multi-ethnic population.

Materials and Methods: Longitudinal study of 51,455 patients with T2DM and without a history of comorbid diseases at diagnosis. Weight changes prior to the diagnosis of T2DM were evaluated and the risk of CVD and all-cause mortality at different BMI levels among three ethnic groups estimated.

Results: White Europeans (n=40,575), African-Caribbeans (n=3,605), and South Asians (n=7,275) were 52 , 49, and 47 years old with a mean BMI of 33.0, 32.0, and 30.0 kg/m^2 at diagnosis, respectively. Among White Europeans, normal weight patients developed CVD significantly earlier by 0.5 years (95% CI: 0.1, 0.9 years; p=0.018) compared to obese patients (mean time to CVD 4.6 years). Furthermore, those with normal body weight at diagnosis were significantly more likely to die earlier by 0.6 years (95% CI: 0.03, 1.2 years; p=0.037) among White Europeans and by 2.5 years (95% CI: 0.3, 4.6 years; p=0.023) among South Asians compared to their respective obese patients.

Conclusion: This study suggests a paradoxical association of BMI with cardiovascular and mortality risks in different ethnic groups. Normal weight White Europeans and South Asians appear to have significantly higher mortality risk compared to those who were obese at the time of T2DM diagnosis.

Keywords: Body mass index; Type 2 Diabetes; Cardiovascular Disease; Obesity; Race and Ethnicity; Weight Change Pattern.

8.2 INTRODUCTION

Recent studies have reported an inverse association of body mass index (BMI) with mortality risk among adults with type 2 diabetes mellitus (T2DM), where patients who were normal weight [BMI 18.5 - 24.9 kg/m^2] at diagnosis had significantly elevated mortality risk compared to their obese counterparts [BMI $\geq$ 30 kg/m^2] [12,19,25,26,87]. While the explanation for this phenomenon, referred to as the obesity paradox in T2DM remains unclear, weight loss before the diagnosis of T2DM as a result of underlying/undiagnosed medical condition was postulated as one of the possible reasons [110,112,232]. However, an analysis of body weight changes over 3 years before diagnosis in patients with T2DM under different BMI categories have shown otherwise [234]. It is possible that ethnicity might play an essential role in understanding the underlying mechanism, as the distribution of adiposity levels in relation to cardiovascular disease (CVD) and mortality risk has been shown to be different for different ethnic groups [28,29,186,199].

Previous studies have evaluated the incidence of CVDs either in different ethnic groups [204,205,207,209,235] or in relation to BMI [236,237]. However, these studies did not evaluate the possible difference in the BMI related risk paradigm in different ethnic groups. Among Asians, a pooled analysis of 20 prospective cohort studies in Asia reported increased cardiovascular mortality risk at lower BMIs [238]. However, this study was conducted in the general population, adjusting for diabetes status where appropriate. Wright and colleagues (2016) reported significantly lower mortality risk in South Asian and African-Caribbean individuals with T2DM compared to White Europeans [239]. However, this UK primary care-based study did not evaluate the interplay of BMI in this context.

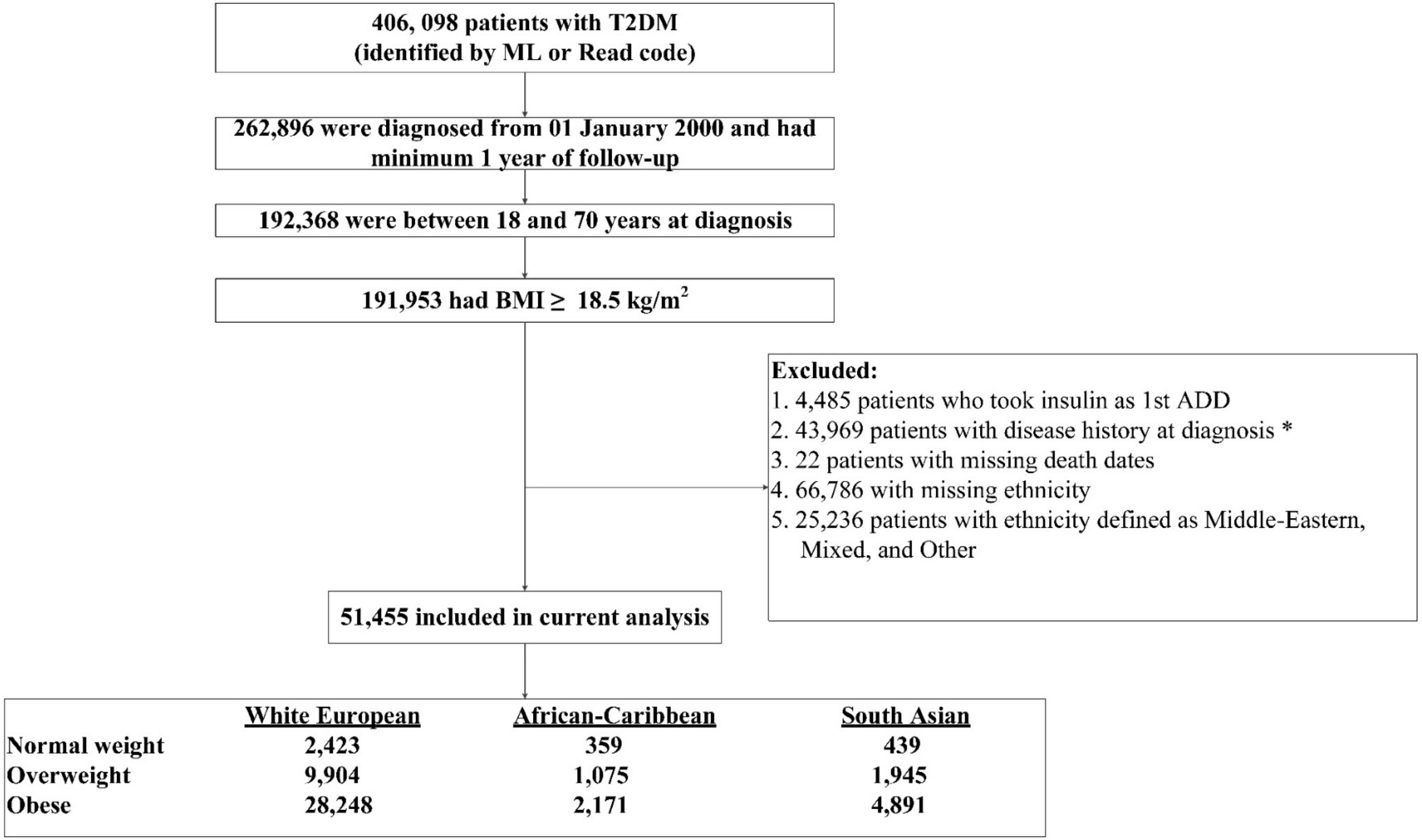
To the best of our knowledge, only one study has examined the modifiable association of ethnicity on the observed phenomenon of the obesity paradox in T2DM. Kokkinos and colleagues [79] used data from two Veteran Affairs Medical Centres in the US to assess the association between BMI, fitness, and mortality in African-Americans and Caucasians. However, this study was based on only male patients, and the BMI measures were not evaluated at the time of diagnosis of diabetes.

A better understanding of the potential role of ethnicity in the obesity paradox in both male and female patients with T2DM is important as this would enable clinicians to better manage diabetes amongst patients of different ethnicity and adiposity. Therefore, to address these knowledge gaps, the aim of this study was to use a cohort of incident T2DM patients from United Kingdom primary care database, to evaluate for each ethnic group, (1) the CVD and mortality rate in each BMI category, by weight change pattern before diagnosis and (2) the association of BMI categories at diagnosis with CVD and mortality risk, controlling for weight change pattern before diagnosis and other risk factors.

8.3 MATERIALS AND METHODS

8.3.1 Identification of T2DM cohort

The data for this study were obtained from The Health Improvement Network (THIN) database. The detailed description of this database has been previously presented [41], and formal access to the database has been obtained from the Independent Scientific Review Committee for the THIN database (Protocol Number: 15THIN030). Patients diagnosed with T2DM between January 1990 and September 2014 (n=406,098) were identified using a robust machine learning algorithm, which uses a combination of Read codes [128], anti-diabetes medications, and lifestyle modification interventions [228]. Those included in this study satisfied the following criteria: (1) complete data on age (18 – 70 years), sex, BMI (≥ 18.5 kg/m^2) and date of diagnosis of T2DM from January 2000 with a minimum 1 year of follow-up), (2) ethnicity identified as White European, African-Caribbean or South Asian, and (3) no history of CVD, renal diseases, cancer, retinopathy, neuropathy or bariatric surgery at diagnosis. South Asians were defined as patients with Indian, Pakistani, Sinhalese, and Bangladeshi origin, while African-Caribbeans were defined as patients with Black-African and Caribbean origin. White, European, Caucasian, and New Zealand European were defined as White Europeans. Those with Read codes for type 1 diabetes mellitus (T1DM) or gestational diabetes, and those who received insulin as the first antidiabetic drug (ADD, highly likely to be patients with T1DM) were excluded. A final cohort of 51,455 patients with T2DM was used for this study. Given the fact that only a proportion of the patients in the THIN database have ethnicity record, to explore the potential selection bias in this study, we have provided a flow-chart and table of basic statistics for patients from the database under the inclusion-exclusion criteria for this study (Figure 8.1 and Table 8.4).



406, 098 patients with T2DM
(identified by ML or Read code)

262,896 were diagnosed from 01 January 2000 and had minimum 1 year of follow-up

192,368 were between 18 and 70 years at diagnosis

191,953 had BMI ≥ 18.5 kg/m^2

Excluded:
1. 4,485 patients who took insulin as 1st ADD
2. 43,969 patients with disease history at diagnosis *
3. 22 patients with missing death dates
4. 66,786 with missing ethnicity
5. 25,236 patients with ethnicity defined as Middle-Eastern, Mixed, and Other

51,455 included in current analysis

	White European	**African-Caribbean**	**South Asian**
Normal weight	**2,423**	**359**	**439**
Overweight	**9,904**	**1,075**	**1,945**
Obese	**28,248**	**2,171**	**4,891**

Figure 8.1: Cohort selection flowchart.

*: History of disease defined as clinical diagnosis of CVD or renal diseases or cancer at diagnosis or retinopathy or neuropathy or bariatric surgery before the diagnosis of T2DM; **ML**: machine learning; **ADD**: Anti-diabetes drug.

8.3.2 Demographic and longitudinal measurements

Data on deprivation score (based on residential address) was extracted where available, and the smoking status for individuals were classified as current, ex, or never smokers. Longitudinal anthropometric, clinical and laboratory measurements including BMI, body weight, glycated haemoglobin (HbA_{1c}), blood pressure and lipids were extracted for all patients. All available measures at or within three months before the diagnosis of T2DM were considered as the baseline measures. If more than one measurement existed within this interval, the closest to the T2DM diagnosis date was taken. After that, longitudinal measures before and after the T2DM diagnosis were arranged in six-monthly windows. BMI categories for White Europeans and African-Caribbeans were defined as normal weight (18.5-24.9 kg/m^2), overweight (25-29.9 kg/m^2), and obese (≥ 30 kg/m^2). For South Asians, BMI in the ranges 18.5-22.9, 23-27.4, ≥ 27.5 kg/m^2 were used to define normal weight, overweight and obese patients respectively [104].

As weight loss before clinical diagnosis of T2DM is a common clinical manifestation, it was hypothesized that a weight loss of at least 2 kg before the diagnosis of diabetes was clinically significant [240]. Therefore, using 6 possible longitudinal body weight measures over 36 months before diagnosis, we classified patients who lost body weight (LBW) by at least 2 kg before diagnosis (if average of 5 prior measurements minus the body weight measure in the 6 months prior to diabetes diagnosis was ≥ 2 kg) and those who did not lose body weight (NWL) – i.e., they remained on the same level or increased body. Complete records on the prescriptions for different classes of ADDs, antihypertensive drugs, weight lowering drugs, anti-depressant drugs, and lipid-modifying drugs were extracted along with the dates of prescriptions.

8.3.3 Mortality and comorbidity data

Records of CVDs, renal diseases (including chronic kidney disease (CKD)), and cancer with dates of diagnoses before and after T2DM diagnosis date were obtained. Information on deaths with dates and possible reasons were extracted. A composite variable for CVD (any CVD) was defined as the occurrence of angina, myocardial infarction, coronary artery disease (including bypass surgery and angioplasty), heart failure, or stroke. Patients with a recorded diagnosis of cancer, any CVD, retinopathy, neuropathy, or renal diseases before the T2DM diagnosis date were considered to have a relevant disease history. Time to a specific disease event and time to death were calculated as the time from T2DM diagnosis date to the first occurrence of the disease event and date of death respectively. Patients who were still alive at the end of the study data collection (September 2014) or dropped out were censored on the respective end date or drop out date.

8.3.4 Statistical analysis

The basic summary statistics were presented by number (percentage), mean (SD) or median (first quartile, third quartile), by ethnicity as appropriate. Among patients without disease history who were identified to have lost body weight (LBW) or not (NWL), age-weighted CVD and ACM rates (per 1000 person-years) were estimated by BMI categories for each ethnic group. Cox proportional hazard regression is a widely used approach to analyse survival time data because of its flexible semi-parametric property. However, the key assumption of the proportional hazards regression model is unlikely to be true for patients with incident T2DM under different adiposity levels. To account for the inherent differences in risk factors between the defined BMI categories and the fact that risk may not be proportional, treatments effects modelling approach was used to provide robust inferences on the time to cardiovascular events or ACM. This modelling approach uses the potential outcomes or counterfactual framework to allow comparison of survival time for CVD and all-cause mortality for patients with different BMI categories, separately for each ethnic group. Briefly, given an observed outcome (Y_0), for a patient with normal weight, the potential outcome or the counterfactual (Y_1) for this same patient is the outcome if the patient had belonged to another BMI category and vice versa. Therefore, the average of the difference between the observed outcomes given a specific BMI category and the potential outcome is the average treatment effect [i.e., average treatment effect (ATE) = average (Y_1-Y_0)] [113-116]. Since the outcome of interest is survival time, a survival model with an inverse-probability weight estimator was used to estimate average time to events for each BMI category. Variables that were conditioned on include sex, weight change pattern before diagnosis, age at diagnosis, smoking status, the incidence of cancer and renal diseases post-diagnosis, and receipt of lifestyle advice before and after diagnosis. Statistical analyses were performed using STATA version 15 MP, at a 2-tailed α level of 0.05.

8.4 RESULTS

8.4.1 Basic demographic and clinical characteristics

In this study of 40,575 White Europeans, 3,605 African-Caribbean, and 7,275 South Asians adults with T2DM, the median follow-up time was 7 years for all three ethnic groups. The demographic and clinical profiles of these patients at diagnosis of T2DM in the three ethnic groups are presented in Table 1. South Asians had the clinical diagnosis of T2DM at a younger age (47 years) and at lower BMI (30.0 kg/m^2) compared to White Europeans (age: 52 years, BMI: 33 kg/m^2) and African-Caribbeans (age: 49 years, BMI: 32.0 kg/m^2). White Europeans had the highest proportion (58%) of ever-smokers (defined as current or ex-smokers) and a higher proportion of patients with systolic blood pressure above 140 mmHg (Table 8.1).

While the proportions of obese patients were 70%, 60%, and 67% in the White European, African-Carribean and South Asian patients respectively, African-Caribbeans had higher proportions of patients in the normal weight (10%) and overweight (30%) groups, as well as highest LDL-cholesterol levels (129 mg/dl) at diagnosis compared to White European and South Asian patients. Furthermore, White Europeans were more likely to receive lifestyle advice before (32%) and after (69%) diagnosis of T2DM compared to African-Caribbeans (25% and 59%) and South Asians (27% and 60%) respectively (Table 8.1).

The distribution of selected clinical characteristics among T2DM patients with no disease history at diagnosis, separately for each ethnic group within the three defined BMI categories are presented in Table 8.2. African-Caribbeans had similar levels of ever-smokers across BMI categories, while South Asians who were normal weight at diagnosis had a significantly higher proportion of ever-smokers (35%) compared to their counterparts who were overweight (26%) and obese (25%) at diagnosis. Furthermore, the proportion of ever-smokers among White Europeans who were normal weight at diagnosis (60%), was significantly higher compared to their White Europeans obese (57%) counterparts (Table 8.2).

Across the three ethnic groups, the proportion of patients with clinically diagnosed hypertension was smaller in normal weight patients compared to obese patients. African-Caribbeans and South Asians who were normal weight at diagnosis underwent more lifestyle intervention than their obese colleagues. The use of statins was significantly higher in normal weight patients compared to obese patients across ethnic groups (Table 8.2). While the proportion of normal weight patients who lost at least 2 kg weight loss before diagnosis was almost double that of obese patients across BMI categories, a similar proportion of patients experienced this weight loss before diagnosis across the three ethnic groups.

Table 8.1: Basic clinical and demographic characteristics of patients with T2DM by ethnicity.

	White European	**African-Caribbean**	**South Asian**
Patients †	40,575	3,605	7,275
Age at diagnosis (years) ‡	52 (12)	49 (11)	47 (12)
age group †			
≤40	6,936 (17)	923 (26)	2,228 (31)
41-50	9,515 (24)	1,129 (31)	2,073 (29)
51-60	12,971 (32)	902 (25)	1,848 (25)
61-70	11,153 (28)	651 (18)	1,126 (16)
Male †	22,534 (56)	1,851 (51)	3,911 (54)
Smoking status, †			
Never smoker	17,132 (42)	2,496 (69)	5,371 (74)
Current smoker	9,718 (24)	461 (13)	952 (13)
Ex-smoker	13,649 (34)	640 (18)	917 (13)
Weight (kg) ‡	95 (19)	89 (16)	82 (15)
BMI (kg/m^2) ‡	33 (6)	32 (5)	30 (5)
BMI categories †			
Normal weight	2,423 (6)	359 (10)	439 (6)
Overweight	9,904 (24)	1,075 (30)	1,945 (27)
Obese	28,248 (70)	2,171 (60)	4,891 (67)
SBP (mmHg) ‡	139 (16)	137 (16)	134 (16)
SBP ≥ 140 mmHg †	19,578 (48)	1,516 (42)	2,442 (34)
HBA$_{1c}$ (%)‡	9 (2)	9 (2)	9 (2)
HBA$_{1c}$ ≥ 7.5%	28,510 (70)	2,651 (74)	5,262 (72)
LDL (mg/dl) ‡	125 (29)	129 (29)	125 (29)
HDL (mg/dl) ‡	45 (11)	47 (11)	44 (10)
Triglycerides (mg/dl) §	170 (136-212)	127 (97-159)	160 (126-201)
LBW prior to diagnosis †	7,595 (18)	662 (18)	1,225 (18)
Lifestyle advice †			
Before diagnosis	12,861 (32)	909 (25)	1,983 (27)
After diagnosis	27,855 (69)	2,130 (59)	4,352 (60)
Follow-up (years) §	7 (4-11)	7 (4-10)	7 (4-10)

†: n (%); ‡: mean (SD); §: median (Q1, Q3); **BMI**: Body mass index; **SPB**: Systolic blood pressure; **LDL**: Low-density lipoprotein cholesterol; **HDL**: High-density lipoprotein cholesterol; **LBW**: Lost at least 2kg body weight before diagnosis;

8.4.2 Cardiovascular disease and mortality event rates

To avoid potential bias resulting from already existing severe diseases that may independently induce weight loss in patients, cardiovascular and mortality risk assessments were carried out excluding patients with clinically diagnosed cancer, any CVD, retinopathy, neuropathy, or chronic kidney disease (CKD) at diagnosis of T2DM. The age-weighted CVD and all-cause mortality rates per 1000 person-years (95 % CI), by BMI categories and weight change pattern, prior to diagnosis in patients without disease history at diagnosis, separately for the three ethnic groups are presented in Figures 8.2 and 8.3. Among White Europeans, CVD event rates per 1000 person-years were significantly higher in normal weight patients (rate: 23.7; 95% CI: 21.3, 26.5) compared to obese patients (rate: 20.3; 95% CI: 19.6, 21.0), independent of weight change pattern before diagnosis. In no other ethnic group did CVD event rates vary across different BMI categories and weight change pattern before diagnosis. However, the CVD event rates in White Europeans with normal weight were significantly higher than the rates in African-Caribbeans with normal weight (rate: 11.6; 95 % CI: 7.6, 18.6), and similar to the rates in South Asians with normal weight (rate: 21.3; 95 % CI: 16.1, 28.7) (Figure 8.2).

Irrespective of weight change pattern before diagnosis (i.e. loss of at least 2 kg of body weight or not), mortality rates per 1000 person-years were significantly higher among White Europeans with normal weight (rate: 12.2; 95% CI: 10.6, 14.1) compared to obese White Europeans (rate: 7.6; 95% CI: 7.2, 8.0). Furthermore, these mortality rates among White Europeans with normal weight were about three-fold higher compared to African-Caribbean (rate: 3.2; 95% CI: 1.5, 8.4) and South Asians (rate: 4.6; 95% CI: 3.1, 7.0) with normal weight (Figure 8.3).

Table 8.2: Distribution of clinical characteristics among patients with T2DM by ethnicity in each BMI category.

	Normal weight (n=3,221)			**Overweight (n=12,924)**			**Obese (n=35,310)**		
	WE (n=2,423)	**AC (n=359)**	**SA (n=439)**	**WE (n=9,904)**	**AC (n=1,075)**	**SA (n=1,945)**	**WE (n=28,248)**	**AC (n=2,171)**	**SA (n=4,891)**
Age at diagnosis (yrs.) ‡	55 (12)	49 (11)	49 (13)	55 (10)	50 (11)	49 (11)	51 (12)	48 (11)	47 (12)
Male †	1,371 (57)	259 (72)	281 (64)	6,529 (66)	686 (64)	1,227 (63)	14,634 (51.8)	906 (41.7)	2,403 (49)
Smoking status†									
Never smokers	982 (41)	236 (66)	285 (65)	3,978 (40)	722 (67)	1,427 (73)	12,172 (43)	1,538 (71)	3,659 (75)
Current smokers	794 (33)	66 (18)	90 (21)	2,411 (24)	137 (13)	278 (14)	6,513 (23)	258 (12)	584 (12)
Ex-smokers	645 (27)	57 (16)	62 (14)	3,498 (35)	212 (20)	240 (12)	9,506 (34)	371 (17)	615 (13)
HbA$_{1c}$ at diagnosis ‡	9 (2)	10 (3)	9 (2)	9 (2)	9 (2)	9 (2)	9 (1)	9 (2)	9 (1)
SBP at diagnosis ‡	136 (19)	134(20)	129(18)	139 (17)	137 (17)	132 (17)	140 (16)	138 (16)	135 (15)
Cardiovascular diseases during follow-up †									
Hypertension	822 (34)	125 (35)	109 (25)	4,091 (41)	452 (42)	610 (31)	11,813 (42)	867 (40)	1,593 (33)
Angina	51 (2)	2 (0.6)	7 (2)	284 (3)	10 (0.9)	49 (3)	714 (3)	17 (0.8)	92 (2)
MI	60 (3)	3 (0.8)	14 (3)	267 (3)	6 (0.6)	44 (2)	536 (2)	17 (0.8)	94 (2)
CHD	111 (5)	6 (2)	25 (6)	563 (6	18 (2)	118 (6)	1,264 (5)	33 (2)	201 (4)
HF	36 (2)	1 (0.3)	1 (0.2)	153 (2)	11 (1)	14 (0.7)	464 (2)	22 (1)	39 (0.8)
Stroke	83 (3)	10 (3)	10 (2)	393 (4)	21 (2)	43 (2)	877 (3)	46 (2)	96 (2)
Any CVD	331 (14)	22 (6)	48 (11)	1,416 (14)	62 (6)	183 (9)	3,183 (11)	107 (5)	395 (8)
Lifestyle advice †	1,676 (69)	226 (63)	258 (59)	6,906 (20)	640 (60)	1,237 (64)	19,273 (68)	1,264 (58)	2,857 (58)
LBW†	699 (29)	105 (29)	125 (29)	2,098 (21)	214 (20)	459 (24)	4,798 (17)	343 (16)	641 (13)

	Normal weight (n=3,221)			Overweight (n=12,924)			Obese (n=35,310)		
Antidiabetic drugs[†]									
OAD	2,033 (84)	325 (91)	402 (92)	8,055 (81)	956 (89)	1,722 (89)	23,306 (83)	1,808 (83)	4,031 (82)
Metformin	1,867 (77)	295 (82)	370 (84)	7,751 (78)	908 (85)	1,658 (85)	22,750 (81)	1,758 (81)	3,897 (80)
Sulphonylureas	1,246 (51)	213 (59)	268 (61)	4,217 (43)	531 (49)	896 (46)	10,357 (37)	855 (39)	1,888 (39)
TZD	310 (13)	41 (11)	71 (16)	1,416 (14)	112 (10)	261 (13)	3,965 (14)	230 (11)	584 (12)
DPP4-i	292 (12)	48 (13)	51 (12)	1,347 (14)	123 (11)	245 (13)	4,228 (15)	260 (12)	645 (13)
GLP1-RA	9 (0.4)	1 (0.3)	1 (0.2)	179 (2)	12 (1)	11 (0.6)	2,007 (7)	71 (3)	130 (3)
SGLT2-i	11 (0.5)	1 (0.3)	0 (0)	55 (0.6)	5 (0.5)	13 (0.7)	352 (1)	10 (0.5)	46 (0.9)
Alpha-glucosidase	11 (0.5)	3 (0.8)	3 (0.7)	46 (0.5)	6 (0.6)	10 (0.5)	125 (0.4)	8 (0.4)	19 (0.4)
Meglitinide	29 (1)	5 (1)	5 (1)	66 (0.7)	7 (0.7)	14 (0.7)	190 (0.7)	20 (0.9)	40 (0.8)
Insulin	433 (18)	48 (13)	44 (10)	1,120 (11)	115 (11)	162 (8)	3,693 (13)	253 (12)	464 (10)
Other medications [†]									
CPM	1,967 (81)	281 (78)	333 (76)	8,609 (87)	852 (79)	1,601 (82)	23,669 (84)	1,631 (75)	3,629 (74)
Diuretics	579 (24)	64 (18)	55 (13)	2,901 (29)	273 (25)	317 (16)	9,419 (33)	582 (27)	965 (20)
Beta-blockers	482 (20)	46 (13)	55 (13)	2,373 (24)	164 (15)	326 (17)	6,759 (24)	332 (15)	839 (17)
Calcium blockers	592 (24)	111 (31)	92 (21)	3,197 (32)	412 (38)	497 (26)	8,883 (31)	912 (42)	1,272 (26)
Renin-angiotensin	1,152 (48)	169 (47)	192 (44)	5,803 (59)	553 (51)	954 (49)	16,863 (60)	1,055 (49)	2,364 (48)
Ace inhibitors	286 (12)	56 (16)	67 (15)	1,593 (16)	167 (15)	330 (17)	4,823 (17)	366 (17)	909 (19)
Statins	1,072 (44)	147 (41)	172 (39)	5,257 (53)	485 (45)	834 (43)	14,778 (52)	898 (41)	1,979 (41)
Lipid-modifiers	1,796 (74)	226 (63)	312 (71)	7,837 (79)	693 (65)	1,454 (75)	20,714 (73)	1,289 (59)	3,129 (64)
Anti-depressants	1,807 (75)	227 (63)	313 (71)	7,918 (80)	695 (65)	1,455 (75)	20,914 (74)	1,295 (60)	3,154 (65)

[†]: n (%);

[‡]: mean (SD);

Data are presented for patients without a history of disease at diagnosis.

MI: Myocardial Infarction; **CHD:** Coronary heart disease (including bypass surgery and angioplasty**)**; **HF:** Heart failure; **Any CVD**: Cardiovascular disease defined as the occurrence angina, myocardial infarction, coronary heart disease (including bypass surgery and angioplasty), heart failure, and stroke on or before diagnosis of T2DM; **OAD**: use of oral antidiabetic drug; **CPM**: use of cardio-protective medications; **LBW**: lost at least 2kg body weight before diagnosis; **TZD**: Thiazolidinedione; **DPP4-i**: Dipeptidyl peptidase 4 inhibitors; **GLP1-RA**: Glucagon-like peptide-1 receptor agonists; **SGLT2-i**: sodium-glucose transport protein 2 inhibitors; **WE**: White European; **AC**: African-Caribbean; **SA**: South Asian.

Table 8.3: Adjusted average time to first CVD event or all-cause mortality (95% CI) in obese patients, and the difference in time to such events in patients with normal body weight or overweight compared to their obese counterpart.

	White European (n=40,575)		**African-Caribbean (n=3,605)**		**South Asian (n=7,275)**	
Any CVD						
Mean time (years) – Obese	4.6 (4.5, 4.7)		4.5 (3.9, 5.2)		4.9 (4.6, 5.3)	
Difference (years)		p-value		p-value		p-value
Normal weight vs Obese	-0.5 (-0.9, -0.1)	0.018	-1.1 (-2.5, 0.3)	0.117	-0.3 (-1.4, 0.9)	0.659
Overweight vs Obese	0.1 (-0.2,0.3)	0.521	1.2 (0.1, 2.2)	0.040	-0.6 (-1.2, 0.0)	0.054
All-cause mortality						
Mean time (years) – Obese	7.0 (6.8,7.2)		6.6 (5.6, 7.7)		7.3 (6.6, 7.9)	
Difference (years)						
Normal weight vs Obese	-0.6 (-1.2, -0.03)	0.037	-1.0 (-2.5, 0.6)	0.207	-2.5 (-4.6, -0.3)	0.023
Overweight vs Obese	-0.3 (-0.6, 0.0)	0.048	-0.02 (-1.7, 1.6)	0.978	0.1 (-1.2, 1.3)	0.899

Patients were without a history of disease at diagnosis.

Any CVD: Cardiovascular disease defined as the occurrence angina, myocardial infarction, coronary heart disease (including bypass surgery and angioplasty), heart failure, and stroke post diagnosis of T2DM.

Table 8.4: Comparison of the distribution of age at diagnosis (in years), sex (%), and follow-up from diagnosis (in years) between patients who meet inclusion criteria for this study (n=51,455), patients excluded, and all patients[†] (n=191,953).

	Study cohort (n=51,455)	Excluded patients (n=140,498)	All patients [†] (n=191,953)
Age at diagnosis (years)			
Mean (SD)	54(11)	51(12)	54(11)
Median (Q1, Q3)	57 (47-64)	52 (43-61)	56(46-63)
Male, %	81,407 (58)	28,296 (55)	109703 (57)
Follow-up from diagnosis (years)			
Mean (SD)	8(4)	7(4)	8(4)
Median (Q1, Q3)	8 (4-11)	7(4-11)	8(4-11)

[†] Patients diagnosed with T2DM who were aged between 18 and 70 with a minimum 1 year of follow-up, and BMI ≥ 18.5 kg/m^2

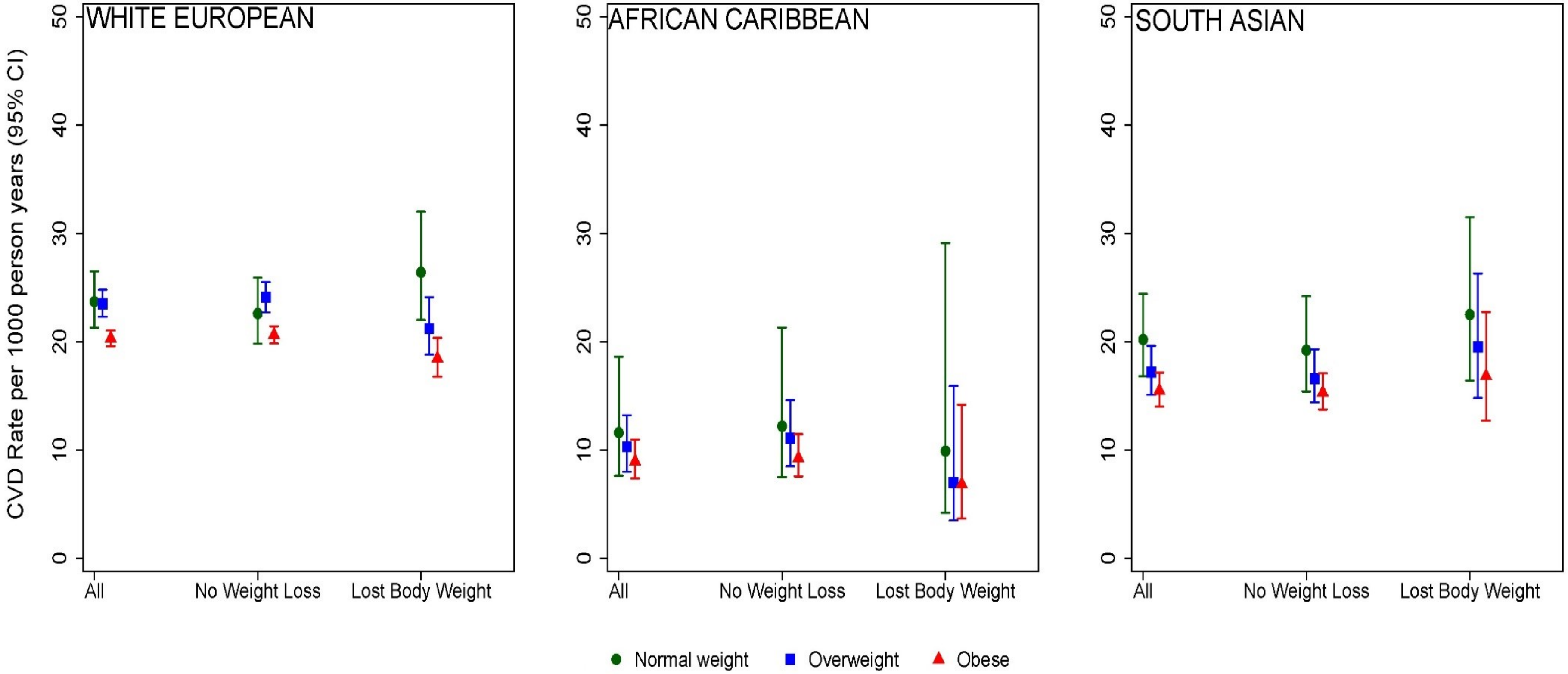


Figure 8.2: Age-weighted CVD event rates per 1000 person-years (95% CI) by BMI categories and weight change pattern before diagnosis in patients without disease history at diagnosis separately for three ethnic groups.

(Legend: **CVD**: Cardiovascular disease defined as the occurrence angina, myocardial infarction, coronary heart disease (including bypass surgery and angioplasty), heart failure, and stroke post diagnosis of T2DM).

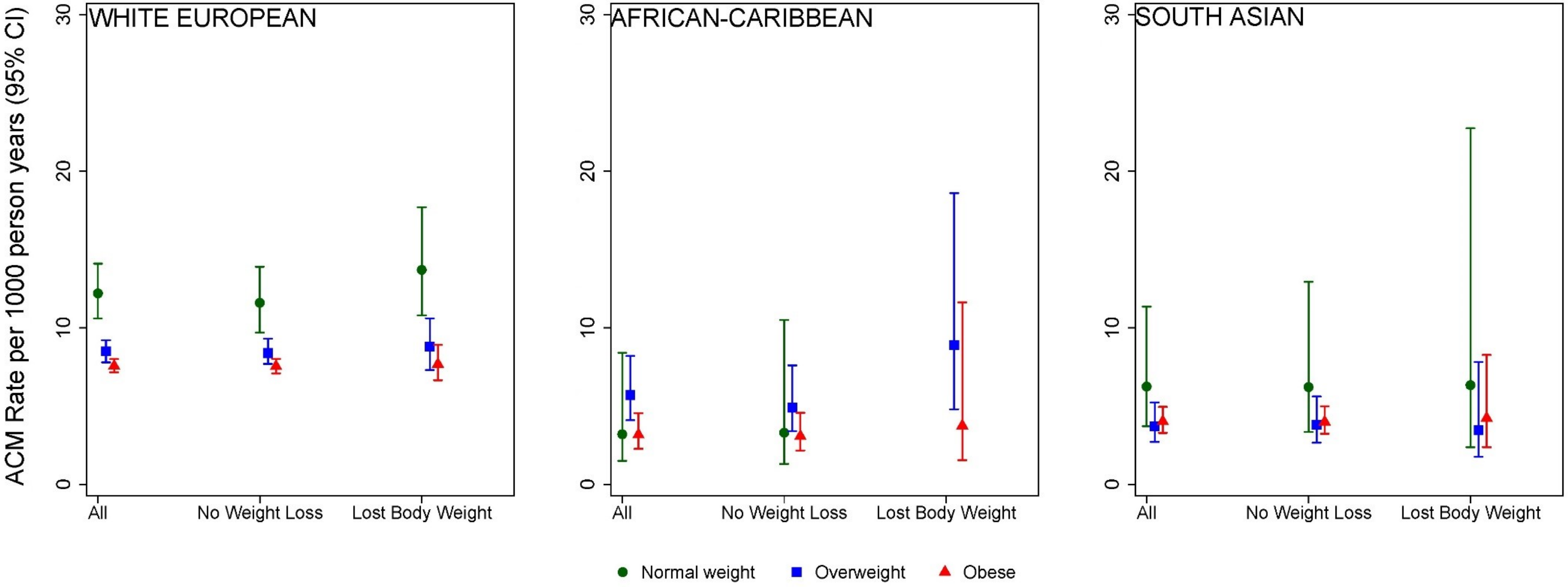


Figure 8.3: Age-weighted all-cause mortality rates per 1000 person-years (95% CI) by BMI categories and weight change pattern before diagnosis in patients without disease history at diagnosis separately for three ethnic groups.

(Legend: **ACM**: All-cause mortality; rates were not calculated for events less than or equal to 5).

8.4.3 Association of BMI categories with survival time for CVD and mortality

The adjusted average time to first CVD event (95% CI) and adjusted average time to ACM (95% CI) in normal weight and overweight patients compared to obese patients with T2DM and without a history of disease at diagnosis, within each ethnic group are presented in Table 8.3. Among White Europeans, compared to obese patients (mean time to CVD of 4.6 years), normal weight patients developed CVD significantly earlier by 0.5 years (95% CI: 0.1, 0.9 years; p=0.018). Furthermore, there was no significant difference between overweight White Europeans and obese White Europeans with regards to time to first CVD event (p > 0.05). The risk of developing CVD was not significantly higher in normal weight African-Caribbeans and South Asians, compared to their obese counterpart. However, overweight African-Caribbeans developed CVDs about 1.2 years (95% CI: 0.6, 2.2 years; p=0.040) later compared to obese African-Carribeans.

With a mean time to death of 7.0 and 7.3 years among obese White Europeans and South Asians respectively, those with normal body weight at diagnosis were significantly more likely to die earlier by 0.6 years (95% CI: 0.03, 1.2 years; p=0.037) in the White European group and by 2.5 years (95% CI: 0.3, 4.6 years; p=0.023) in the South Asian group.

8.5 DISCUSSION

The novelty of this electronic medical record-based study from a nationally representative primary care database in incident T2DM patients include assessment of risk profile for different ethnic groups at the time of clinical diagnosis of diabetes by different adiposity level, extensive exploration of weight change patterns prior to diagnosis of diabetes, a robust evaluation of the rates and risk of cardiovascular disease and all-cause mortality in different ethnic groups with different adiposity levels.

In this longitudinal outcome study based on three well-defined ethnic groups, we found that (1) the paradoxical association of lower BMI with high CVD rate appeared only in White Europeans, and this was not modified by weight change pattern before diagnosis, (2) normal weight White Europeans and South Asians appear to have significantly higher mortality risk compared to their obese counterparts, independent of weight change patterns prior to diagnosis of T2DM, and (3) the BMI at diagnosis was not associated with increased risk of CVD and death among African-Caribbeans.

Obesity is a strong risk factor for cardiovascular diseases in the general population and in some clinical populations. However, increasing evidence is pointing to a paradoxical phenomenon where overweight or obese patients may have better survival outcomes regarding developing heart failure or coronary heart disease, compared to normal weight patients [237,241]. Our analysis in patients with T2DM goes further to show that this paradoxical association between lower BMI and higher CVD risk was strong among White Europeans. In keeping with previous studies, our data show higher proportions of current smokers among normal weight White Europeans [242,243] and this could contribute to increased CVD risk among this group of patients. We previously reported that contrary to the notion that the observed obesity paradox could be due to weight loss from latent diseases, weight loss before the diagnosis of T2DM was not associated with increased mortality in normal weight patients [234]. The current study shows that the significantly higher event rates for CVD and mortality among normal weight patients were independent of weight change pattern before the diagnosis of T2DM. This clearly supports the fact that weight change pattern before diagnosis does not impact the observed obesity paradox in patients with T2DM.

Patients with many types of CVD may have a better prognosis if classified as overweight or obese. However, a previously published study with the same database has reported that among patients with a history of cardiovascular diseases at diagnosis of diabetes, those with normal weight at diagnosis had 30% (CI of HR: 1.11, 1.53) significantly higher adjusted mortality risk compared to the obese counterpart [12]. While conducting exploratory analyses during my PhD project, I observed that: among those with history of CVD, CKD or cancer at diagnosis, normal weight patients who did not / did lose body weight prior to diagnosis had 18% (CI of HR: 1.02, 1.37) / 48% (95% CI of HR: 1.24, 1.77) higher adjusted mortality risk (this data was not presented in the thesis). This clearly reflects the overall higher mortality risk paradigm in patients with established CVD / secondary care population. It is important to mention here that the people who are sicker tend to lose weight, which can artificially make obesity look protective. The relationship becomes confounded in this case and is difficult to draw a robust inference in this scenario.

In evaluating the association of BMI levels with mortality risk, we adjusted for weight loss pattern before the diagnosis of T2DM, in addition to other known confounders in the risk estimation models and found the paradoxical association of lower BMI with higher mortality risk was more prominent in South Asians than White Europeans. Some studies in patients with T2DM have compared all-cause mortality in different ethnic groups and different BMI categories, but to the best of our knowledge, none have provided ethnicity-specific mortality risk estimates by BMI categories at diagnosis. While the study by Kokkinos and colleagues [79] reported significantly higher mortality risk among African-Caribbeans and Caucasians with normal weight compared to patients with BMI ≥ 35 kg/m^2 (reference group), the BMI measure used in this study was not obtained at diagnosis of diabetes. Our risk assessments were based on BMI measured at diagnosis of T2DM and a more pragmatic approach of estimating the time to the events under consideration compared at different adiposity levels with an average 7 years of follow-up time, rather than estimating the hazard ratios which might provide misleading inference under highly heterogeneous characteristics in different ethnic groups. We also ensured the exclusion of patients with already existing diseases at diagnosis that are associated with increased mortality risk.

One of the novel findings of this study was that South Asians with normal body weight at diagnosis were significantly more likely to die earlier by about 2.5 years compared to their counterparts who were obese at diagnosis. One may argue that the above finding was due to the fact that the proportion of ever-smokers in normal weight South Asians (35%) was significantly higher than that in the obese group (25%), while the distribution of ever-smokers was similar between normal weight and obese White Europeans. However, while we do not know the change in smoking behaviour during the 7 years of median follow-up, our analyses were balanced for such possible baseline differences. Our result is contrary to that by Wright and colleagues [239] who reported longer life expectancy and reduced all-cause mortality for older South Asians with diabetes compared to White Europeans. Our current observation of lower prevalence of current or ex-smokers among South Asians compared to White Europeans was consistent with our previous study [41] and the study by Wright and colleagues [239]. Nonetheless, our observation agrees in principle with the study by Bellary and colleagues [209] who reported mean age at death for South Asians to be significantly lower by seven years compared to that in White Europeans. Findings from our study also suggest disparities in the receipt of lifestyle intervention advice, prescription of antidiabetic and cardio-protective drugs among patients with T2DM. These disparities are mostly skewed towards the majority White European population and could be the reason the obesity paradox was more prominent in South Asians than White Europeans.

Our findings should be interpreted considering the limitations of this study, which include: (1) availability of ethnicity data on a limited number of patients; (2) non-availability of longitudinal data on smoking cessation, and (3) potential for residual confounding as with all observational studies. Despite the issue of limited ethnicity data on some patients, previous work with this cohort by our research group showed that the distribution of sex, smoking status and BMI among persons with missing information on ethnicity was similar to the respective distributions among those with available information on ethnicity [41]. Furthermore, we also attempted to minimize bias introduced by confounders by using the "treatment effect" modelling approach. With this approach, robust inferences are provided through appropriate adjustments and balancing of a detailed list of confounders. However, patient-level data from electronic health records still present challenges regarding accuracy and completeness.

In conclusion, our study confirms a paradoxical association BMI and mortality among patients with T2DM and provides new insight into the possible role of ethnicity in explaining the obesity paradox both regarding CVD and total mortality.

8.5.1 Acknowledgements

S.K.P. and E.S.O.A conceived the idea and contributed to the study design. E.S.O.A. conducted the data extraction, data manipulation, statistical analyses and developed the first draft of the manuscript. E.S.O.A., K.K.R, and S.K.P. contributed to the finalization of the manuscript. S.K.P. contributed to the statistical analyses and had full access to all the data in the study and is the guarantor, taking responsibility for the integrity of the data and the accuracy of the data analysis. E.S.O.A. was supported by QIMR Berghofer International Ph.D. Scholarship and The University of Queensland International Scholarship. No separate funding was obtained for this study. Melbourne EpiCentre gratefully acknowledges the support from the Australian Government's National Collaborative Research Infrastructure Strategy (NCRIS) initiative through Therapeutic Innovation Australia and the research project funding from the National Health and Medical Research Council of Australia (Project Number: GNT1063477).

8.5.2 Conflicts of interest

S. K. P. has acted as a consultant and/or speaker for Novartis, GI Dynamics, Roche, AstraZeneca, Guangzhou Zhongyi Pharmaceutical and Amylin Pharmaceuticals LLC. He has received grants in support of investigator and investigator-initiated clinical studies from Merck, Novo Nordisk, AstraZeneca, Hospira, Amylin Pharmaceuticals, Sanofi Aventis and Pfizer. K.K.R. has acted as a speaker or consultant for Abbvie, Amgen, Pfizer, Astra Zeneca, Sanofi Resverlogix, Regeneron, Esperion, ACKCEA, Medicines Company, BI, Novo Nordisk. E. S. O. A. has no conflicts of interest to declare.

Chapter 9: General Discussion and Conclusion

In evaluating the obesity paradox in patients with T2DM, a robust methodological framework that incorporates several biostatistical and epidemiological methods was used to address the aims of this thesis. The obesity paradox is a phenomenon where patients with T2DM who were normal weight at diagnosis were found to have significantly higher mortality risk compared to their obese counterparts. As pointed out in the literature review section (Section 1.3.4), there are some methodological limitations of previous studies that evaluated the association of adiposity levels with mortality risk in patients with diabetes, which can be grouped into design and analytical issues. The current thesis has considered these limitations in addressing its main aims and the summary of primary results are given below.

9.1.1 Difficulty in identifying patients from the THIN database.

This thesis used a primary care-based EMR database from the UK called THIN (n=11,018,025 patients). THIN, like other primary care databases, present clinical researchers with the opportunity to conduct epidemiologic studies on a host of disease conditions of interest. Disease events are recorded using Read codes in THIN. While most of the diagnosis codes are well recorded, data entry errors like omissions, typing, or communicating errors may result in undiagnosed, misdiagnosed and misclassification of disease status.

Chapter 3 confirms the need for an extensive data mining/machine learning approach to correctly identifying patients with T2DM in a holistic way from the THIN database. Deterministic approaches based on disease Read codes were compared to a logistic regression classification algorithm to identify patients with T2DM. Of the patients identified by the classification algorithm to be living with T2DM, 17% did not have a T2DM Read code and 16% of those identified by the deterministic approach were not identified by the classification algorithm. Also, complement cohort-specific analyses based on markers of elevated glucose (at least two measurements of $HbA_{1c} > 6\%$ or fasting blood glucose > 7 mmol/l or random blood glucose > 11.1 mmol/l within 1 year) clearly showed that the ML classification algorithm reduces the number of patients with uncoded or undiagnosed T2DM. This formed a crucial part of the current thesis because, by design, the time of clinical diagnosis of diabetes (or close time-window around it) was set as the baseline. By correctly identifying patients with T2DM, the corresponding date of diagnosis helped in obtaining BMI and other cardiovascular and glycaemic risk factors within 3 months of diagnosis. These measurements obtained within this time window were considered as the baseline measures.

9.1.2 Multiple imputation of missing longitudinal risk factor

One of the critical problems with EMR data from a primary care setting, as with all longitudinal observational data, is the issue of missing data [106,244-246]. It was a condition of inclusion that a patient must have at least two measures of a risk factor longitudinally. This thesis seeks to explore the longitudinal trajectory of risk factors, hence the reason to impose this condition before the multiple imputation of missing longitudinal. Among patients with a minimum 2 years of follow-up, the proportion of patients who had at least 2 of the 5 measures missing was 16 % for weight, BMI, and SBP and 11% for HbA_{1c}. Although the problem of having significant proportions of missing data in longitudinal studies can be minimised through careful design, it is almost unavoidable in clinical and epidemiological studies [247-249].

Several factors may affect the frequency of recording longitudinal risk factor data within the primary care as data entry in EMRs depends on the nature and level of engagement between the individual and the clinical service provider. First, younger patients are less likely to get blood tests done because of perceived low-risk profile. Second, full panel blood tests may be requested and performed more frequently given the severity of disease (e.g. patients with T2DM on dynamic anti-diabetic drug treatment regimens). Third, missing communications from pathology laboratories could lead to missing values on some risk factors for some patients. Fourth, the ease with which a test is performed also affects the frequency of recording results from the test within the EMR. For example, systolic and diastolic blood pressure measurements may be recorded at every GP encounter because of the relative ease with which it can be measured. Also, the capture of the common adiposity measurements may vary as body weight is measured more often than waist circumference due to the simplicity and standard way it is measured. Finally, the missing data may also arise simply because a patient failed to attend the scheduled consultation.

These aspects complicate the process of evaluating the nature of missing data in EMRs making it difficult to appropriately differentiate between random and non-random missingness patterns. Before investigating and imputing for the missing data, understanding the mechanisms behind the missing data is crucial. In practice, incomplete data are typically considered as MAR even if they may not be [248,250]. In most EMRs, some variables would be expected to partially explain some of the variation in missingness, which indicates imputation under MAR setting [248].

As outlined in Chapter 4, I tried to account for variation in missingness by adjusting for age at diagnosis, sex, smoking status, deprivation status, the usages of ADDs during the multiple imputation process. Subsequently, the longitudinal distribution of body weight, BMI, SBP was similar for both complete and imputed datasets indicating that the multiple imputation via PMM captured the true longitudinal distribution of these risk factors. This result guarantees reliable inferences from any analysis that uses the imputed risk factor data.

9.1.3 BMI, ethnicity, and the risk of developing T2DM

Some studies have shown a link between obesity and an increased risk of developing T2DM, but little was known about the differences in risk of T2DM across BMI levels among a multi-ethnic group of patients. One of the novel components of this thesis was the extensive evaluation of the differences in risk of T2DM over the entire spectrum of BMI among a multi-ethnic group of patients in Chapter 5. A comparison of the distribution of BMI at diagnosis between patients with T2DM and their age-sex-ethnicity matched non-diabetic controls revealed significant differences in cardiovascular and glycaemic risk profiles at different BMI levels at diagnosis of T2DM among different ethnic groups. Most notably, South Asians developed T2DM significantly early (~2-10 years) and at a lower BMI compared to African Caribbeans and White Europeans respectively. These results are consistent with previous studies as well and are important for the evaluation of the obesity paradox in patients with T2DM, as ethnicity has been suggested as a possible reason for the observed increased mortality risk in normal weight patients with T2DM compared to obese patients with T2DM.

9.1.4 Pre-existing disease conditions

One proposed reason for the obesity paradox in patients with T2DM is that some pre-existing disease conditions are over-represented in the normal weight group and lead to weight loss before the diagnosis of a T2DM, hence the increased mortality rate in the normal weight group. Several studies have compared the prevalence and severity of diabetes complications between South Asians and White Europeans [202-207], but no separate assessment of the potential differences in the risk paradigm by adiposity levels was evaluated. The second novel aspect of this thesis is a dedicated evaluation of pre-existing cardiovascular and non-cardiovascular diseases before and after the diagnosis of T2DM in different ethnic groups for each BMI category in Chapter 6. The prevalence of pre-existing cardiovascular diseases among normal weight patients with T2DM ranged from 4.0% to 11.3% across White Europeans, African-Caribbeans and South Asians. Furthermore, obese White Europeans have significantly higher prevalence compared to their normal weight population and also compared to other ethnic groups. These findings add to current literature and provide a greater understanding of the relationship between levels of adiposity and diabetes complications in different ethnic groups. The results of this study should enable clinicians to better diagnose and manage diabetes amongst people of different ethnicities. Given this interplay between ethnicity, BMI and cardiovascular diseases at diagnosis, patients with T2DM who had established pre-existing disease conditions were separated from those without these conditions at diagnosis in order to disentangle the contribution of pre-existing disease to weight loss before the diagnosis of T2DM. These analyses are presented in Chapter 7 and 8.

9.1.5 Weight loss before diagnosis and the obesity paradox in patients with T2DM

Weight loss before diagnosis as a result of pre-existing disease conditions could have an impact on the association between BMI and mortality. Chapter 7 of this thesis addresses the obesity paradox in a two-step approach. First, the influence of the presence or absence of pre-existing disease conditions (latent/underlying/undiagnosed) on changes in body weight before the diagnosis of T2DM were investigated. Second, an assessment of the possible impact of weight change pattern before diagnosis on the association between BMI at diagnosis and long-term mortality risk was conducted.

An analysis of weight trajectory before diagnosis was conducted in patients without pre-existing disease conditions at diagnosis. This was the third novel aspects of the current thesis as it revealed that patients with T2DM who were normal weight and overweight at diagnosis experienced a small but significant reduction in body weight six months before diagnosis. However, unlike overweight patients who continued in the downward trend at 6 months after diagnosis before increasing at 12 months after diagnosis, normal weight patients had a steady increase in body weight throughout the respective time post-diagnosis. If the observed weight loss in normal weight patients before diagnosis was due to pre-existing disease conditions, then their trajectory after diagnosis would have continued downward. This observation coupled with the fact that pre-existing disease conditions were not over-represented in the normal weight group contradicts the assertion of possible weight loss due to pre-existing disease.

Furthermore, among patients with no established disease conditions at diagnosis, the association of BMI at diagnosis with mortality risk, separately for patients who lost body weight before diagnosis and those who did not was evaluated. This was the fourth novel analysis done to estimate the possible influence of weight change pattern before diagnosis on the association between BMI at diagnosis and long-term mortality risk. This analysis demonstrates that among those who did not lose body weight before diagnosis, patients with normal weight at diagnosis had significantly higher mortality risk compared with grade 1 obese patients. However, among those who lost body weight before diagnosis, BMI at diagnosis was not associated with mortality risk. These results provided enough evidence that weight loss due to pre-existing diseases could not explain the obesity paradox in patients with T2DM.

9.1.6 Survival-time treatment effects model

Cox proportional hazard regression is a widely used approach to analyse survival time data because of its flexible semi-parametric property. This means that in using this regression method, an assumption of the underlying distribution of the outcome is not required. However, to develop a regression model, a metric influence of covariates is required. Therefore, an assumption that covariates modify a shared underlying hazard function is made (proportionality assumption). If a variable(s) violates the proportionality assumption, the options for robust results are to (1) include the variable(s) as stratification factor, (2) include the variable as time-varying variables (i.e. adding a term for the interaction of covariates with time), and (3) separate scale of analysis time into equal bands and perform the proportional hazards regression within each band.

The key assumption of the proportional hazards regression model is unlikely to be true for patients with incident T2DM under different adiposity levels. As part of the preliminary model diagnostic test for the Cox proportional hazards model used in Chapter 7 of this thesis, the effect of some variables were not constant over time (i.e. violated the assumption). Therefore, the final model used to obtain the estimates of mortality risk reported in Chapter 7 was a stratified Cox regression model with age group at diagnosis as stratification factor and included terms for the interaction of covariates with time.

To account for the inherent differences in risk factors between the defined BMI categories and the fact that risk may not be proportional, a method that allows for the balancing of categories is required. By means of weighted propensity-score type adjustments, the survival time treatments effects modelling can provide robust inferences [113-116] by adjusting and balancing comparison categories based on global risk paradigm within the cohort. The results from Chapter 8 have confirmed the use of this novel modelling approach to account for the limitations of the traditional proportional Cox regression model. To the best of our knowledge, no study investigating the obesity paradox in T2DM has adopted this robust approach.

9.1.7 Ethnicity and the obesity paradox in patients with T2DM

For clinical management of diabetes among patients of different ethnicity to improve, a better understanding of the relationship between levels of adiposity and diabetes complications in different ethnic groups is required. Chapter 6 showed that the overall risk of developing MACE was significantly higher for patients with T2DM compared to non-diabetic controls at all levels of BMI within each ethnic group. While the risk was similar for White Europeans and African Caribbeans, it was significantly higher for South Asians compared to White Europeans.

Furthermore, using data on patients with T2DM patients only, the potential role of ethnicity in the association of BMI with mortality was evaluated in Chapter 8. This work represents a major update on previous studies evaluating the observed phenomenon of the obesity paradox in T2DM. To the best of our knowledge, only one study has examined the modifiable effect of ethnicity on long-term mortality risks at different adiposity levels, but only male participants were included in this study [79]. Survival time treatment effects modelling was used to examine cardiovascular and mortality risk for each ethnic group. White Europeans who were normal weight at diagnosis developed CVDs significantly earlier compared to their obese colleagues. However, BMI at diagnosis was not associated with increased risk of CVD among African-Caribbeans and South Asians. This clearly suggests that the paradoxical association of lower BMI with high CVD rate appeared only among White Europeans.

9.2 FUTURE DIRECTIONS

Exposure to or use of anti-diabetic therapy may lead to weight loss or gain after diagnosis [117,118], which may have different effects on the association of BMI at diagnosis with mortality and cardiovascular risk. In this thesis, the potential for confounding by medication use was reduced by adjusting for weight, use of insulin, as well as the use of oral ADDs during follow-up. However, within the context of evaluating the obesity paradox in patients with T2DM, it is possible that mortality risk may be reduced for individuals in the overweight/obese category because of more aggressive therapy for patients in this group. There is the need to explore the possible association of weight changes with cardiovascular or mortality outcomes in patients treated with different antidiabetic medications. With complete information on classes of ADD (including start and stop dates), exposure to different combination of ADDs can be defined and future studies might assess (1) the influence of specific drug class on the cardiovascular and mortality “risk spectrum” observed between different BMI categories, and (2) if time spent on a particular drug class influences the “risk spectrum”, with its residual effect over-influencing the body weight factor.

Another research question that may arise evolves around if there are any long-term interactions of cardiovascular (blood pressure, lipids) and glycaemic risk factors (e.g. glucose levels measured by HbA_{1c} and hypoglycaemia) with body weight that modify the risks? Therefore, it is very important to evaluate the possible interactions of time-varying blood pressure, lipid measures and glucose control (measured by HbA_{1c}) with the changes in body weight, while evaluating the associated risk. Further studies should also examine the level of interactions between body weight and HbA_{1c} in relation to mortality risk. Identification of the patterns of possible interactions over time and their effects on cardiovascular and mortality risks, by baseline BMI status, will provide new clinical information to better manage the patients with diabetes. Further studies can be conducted by using measures of abdominal girth instead of BMI. This however depends on improved reporting of measures of abdominal girth within the UK primary care settings, in such a way that there will be enough longitudinal data for patients with T2DM. Opportunity also exists for further research on elucidating the genetic basis of the obesity paradox in patients with T2DM.

9.3 CONCLUSION

Overall, the findings of this thesis add to the evidence base that patients with T2DM, who were normal weight at the time of clinical diagnosis have significantly higher mortality risk compared to those who were obese, and this may partially be driven by different cardiovascular and glycaemic risk profiles of different ethnic groups. Empirical results from this thesis suggest that there was no evidence of pre-existing latent or severe disease conditions being overrepresented in normal weight patients. Infect, dynamic changes in body weight before clinical diagnosis of T2DM were independent of pre-existing latent or severe disease conditions. After untangling the roles of pre-existing severe disease conditions in dynamic changes in body weight before clinical diagnosis of T2DM, the increased mortality risk in the normal weight group may reflect differences in the aetiology of diabetes in normal-weight people and emphasises the importance of addressing risk factors for excess mortality in this group.

Bibliography

1. Nathan DM. Diabetes: Advances in diagnosis and treatment. *JAMA* 2015;**314**(10):1052-1062.
2. Ghosh Sujoy, Andrew Collier. *Chruchill's Pocketbook of Diabetes*. 2 ed Elsevier Ltd, 2012.
3. International Diabetes Federation. IDF Diabetes Atlas. http://www.diabetesatlas.org Accessed 22-Nov, 2017.
4. American Diabetes Association. Diagnosis and classification of diabetes mellitus. *Diabetes Care* 2014;**37 Suppl 1**:S81-90.
5. World Health Organization. Obesity: preventing and managing the global epidemic: report of a WHO consultation on obesity,. Geneva, 1998.
6. Eckel RH, Kahn SE, Ferrannini E, Goldfine AB, Nathan DM, Schwartz MW, Smith RJ, Smith SR. Obesity and Type 2 Diabetes: What Can Be Unified and What Needs to Be Individualized? *Diabetes Care* 2011;**34**(6):1424-1430.
7. Zoppini G, Verlato G, Leuzinger C, Zamboni C, Brun E, Bonora E, Muggeo M. Body mass index and the risk of mortality in type II diabetic patients from Verona. *International Journal of Obesity* 2003;**27**(2):281-285.
8. Flegal KM, Kit BK, Orpana H, Graubard BI. Association of all-cause mortality with overweight and obesity using standard body mass index categories: a systematic review and meta-analysis. *JAMA* 2013;**309**(1):71-82.
9. Kalantar-Zadeh K, Streja E, Kovesdy CP, Oreopoulos A, Noori N, Jing J, Nissenson AR, Krishnan M, Kopple JD, Mehrotra R, Anker SD. The Obesity Paradox and Mortality Associated With Surrogates of Body Size and Muscle Mass in Patients Receiving Hemodialysis. *Mayo Clinic Proceedings* 2010;**85**(11):991-1001.
10. Lavie CJ, Milani RV, Ventura HO. Obesity and cardiovascular disease: risk factor, paradox, and impact of weight loss. *Journal of the American College of Cardiology* 2009;**53**(21):1925-32.
11. Navaneethan SD, Kirwan JP, Arrigain S, Schold JD. Adiposity measures, lean body mass, physical activity and mortality: NHANES 1999–2004. *BMC Nephrology* 2014;**15**:108-108.
12. Thomas G, Khunti K, Curcin V, Molokhia M, Millett C, Majeed A, Paul S. Obesity paradox in people newly diagnosed with type 2 diabetes with and without prior cardiovascular disease. *Diabetes, Obesity & Metabolism* 2014;**16**(4):317-25.
13. Oreopoulos A, Padwal R, Kalantar-Zadeh K, Fonarow GC, Norris CM, McAlister FA. Body mass index and mortality in heart failure: a meta-analysis. *American Heart Journal* 2008;**156**(1):13-22.
14. Angerås O, Albertsson P, Karason K, Råmunddal T, Matejka G, James S, Lagerqvist B, Rosengren A, Omerovic E. Evidence for obesity paradox in patients with acute coronary syndromes: a report from the Swedish Coronary Angiography and Angioplasty Registry. *European Heart Journal* 2013;**34**(5):345-353.
15. Weber MA, Jamerson K, Bakris GL, Weir MR, Zappe D, Zhang Y, Dahlof B, Velazquez EJ, Pitt B. Effects of body size and hypertension treatments on cardiovascular event rates: subanalysis of the ACCOMPLISH randomised controlled trial. *The Lancet* 2013;**381**(9866):537-545.
16. Eknoyan G. Obesity, diabetes, and chronic kidney disease. *Current Diabetes Reports* 2007;**7**(6):449-53.
17. Eknoyan G. Obesity and chronic kidney disease. *Nefrologia* 2011;**31**(4):397-403.
18. Florez H, Castillo-Florez S. Beyond the obesity paradox in diabetes: fitness, fatness, and mortality. *JAMA* 2012;**308**(6):619-620.
19. Carnethon MR, De Chavez PJD, Biggs ML, Lewis CE, Pankow JS, Bertoni AG, Golden SH, Liu K, Mukamal KJ, Campbell-Jenkins B, Dyer AR. Association of weight Status with mortality in adults with incident diabetes. *JAMA* 2012;**308**(6):581-590.

20. Landman G, Van Hateren K, Kleefstra N, Bilo H. The relationship between obesity and cancer mortality in type 2 diabetes: a ten-year follow-up study (ZODIAC-21). *Anticancer Research* 2010;**30**(2):681-682.
21. Church TS, Cheng YJ, Earnest CP, Barlow CE, Gibbons LW, Priest EL, Blair SN. Exercise capacity and body composition as predictors of mortality among men with diabetes. *Diabetes Care* 2004;**27**(1):83-88.
22. Costanzo P, Cleland JG, Pellicori P, Clark AL, Hepburn D, Kilpatrick ES, Perrone-Filardi P, Zhang J, Atkin SL. The obesity paradox in type 2 diabetes mellitus: relationship of body mass index to prognosis: a cohort study. *Annals of Internal Medicine* 2015;**162**(9):610-8.
23. Ford ES, DeStefano F. Risk Factors for Mortality from All Causes and from Coronary Heart Disease among Persons with Diabetes: Findings from the National Health and Nutrition Examination Survey I Epidemiologic Follow-up Study. *American Journal of Epidemiology* 1991;**133**(12):1220-1230.
24. Tobias DK, Pan A, Jackson CL, O'Reilly EJ, Ding EL, Willett WC, Manson JE, Hu FB. Body-mass index and mortality among adults with incident type 2 diabetes. *New England Journal of Medicine* 2014;**370**(3):233-44.
25. Logue J, Walker JJ, Leese G, Lindsay R, Mcknight J, Morris A, Philip S, Wild S, Sattar N, on behalf of the Scottish Diabetes Research Network Epidemiology Group Association between BMI measured within a year after diagnosis of type 2 diabetes and mortality. *Diabetes Care* 2013;**36**(4):887–893.
26. Zhao W, Katzmarzyk PT, Horswell R, Wang Y, Li W, Johnson J, Heymsfield SB, Cefalu WT, Ryan DH, Hu G. Body Mass Index and the Risk of All-Cause Mortality Among Patients with Type 2 Diabetes. *Circulation* 2014;**130**(24):2143-51.
27. Abell JE, Egan BM, Wilson PW, Lipsitz S, Woolson RF, Lackland DT. Differences in cardiovascular disease mortality associated with body mass between Black and White persons. *American Journal of Public Health* 2008;**98**(1):63-66.
28. Tillin T, Sattar N, Godsland IF, Hughes AD, Chaturvedi N, Forouhi NG. Ethnicity-specific obesity cut-points in the development of Type 2 diabetes - a prospective study including three ethnic groups in the United Kingdom. *Diabetic Medicine* 2015;**32**(2):226-34.
29. Shai I, Jiang R, Manson JE, Stampfer MJ, Willett WC, Colditz GA, Hu FB. Ethnicity, Obesity, and Risk of Type 2 Diabetes in Women: A 20-year follow-up study. *Diabetes Care* 2006;**29**(7):1585-1590.
30. Ganz ML, Wintfeld N, Li Q, Alas V, Langer J, Hammer M. The association of body mass index with the risk of type 2 diabetes: a case-control study nested in an electronic health records system in the United States. *Diabetology & Metabolic Syndrome* 2014;**6**(1):50.
31. Rejeski WJ, Ip EH, Bertoni AG, Bray GA, Evans G, Gregg EW, Zhang Q. Lifestyle change and mobility in obese adults with type 2 diabetes. *New England Journal of Medicine* 2012;**366**(13):1209-1217.
32. Myers J, Lata K, Chowdhury S, McAuley P, Jain N, Froelicher V. The obesity paradox and weight loss. *The American Journal of Medicine* 2011;**124**(10):924-30.
33. The Look AHEAD Research Group. Long term effects of a lifestyle intervention on weight and cardiovascular risk factors in individuals with type 2 diabetes: four year results of the Look AHEAD trial. *Archives of Internal Medicine* 2010;**170**(17):1566.
34. Alberti KGMM, Zimmet PZ. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: Diagnosis and classification of diabetes mellitus. Provisional report of a WHO consultation. *Diabetic Medicine* 1998;**15**(7):539-553.
35. Ahima RS. *Metabolic basis of obesity*. New York: Springer, 2011.
36. Evans M, Vora J. *Managing Diabetes*. London: Springer Verlag, 2012.
37. Wild S, Roglic G, Green A, Sicree R, King H. Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. *Diabetes Care* 2004;**27**(5):1047-1053.
38. Zimmet PZ. Diabetes and its drivers: the largest epidemic in human history? *Clinical Diabetes and Endocrinology* 2017;**3**(1):1.

39. Zheng Y, Ley SH, Hu FB. Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. *Nature Reviews Endocrinology* 2017.
40. Hu FB. Globalization of Diabetes: The role of diet, lifestyle, and genes. *Diabetes Care* 2011;**34**(6):1249-1257.
41. Paul SK, Owusu Adjah ES, Samanta M, Patel K, Bellary S, Hanif W, Khunti K. Comparison of body mass index at diagnosis of diabetes in a multi-ethnic population: A case-control study with matched non-diabetic controls. *Diabetes, Obesity and Metabolism* 2017;**19**(7):1014-1023.
42. Xu Y, Wang L, He J, Bi Y, Li M, Wang T, Wang L, Jiang Y, Dai M, Lu J. Prevalence and control of diabetes in Chinese adults. *JAMA* 2013;**310**(9):948-959.
43. International Diabetes Federation. IDF Diabetes Atlas. http://www.diabetesatlas.org/resources/previous-editions.html Accessed 01/14, 2019.
44. Skyler JS. *Atlas of diabetes*. New York: Springer, 2012.
45. International Diabetes Federation. IDF Diabetes Atlas. http://www.idf.org/diabetesatlas Accessed 1/14, 2019.
46. Bellamy L, Casas J-P, Hingorani AD, Williams D. Type 2 diabetes mellitus after gestational diabetes: a systematic review and meta-analysis. *The Lancet* 2009;**373**(9677):1773-1779.
47. Rayanagoudar G, Hashi AA, Zamora J, Khan KS, Hitman GA, Thangaratinam S. Quantification of the type 2 diabetes risk in women with gestational diabetes: a systematic review and meta-analysis of 95,750 women. *Diabetologia* 2016;**59**(7):1403-1411.
48. Wendland EM, Torloni MR, Falavigna M, Trujillo J, Dode MA, Campos MA, Duncan BB, Schmidt MI. Gestational diabetes and pregnancy outcomes-a systematic review of the World Health Organization (WHO) and the International Association of Diabetes in Pregnancy Study Groups (IADPSG) diagnostic criteria. *BMC Pregnancy and Childbirth* 2012;**12**(1):23.
49. Clausen TD, Mathiesen ER, Hansen T, Pedersen O, Jensen DM, Lauenborg J, Damm P. High prevalence of type 2 diabetes and pre-diabetes in adult offspring of women with gestational diabetes mellitus or type 1 diabetes: the role of intrauterine hyperglycemia. *Diabetes Care* 2008;**31**(2):340-346.
50. Krentz A, Bailey C. Oral Antidiabetic Agents. *Drugs* 2005;**65**(3):385-411.
51. Nathan DM, Buse JB, Davidson MB, Ferrannini E, Holman RR, Sherwin R, Zinman B, American Diabetes A, Diabetes EAfSo. Medical management of hyperglycemia in type 2 diabetes: a consensus algorithm for the initiation and adjustment of therapy: a consensus statement of the American Diabetes Association and the European Association for the Study of Diabetes. *Diabetes Care* 2009;**32**(1):193-203.
52. Lopez AD, Mathers CD, Ezzati M, Jamison DT, Murray CJL. *Global Burden of Disease and Risk Factors*. Washington, DC;New York, NY;: The World Bank, 2006.
53. Nestle M, Jacobson MF. Halting the obesity epidemic: a public health policy approach. *Public Health Reports* 2000;**115**(1):12-24.
54. Lau DC, Douketis JD, Morrison KM, Hramiak IM, Sharma AM, Ur E, Obesity Canada Clinical Practice Guidelines Expert Panel. 2006 Canadian clinical practice guidelines on the management and prevention of obesity in adults and children [summary]. *CMAJ* 2007;**176**(8):S1-13.
55. Bray GA. Pathophysiology of obesity. *The American Journal of Clinical Nutrition* 1992;**55**(2):488S-494S.
56. Keith SW, Redden DT, Katzmarzyk PT, Boggiano MM, Hanlon EC, Benca RM, Ruden D, Pietrobelli A, Barger JL, Fontaine KR, Wang C, Aronne LJ, Wright SM, Baskin M, Dhurandhar NV, Lijoi MC, Grilo CM, DeLuca M, Westfall AO, Allison DB. Putative contributors to the secular increase in obesity: exploring the roads less traveled. *International Journal of Obesity* 2006;**30**(11):1585-94.
57. James WP. The fundamental drivers of the obesity epidemic. *Obesity Reviews* 2008;**9 Suppl 1**:6-13.

58. Bleich S, Cutler D, Murray C, Adams A. Why is the developed world obese? *Annual Review of Public Health* 2008;**29**:273-95.
59. Klonoff DC, Prahalad P. Performance of Cleared Blood Glucose Monitors. *Journal of Diabetes Science and Technology* 2015;**9**(4):895-910.
60. Rothney MP, Brychta RJ, Schaefer EV, Chen KY, Skarulis MC. Body Composition Measured by Dual-energy X-ray Absorptiometry Half-body Scans in Obese Adults. *Obesity (Silver Spring, Md.)* 2009;**17**(6):1281-1286.
61. Mei Z, Grummer-Strawn LM, Pietrobelli A, Goulding A, Goran MI, Dietz WH. Validity of body mass index compared with other body-composition screening indexes for the assessment of body fatness in children and adolescents. *The American Journal of Clinical Nutrition* 2002;**75**(6):978-985.
62. Cornier MA, Despres JP, Davis N, Grossniklaus DA, Klein S, Lamarche B, Lopez-Jimenez F, Rao G, St-Onge MP, Towfighi A, Poirier P, American Heart Association Obesity Committee of the Council on Nutrition, Physical and Activity Metabolism, Council on Arteriosclerosis, Thrombosis and Vascular Biology, Council on Cardiovascular Disease in the Young, Council on Cardiovascular Radiology and Intervention, Council on Cardiovascular Nursing, Council on Epidemiology and Prevention, Council on the Kidney in Cardiovascular Disease, Stroke Council. Assessing adiposity: a scientific statement from the American Heart Association. *Circulation* 2011;**124**(18):1996-2019.
63. World Health Organization. *Global health risks: mortality and burden of disease attributable to selected major risks*. Geneva, Switzerland: World Health Organization, 2009.
64. Wells JCK, Fewtrell MS. Measuring body composition. *Archives of Disease in Childhood* 2006;**91**(7):612-617.
65. Wellens RI, Roche AF, Khamis HJ, Jackson AS, Pollock ML, Siervogel RM. Relationships between the body mass index and body composition. *Obesity Research* 1996;**4**(1):35-44.
66. Garn SM, Leonard WR, Hawthorne VM. Three limitations of the body mass index. *The American Journal of Clinical Nutrition* 1986;**44**(6):996-7.
67. Ross R, Berentzen T, Bradshaw AJ, Janssen I, Kahn HS, Katzmarzyk PT, Kuk JL, Seidell JC, Snijder MB, Sørensen TIA, Després J-P. Does the relationship between waist circumference, morbidity and mortality depend on measurement protocol for waist circumference? *Obesity Reviews* 2008;**9**(4):312-325.
68. de Koning L, Merchant AT, Pogue J, Anand SS. Waist circumference and waist-to-hip ratio as predictors of cardiovascular events: meta-regression analysis of prospective studies. *European Heart Journal* 2007;**28**(7):850-856.
69. World Health Organization. Global status report on noncommunicable diseases 2014. *WHO*. Geneva: World Health Organization, 2015.
70. Harvard T.H. Chan School of Public Health. Obesity Prevention Source. https://www.hsph.harvard.edu/obesity-prevention-source/ Accessed 05/03/2016, 2016.
71. Popkin BM, Adair LS, Ng SW. Global nutrition transition and the pandemic of obesity in developing countries. *Nutrition Reviews* 2012;**70**(1):3-21.
72. Health and Social Care Information Centre. Health Survey for England 2015. https://webarchive.nationalarchives.gov.uk/20180328130330/http://digital.nhs.uk/catalogue/PUB22616 Accessed 10-March, 2019.
73. Badrick E, Sperrin M, Buchan IE, Renehan AG. Obesity paradox and mortality in adults with and without incident type 2 diabetes: a matched population-level cohort study. *BMJ Open Diabetes Research & Care* 2017;**5**(1).
74. Balkau B, Eschwège E, Papoz L, Richard JL, Claude JR, Warnet JM, Ducimetière P. Risk factors for early death in non-insulin dependent diabetes and men with known glucose tolerance status. *BMJ* 1993;**307**(6899):295-299.
75. Chaturvedi N, Fuller JH, Jarrett RJ, Keen H, Morrish NJ, Watkins PJ, Teuscher A, Teuscher T, Studer PP, Diem P, Czyzyk A, Janeczko D, Kopczynski J, Raskovic M, Schliack V, Ratzmann KP, Aganovic I, Skrabolo A, Stavljenic A. Mortality risk by body weight and

weight change in people with NIDDM: The WHO Multinational Study of Vascular Disease in Diabetes. *Diabetes Care* 1995;**18**(6):766-774.

76. Dallongeville J, Bhatt DL, Steg PH, Ravaud P, Wilson PW, Eagle KA, Goto S, Mas JL, Montalescot G. Relation between body mass index, waist circumference, and cardiovascular outcomes in 19,579 diabetic patients with established vascular disease: the REACH Registry. *European Journal of Preventive Cardiology* 2012;**19**(2):241-9.

77. Jackson CL, Yeh H-C, Szklo M, Hu FB, Wang N-Y, Dray-Spira R, Brancati FL. Body-Mass Index and All-Cause Mortality in US Adults With and Without Diabetes. *Journal of General Internal Medicine* 2014;**29**(1):25-33.

78. Khalangot M, Tronko M, Kravchenko V, Kulchinska J, Hu G. Body mass index and the risk of total and cardiovascular mortality among patients with type 2 diabetes: a large prospective study in Ukraine. *Heart* 2009;**95**(6):454-460.

79. Kokkinos P, Myers J, Faselis C, Doumas M, Kheirbek R, Nylen E. BMI–mortality paradox and fitness in African American and Caucasian men with type 2 diabetes. *Diabetes Care* 2012;**35**(5):1021-1027.

80. McEwen LN, Kim C, Karter AJ, Haan MN, Ghosh D, Lantz PM, Mangione CM, Thompson TJ, Herman WH. Risk Factors for Mortality Among Patients With Diabetes: The Translating Research Into Action for Diabetes (TRIAD) Study. *Diabetes Care* 2007;**30**(7):1736-1741.

81. Sasaki A, Horiuchi N, Hasegawa K, Uehara M. Mortality and causes of death in type 2 diabetic patients: A long-term follow-up study in Osaka District, Japan. *Diabetes Research and Clinical Practice* 1989;**7**(1):33-40.

82. Weiss A, Boaz M, Beloosesky Y, Kornowski R, Grossman E. Body mass index and risk of all-cause and cardiovascular mortality in hospitalized elderly patients with diabetes mellitus. *Diabetic Medicine* 2009;**26**(3):253-9.

83. Cho E, Manson JE, Stampfer MJ, Solomon CG, Colditz GA, Speizer FE, Willett WC, Hu FB. A prospective study of obesity and risk of coronary heart disease among diabetic women. *Diabetes Care* 2002;**25**(7):1142-8.

84. Edqvist J, Rawshani A, Adiels M, Björck L, Lind M, Svensson A-M, Gudbjörnsdottir S, Sattar N, Rosengren A. BMI and Mortality in Patients With New-Onset Type 2 Diabetes: A Comparison With Age-and Sex-Matched Control Subjects From the General Population. *Diabetes Care* 2018;**41**(3):dc171309.

85. Kuo J-F, Hsieh Y-T, Mao IC, Lin S-D, Tu S-T, Hsieh M-C. The Association Between Body Mass Index and All-Cause Mortality in Patients With Type 2 Diabetes Mellitus: A 5.5-Year Prospective Analysis. *Medicine* 2015;**94**(34):e1398.

86. McAuley PA, Myers JN, Abella JP, Tan SY, Froelicher VF. Exercise Capacity and Body Mass as Predictors of Mortality Among Male Veterans With Type 2 Diabetes. *Diabetes Care* 2007;**30**(6):1539-1543.

87. Mulnier HE, Seaman HE, Raleigh VS, Soedamah-Muthu SS, Colhoun HM, Lawrenson RA. Mortality in people with type 2 diabetes in the UK. *Diabetic Medicine* 2006;**23**(5):516-521.

88. Pettitt DJ, Lisse JR, Knowler WC, Bennett PH. Mortality as a function of obesity and diabetes mellitus. *American Journal of Epidemiology* 1982;**115**(3):359-366.

89. Rosengren A, Welin L, Tsipogianni A, Wilhelmsen L. Impact of cardiovascular risk factors on coronary heart disease and mortality among middle aged diabetic men: a general population study. *BMJ* 1989;**299**(6708):1127-1131.

90. Ross C, Langer RD, Barrett-Connor E. Given diabetes, is fat better than thin? *Diabetes Care* 1997;**20**(4):650-652.

91. Sluik D, Boeing H, Montonen J, Pischon T, Kaaks R, Teucher B, Tjonneland A, Halkjaer J, Berentzen TL, Overvad K, Arriola L, Ardanaz E, Bendinelli B, Grioni S, Tumino R, Sacerdote C, Mattiello A, Spijkerman AM, van der AD, Beulens JW, van der Schouw YT, Nilsson PM, Hedblad B, Rolandsson O, Franks PW, Nothlings U. Associations between general and abdominal adiposity and mortality in individuals with diabetes mellitus. *American Journal of Epidemiology* 2011;**174**(1):22-34.

92. Lajous M, Bijon A, Fagherazzi G, Boutron-Ruault M-C, Balkau B, Clavel-Chapelon F, Hernán MA. Body mass index, diabetes, and mortality in French women: explaining away a "paradox". *Epidemiology* 2014;**25**(1):10-14.
93. Doehner W, Erdmann E, Cairns R, Clark AL, Dormandy JA, Ferrannini E, Anker SD. Inverse relation of body weight and weight change with mortality and morbidity in patients with type 2 diabetes and cardiovascular co-morbidity: an analysis of the PROactive study population. *International Journal of Cardiology* 2012;**162**(1):20-6.
94. Tseng C-H. Obesity paradox: Differential effects on cancer and noncancer mortality in patients with type 2 diabetes mellitus. *Atherosclerosis* 2013;**226**(1):186-192.
95. Ma SH, Park B-Y, Yang JJ, Jung E-J, Yeo Y, Whang Y, Chang S-H, Shin H-R, Kang D, Yoo K-Y. Interaction of body mass index and diabetes as modifiers of cardiovascular mortality in a cohort study. *Journal of Preventive Medicine and Public Health* 2012;**45**(6):394.
96. Perotto M, Panero F, Gruden G, Fornengo P, Lorenzati B, Barutta F, Ghezzo G, Amione C, Cavallo-Perin P, Bruno G. Obesity is associated with lower mortality risk in elderly diabetic subjects: the Casale Monferrato study. *Acta Diabetologica* 2013;**50**(4):563-568.
97. Murphy RA, Reinders I, Garcia ME, Eiriksdottir G, Launer LJ, Benediktsson R, Gudnason V, Jonsson PV, Harris TB. Adipose tissue, muscle, and function: potential mediators of associations between body weight and mortality in older adults with type 2 diabetes. *Diabetes Care* 2014:DC_140293.
98. Lee EY, Lee Y-h, Yi S-W, Shin S-A, Yi J-J. BMI and All-Cause Mortality in Normoglycemia, Impaired Fasting Glucose, Newly Diagnosed Diabetes, and Prevalent Diabetes: A Cohort Study. *Diabetes Care* 2017;**40**(8):1026-1033.
99. Xu H, Zhang M, Xu D, Zhang F, Yao B, Yan Y, Zhao N, Xu W, Qin G. Body mass index and the risk of mortality among Chinese adults with Type 2 diabetes. *Diabetic Medicine* 2018;**0**(ja).
100. Jenkins DA, Bowden J, Robinson HA, Sattar N, Loos RJF, Rutter MK, Sperrin M. Adiposity-Mortality Relationships in Type 2 Diabetes, Coronary Heart Disease and Cancer Subgroups in the UK Biobank, and Their Modification by Smoking. *Diabetes Care* 2018;**41**(9):1878-1886.
101. Bozorgmanesh M, Arshi B, Sheikholeslami F, Azizi F, Hadaegh F. No Obesity Paradox-BMI Incapable of Adequately Capturing the Relation of Obesity with All-Cause Mortality: An Inception Diabetes Cohort Study. *International Journal of Endocrinology* 2014;**2014**:9.
102. Paul S, Klein K, Majeed A, Khunti K. Longitudinal profiles of blood pressure, lipies and HbA1c and their association with vascular and mortality risks in patients with T2DM under cardioprotective medications. American Diabetes Association Scientific Congress. San Francisco, CA, USA, 2014.
103. Paul S, Best J, Klein K, Maggs D. Risk factors associated with hypoglycemia in patients treated with either exenatide once weekly or insulin glargine. 72nd Scientific Sessions of the American Diabetes Association, 2012.
104. WHO Expert Consultation. Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies. *Lancet* 2004;**363**(9403):157.
105. Kwon Y, Kim HJ, Park S, Park Y-G, Cho K-H. Body Mass Index-Related Mortality in Patients with Type 2 Diabetes and Heterogeneity in Obesity Paradox Studies: A Dose-Response Meta-Analysis. *PLoS ONE* 2017;**12**(1):e0168247.
106. Thomas G, Klein K, Paul S. Statistical challenges in analysing large longitudinal patient-level data: The danger of misleading clinical inferences with imputed data. *Journal of the Indian Society of Agricultural Statistics* 2014;**68**(2):39-54.
107. Adams KF, Schatzkin A, Harris TB, Kipnis V, Mouw T, Ballard-Barbash R, Hollenbeck A, Leitzmann MF. Overweight, Obesity, and Mortality in a Large Prospective Cohort of Persons 50 to 71 Years Old. *New England Journal of Medicine* 2006;**355**(8):763-78.
108. Ma SH, Park B-Y, Yang JJ, Jung E-J, Yeo Y, Whang Y, Chang S-H, Shin H-R, Kang D, Yoo K-Y, Park SK. Interaction of Body Mass Index and Diabetes as Modifiers of Cardiovascular

Mortality in a Cohort Study. *Journal of Preventive Medicine and Public Health* 2012;**45**(6):394-401.

109. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Statistics John Wiley & Sons, Inc., 2008.
110. Banack HR, Kaufman JS. Does selection bias explain the obesity paradox among individuals with cardiovascular disease? *Annals of Epidemiology* 2015;**25**(5):342-349.
111. Flegal KM, Graubard BI, Williamson DF, Cooper RS. Reverse Causation and Illness-related Weight Loss in Observational Studies of Body Weight and Mortality. *American Journal of Epidemiology* 2011;**173**(1):1-9.
112. Hainer V, Aldhoon-Hainerová I. Obesity paradox does exist. *Diabetes Care* 2013;**36**(Supplement 2):S276-S281.
113. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 1974;**66**(5):688.
114. Austin PC, Stuart EA. The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. *Statistical Methods in Medical Research* 2015;**26**(4):1654-1670.
115. Cattaneo MD. Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics* 2010;**155**(2):138-154.
116. Lee M-j. Treatment effects in sample selection models and their nonparametric estimation. *Journal of Econometrics* 2012;**167**(2):317-329.
117. Bonora E. Antidiabetic medications in overweight/obese patients with type 2 diabetes: drawbacks of current drugs and potential advantages of incretin-based treatment on body weight. *International Journal of Clinical Practice* 2007;**61**(154):19-28.
118. McFarlane SI. Antidiabetic medications and weight gain: Implications for the practicing physician. *Current Diabetes Reports* 2009;**9**(3):249-254.
119. Chrysant SG, Chrysant GS. New insights into the true nature of the obesity paradox and the lower cardiovascular risk. *Journal of the American Society of Hypertension* 2013;**7**(1):85-94.
120. Jee SH, Sull JW, Park J, Lee S-Y, Ohrr H, Guallar E, Samet JM. Body-mass index and mortality in Korean men and women. *New England Journal of Medicine* 2006;**355**(8):779-787.
121. Hjellvik V, Selmer R, Gjessing HK, Tverdal A, Vollset SE. Body mass index, smoking, and risk of death between 40 and 70 years of age in a Norwegian cohort of 32,727 women and 33,475 men. *European Journal of Epidemiology* 2013;**28**(1):35-43.
122. IMS Health Incorporated. The Health Improvement Network (THIN) database. http://www.csdmruk.imshealth.com/index.html Accessed 09-May, 2017.
123. Maguire A, Blak BT, Thompson M. The importance of defining periods of complete mortality reporting for research using automated data from primary care. *Pharmacoepidemiology and Drug Safety* 2009;**18**(1):76-83.
124. Blak BT, Thompson M, Dattani H, Bourke A. Generalisability of The Health Improvement Network (THIN) database: demographics, chronic disease prevalence and mortality rates. *Informatics in Primary Care* 2011;**19**(4):251-255.
125. Denburg MR, Haynes K, Shults J, Lewis JD, Leonard MB. Validation of The Health Improvement Network (THIN) database for epidemiologic studies of chronic kidney disease. *Pharmacoepidemiology and Drug Safety* 2011;**20**(11):1138-1149.
126. Townsend P, Phillimore P, Beattie A. *Health and Deprivation: Inequality and the North* Croom Helm, 1988.
127. Saint-Yves I. The Read Clinical Classification. *Health bulletin* 1992;**50**(6):422-427.
128. Read J. The Read clinical classification (Read codes). *British Homoeopathic Journal* 1991;**80**(1):14-20.

129. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, Smeeth L. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *International Journal of Epidemiology* 2015;**44**(3):827-836.

130. de Lusignan S, Liaw S-T, Dedman D, Khunti K, Sadek K, Jones S. An algorithm to improve diagnostic accuracy in diabetes in computerised problem orientated medical records (POMR) compared with an established algorithm developed in episode orientated records (EOMR). *Journal of Innovation in Health Informatics* 2015;**22**(2):255-264.

131. de Lusignan S, Sadek K, McDonald H, Horsfield P, Sadek NH, Tahir A, Desombre T, Khunti K. Call for consistent coding in diabetes mellitus using the Royal College of General Practitioners and NHS pragmatic classification of diabetes. *Informatics in Primary Care* 2012;**20**(2):103–13.

132. Hassan Sadek N, Sadek AR, Tahir A, Khunti K, Desombre T, de Lusignan S. Evaluating tools to support a new practical classification of diabetes: excellent control may represent misdiagnosis and omission from disease registers is associated with worse control. *International Journal of Clinical Practice* 2012;**66**(9):874-882.

133. Sadek AR, van Vlymen J, Khunti K, de Lusignan S. Automated identification of miscoded and misclassified cases of diabetes from computer records. *Diabetic Medicine* 2012;**29**(3):410-4.

134. Kang EM, Pinheiro SP, Hammad TA, Abou-Ali A. Evaluating the validity of clinical codes to identify cataract and glaucoma in the UK Clinical Practice Research Datalink. *Pharmacoepidemiology and Drug Safety* 2015;**24**(1):38-44.

135. Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: a systematic review. *British Journal of General Practice* 2010;**60**(572):e128-e136.

136. Moreno-Iribas C, Sayon-Orea C, Delfrade J, Ardanaz E, Gorricho J, Burgui R, Nuin M, Guevara M. Validity of type 2 diabetes diagnosis in a population-based electronic health record database. *BMC Medical Informatics and Decision Making* 2017;**17**(1):34.

137. Mamtani R, Haynes K, Finkelman BS, Scott FI, Lewis JD. Distinguishing incident and prevalent diabetes in an electronic medical records database. *Pharmacoepidemiology and Drug Safety* 2014;**23**(2):111-118.

138. de Lusignan S, Khunti K, Belsey J, Hattersley A, Van Vlymen J, Gallagher H, Millett C, Hague NJ, Tomson C, Harris K, Majeed A. A method of identifying and correcting miscoding, misclassification and misdiagnosis in diabetes: A pilot and validation study of routinely collected data. *Diabetic Medicine* 2010;**27**(2):203-209.

139. Friedman N, Geiger D, Goldszmidt M. Bayesian Network Classifiers. *Machine Learning*;**29**(2):131-163.

140. John GH, Langley P. Estimating continuous distributions in Bayesian classifiers. Proceedings of the Eleventh conference on Uncertainty in artificial intelligence: Morgan Kaufmann Publishers Inc., 1995;338-345.

141. Le Cessie S, Van Houwelingen JC. Ridge estimators in logistic regression. *Applied Statistics* 1992:191-201.

142. Hastie T, Tibshirani R. Classification by pairwise coupling. *The Annals of Statistics* 1998;**26**(2):451-471.

143. Keerthi SS, Shevade SK, Bhattacharyya C, Murthy KRK. Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Computation* 2001;**13**(3):637-649.

144. Platt JC. Fast training of support vector machines using sequential minimal optimization. In: Schoelkopf B, Burges C, Smola A, eds. *Advances in kernel methods- Support Vector Learning*. Cambridge, Massachusets: MIT Press, 1999;185-208.

145. Ruck DW, Rogers SK, Kabrisky M. Feature selection using a multilayer perceptron. *Journal of Neural Network Computing* 1990;**2**(2):40-48.

146. Witten IH, Frank E. *Data mining: practical machine learning tools and techniques*. Vol. 2nd. San Francisco, Calif: Morgan Kaufman, 2005.

147. Loh W-Y. Improving the precision of classification trees. *The Annals of Applied Statistics* 2009;**3**(4):1710-1737.

148. Drazin S, Montag M. Decision tree analysis using WEKA. *Machine Learning-Project II, University of Miami* 2012:1-3.

149. Holte RC. Very simple classification rules perform well on most commonly used datasets. *Machine Learning* 1993;**11**(1):63-90.

150. Sagreiya H, Altman RB. The utility of general purpose versus specialty clinical databases for research: warfarin dose estimation from extracted clinical variables. *Journal of Biomedical Informatics* 2010;**43**(5):747-751.

151. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, Lai AM. A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association* 2014;**21**(2):221-230.

152. Tate AR, Beloff N, Al-Radwan B, Wickson J, Puri S, Williams T, Van Staa T, Bleach A. Exploiting the potential of large databases of electronic health records for research using rapid search algorithms and an intuitive query interface. *Journal of the American Medical Informatics Association* 2014;**21**(2):292-298.

153. Kandula S, Zeng-Treitler Q, Chen L, Salomon WL, Bray BE. A bootstrapping algorithm to improve cohort identification using structured data. *Journal of Biomedical Informatics* 2011;**44**:S63-S68.

154. Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. *British Journal of Clinical Pharmacology* 2010;**69**(1):4-14.

155. Hammad TA, Margulis AV, Ding Y, Strazzeri MM, Epperly H. Determining the predictive value of Read codes to identify congenital cardiac malformations in the UK Clinical Practice Research Datalink. *Pharmacoepidemiology and Drug Safety* 2013;**22**(11):1233-1238.

156. Stone MA, Camosso-Stefinovic J, Wilkinson J, de Lusignan S, Hattersley AT, Khunti K. Incorrect and incomplete coding and classification of diabetes: a systematic review. *Diabetic Medicine* 2010;**27**(5):491-497.

157. Seidu S, Davies MJ, Mostafa S, de Lusignan S, Khunti K. Prevalence and characteristics in coding, classification and diagnosis of diabetes in primary care. *Postgraduate Medical Journal* 2014;**90**(1059):13-17.

158. Holt TA, Gunnarsson CL, Cload PA, Ross SD. Identification of undiagnosed diabetes and quality of diabetes care in the United States: cross-sectional study of 11.5 million primary care electronic records. *CMAJ Open* 2014;**2**(4):E248-E255.

159. Holt TA, Stables D, Hippisley-Cox J, O'Hanlon S, Majeed A. Identifying undiagnosed diabetes: cross-sectional survey of 3.6 million patients' electronic records. *The British Journal of General Practice* 2008;**58**(548):192-196.

160. Magliano DJ, Zimmet P, Shaw J. US trends for diabetes prevalence among adults. *JAMA* 2016;**315**(7):705-705.

161. Gray J, Orr D, Majeed A. Use of Read codes in diabetes management in a south London primary care group: implications for establishing disease registers. *BMJ* 2003;**326**(7399):1130.

162. Rollason W, Khunti K, de Lusignan S. Variation in the recording of diabetes diagnostic data in primary care computer systems: implications for the quality of care. *Informatics in Primary Care* 2009;**17**(2):113.

163. Lycett D, Nichols L, Ryan R, Farley A, Roalfe A, Mohammed MA, Szatkowski L, Coleman T, Morris R, Farmer A, Aveyard P. The association between smoking cessation and glycaemic control in patients with type 2 diabetes: A THIN database cohort study. *The Lancet Diabetes and Endocrinology* 2015;**3**(6):423-430.

164. American Diabetes Association. Standards of Medical Care in Diabetes—2015 Abridged for Primary Care Providers. *Clinical Diabetes* 2015;**33**(2):97-111.

165. Hall MA, Smith LA. Practical feature subset selection for machine learning. *Proceedings of the 21st Australasian Computer Science Conference (ACSC)*. Vol. 20. Perth: Springer, 1998;181-191.

166. Senliol B, Gulgezen G, Yu L, Cataltepe Z. Fast Correlation Based Filter (FCBF) with a different search strategy. 2008 23rd International Symposium on Computer and Information Sciences, 2008;1-4.

167. Schmidt M, Roux NL, Bach F. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming* 2017;**162**(1-2):83-112.

168. Cortes C, Vapnik V. Support-vector networks. *Machine Learning* 1995;**20**.

169. Wu T-F, Lin C-J, Weng RC. Probability estimates for multi-class classification by pairwise coupling. *The Journal of Machine Learning Research* 2004;**5**:975-1005.

170. Tapak L, Mahjub H, Hamidi O, Poorolajal J. Real-Data Comparison of Data Mining Methods in Prediction of Diabetes in Iran. *Healthcare Informatics Research* 2013;**19**(3):177-185.

171. Mani S, Chen Y, Elasy T, Clayton W, Denny J. Type 2 diabetes risk forecasting from EMR data using machine learning. *AMIA Annu Symp Proc* 2012;**2012**:606-15.

172. Zheng T, Xie W, Xu L, He X, Zhang Y, You M, Yang G, Chen Y. A machine learning-based framework to identify type 2 diabetes through electronic health records. *International Journal of Medical Informatics* 2017;**97**:120-127.

173. Razavian N, Blecker S, Schmidt AM, Smith-McLallen A, Nigam S, Sontag D. Population-Level Prediction of Type 2 Diabetes From Claims Data and Analysis of Risk Factors. *Big Data* 2015;**3**(4):277-287.

174. Khunti K, Davies M, Majeed A, Thorsted BL, Wolden ML, Paul SK. Hypoglycemia and Risk of Cardiovascular Disease and All-Cause Mortality in Insulin-Treated People With Type 1 and Type 2 Diabetes: A Cohort Study. *Diabetes Care* 2014.

175. Parsons LS. Reducing bias in a propensity score matched-pair sample using greedy matching techniques. Proceedings of the Twenty-sixth Annual SAS Users group international conference: SAS Institute Inc Cary, NC, 2001;214-226.

176. Iacus SM, King G, Porro G. Multivariate Matching Methods That Are Monotonic Imbalance Bounding. *Journal of the American Statistical Association* 2011;**106**(493):345-361.

177. Carpenter JR, Kenward MG. *Multiple Imputation and its Application* John Wiley & Sons, Ltd, 2013.

178. Molenberghs G, Kenward MG. *Missing Data in Clinical Studies*. Missing Data in Clinical Studies John Wiley & Sons, Ltd, 2007.

179. Welch CA, Petersen I, Bartlett JW, White IR, Marston L, Morris RW, Nazareth I, Walters K, Carpenter J. Evaluation of two-fold fully conditional specification multiple imputation for longitudinal electronic health record data. *Statistics in Medicine* 2014;**33**(21):3725-3737.

180. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. 2nd ed. Hoboken, NJ, USA:: John Wiley & Sons, Inc., 2014.

181. Ng M, Fleming T, Robinson M, Thomson B, Graetz N, Margono C, Mullany EC, Biryukov S, Abbafati C, Abera SF. Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet* 2014;**384**(9945):766-781.

182. Kodama S, Horikawa C, Fujihara K, Heianza Y, Hirasawa R, Yachi Y, Sugawara A, Tanaka S, Shimano H, Iida KT, Saito K, Sone H. Comparisons of the Strength of Associations With Future Type 2 Diabetes Risk Among Anthropometric Obesity Indicators, Including Waist-to-Height Ratio: A Meta-Analysis. *American Journal of Epidemiology* 2012;**176**(11):959-969.

183. Garber AJ. Obesity and type 2 diabetes: which patients are at risk? *Diabetes, Obesity and Metabolism* 2012;**14**(5):399-408.

184. Misra A. Ethnic-Specific Criteria for Classification of Body Mass Index: A Perspective for Asian Indians and American Diabetes Association Position Statement. *Diabetes Technology & Therapeutics* 2015;**17**(9):667-71.

185. Chiu M, Austin PC, Manuel DG, Shah BR, Tu JV. Deriving Ethnic-Specific BMI Cutoff Points for Assessing Diabetes Risk. *Diabetes Care* 2011;**34**(8):1741-1748.

186. Misra A, Vikram NK, Gupta R, Pandey RM, Wasir JS, Gupta VP. Waist circumference cutoff points and action levels for Asian Indians for identification of abdominal obesity. *International Journal of Obesity* 2006;**30**(1):106-11.

187. Bodicoat DH, Gray LJ, Henson J, Webb D, Guru A, Misra A, Gupta R, Vikram N, Sattar N, Davies MJ, Khunti K. Body Mass Index and Waist Circumference Cut-Points in Multi-Ethnic Populations from the UK and India: The ADDITION-Leicester, Jaipur Heart Watch and New Delhi Cross-Sectional Studies. *PLoS One* 2014;**9**(3):e90813.

188. Decode-Decoda Study Group. Age, body mass index and type 2 diabetes—associations modified by ethnicity. *Diabetologia* 2003;**46**(8):1063-1070.

189. Gary King, Michael Tomz, Wittenberg J. Making the Most of Statistical Analyses: Improving Interpretation and Presentation. *American Journal of Political Science* 2000;**44**(2):347-361.

190. Zelner BA. Using simulation to interpret results from logit, probit, and other nonlinear models. *Strategic Management Journal* 2009;**30**(12):1335-1348.

191. James GD, Baker P, Badrick E, Mathur R, Hull S, Robson J. Ethnic and social disparity in glycaemic control in type 2 diabetes; cohort study in general practice 2004–9. *Journal of the Royal Society of Medicine* 2012;**105**(7):300-308.

192. Paul SK, Klein K, Thorsted BL, Wolden ML, Khunti K. Delay in treatment intensification increases the risks of cardiovascular events in patients with type 2 diabetes. *Cardiovascular Diabetology* 2015;**14**(1):1.

193. Paul S, Klein K, Majeed A, Khunti K. Association of smoking and concomitant use of metformin with cardiovascular events and mortality in people newly diagnosed with type 2 diabetes. *Journal of Diabetes* 2015;**201**(5).

194. Paul S, Thomas G, Majeed A, Khunti K, Klein K. Women develop type 2 diabetes at a higher body mass index than men. *Diabetologia* 2012;**55(**5):1556-1557.

195. Peters SA, Huxley RR, Woodward M. Sex differences in body anthropometry and composition in individuals with and without diabetes in the UK Biobank. *BMJ Open* 2016;**6**(1):e010007.

196. Ntuk UE, Gill JMR, Mackay DF, Sattar N, Pell JP. Ethnic-Specific Obesity Cutoffs for Diabetes Risk: Cross-sectional Study of 490,288 UK Biobank Participants. *Diabetes Care* 2014;**37**(9):2500-2507.

197. Mathur R, Bhaskaran K, Chaturvedi N, Leon DA, vanStaa T, Grundy E, Smeeth L. Completeness and usability of ethnicity data in UK-based primary care and hospital databases. *Journal of Public Health* 2014;**36**(4):684-692.

198. Oldroyd J, Banerjee M, Heald A, Cruickshank K. Diabetes and ethnic minorities. *Postgraduate Medical Journal* 2005;**81**(958):486-490.

199. Hsu WC, Araneta MRG, Kanaya AM, Chiang JL, Fujimoto W. BMI Cut Points to Identify At-Risk Asian Americans for Type 2 Diabetes Screening. *Diabetes Care* 2015;**38**(1):150-158.

200. Carulli L, Rondinella S, Lombardini S, Canedi I, Loria P, Carulli N. Review article: diabetes, genetics and ethnicity. *Alimentary Pharmacology & Therapeutics* 2005;**22**(Suppl 2):16-9.

201. Murea M, Ma L, Freedman BI. Genetic and environmental factors associated with type 2 diabetes and diabetic vascular complications. *Review of Diabetic Studies* 2012;**9**(1):6-22.

202. McBean AM, Li S, Gilbertson DT, Collins AJ. Differences in Diabetes Prevalence, Incidence, and Mortality Among the Elderly of Four Racial/Ethnic Groups: Whites, Blacks, Hispanics, and Asians. *Diabetes Care* 2004;**27**(10):2317-2324.

203. McNeely MJ, Boyko EJ. Type 2 Diabetes Prevalence in Asian Americans: Results of a national health survey. *Diabetes Care* 2004;**27**(1):66-69.

204. Shah AD, Langenberg C, Rapsomaniki E, Denaxas S, Pujades-Rodriguez M, Gale CP, Deanfield J, Smeeth L, Timmis A, Hemingway H. Type 2 diabetes and incidence of cardiovascular diseases: a cohort study in 1·9 million people. *The Lancet Diabetes & Endocrinology* 2015;**3**(2):105-113.

205. U.K. Prospective Diabetes Study Group. Ethnicity and Cardiovascular Disease: The incidence of myocardial infarction in White, South Asian, and Afro-Caribbean patients with type 2 diabetes (U.K. Prospective Diabetes Study 32). *Diabetes Care* 1998;**21**(8):1271-1277.

206. Spanakis EK, Golden SH. Race/ethnic difference in diabetes and diabetic complications. *Current Diabetes Reports* 2013;**13**(6):10.1007/s11892-013-0421-9.

207. Lanting LC, Joung IMA, Mackenbach JP, Lamberts SWJ, Bootsma AH. Ethnic Differences in Mortality, End-Stage Complications, and Quality of Care Among Diabetic Patients: A review. *Diabetes Care* 2005;**28**(9):2280-2288.

208. Young BA, Maynard C, Boyko EJ. Racial Differences in Diabetic Nephropathy, Cardiovascular Disease, and Mortality in a National Population of Veterans. *Diabetes Care* 2003;**26**(8):2392-2399.

209. Bellary S, O'Hare JP, Raymond NT, Mughal S, Hanif WM, Jones A, Kumar S, Barnett AH. Premature cardiovascular events and mortality in South Asians with type 2 diabetes in the United Kingdom Asian Diabetes Study - effect of ethnicity on risk. *Current Medical Research and Opinion* 2010;**26**(8):1873-9.

210. Kou S, Cao JY, Yeo S, Holmes-Walker DJ, Lau SL, Gunton JE. Ethnicity influences cardiovascular outcomes and complications in patients with type 2 diabetes. *Journal of Diabetes and Its Complications* 2018;**32**(2):144-149.

211. Tillin T, Hughes AD, Mayet J, Whincup P, Sattar N, Forouhi NG, McKeigue PM, Chaturvedi N. The Relationship Between Metabolic Risk Factors and Incident Cardiovascular Disease in Europeans, South Asians, and African Caribbeans: SABRE (Southall and Brent Revisited)—A Prospective Population-Based Study. *Journal of the American College of Cardiology* 2013;**61**(17):1777-1786.

212. Koshizaka M, Lopes RD, Newby LK, Clare RM, Schulte PJ, Tricoci P, Mahaffey KW, Ogawa H, Moliterno DJ, Giugliano RP, Huber K, James S, Harrington RA, Alexander JH. Obesity, Diabetes, and Acute Coronary Syndrome: Differences Between Asians and Whites. *The American Journal of Medicine* 2017;**130**(10):1170-1176.

213. Gholap N, Davies M, Patel K, Sattar N, Khunti K. Type 2 diabetes and cardiovascular disease in South Asians. *Primary Care Diabetes* 2011;**5**(1):45-56.

214. Fernando E, Razak F, Lear SA, Anand SS. Cardiovascular Disease in South Asian Migrants. *Canadian Journal of Cardiology* 2015;**31**(9):1139-1150.

215. Hubert HB, Feinleib M, McNamara PM, Castelli WP. Obesity as an independent risk factor for cardiovascular disease: a 26-year follow-up of participants in the Framingham Heart Study. *Circulation* 1983;**67**(5):968-977.

216. Poirier P, Giles TD, Bray GA, Hong Y, Stern JS, Pi-Sunyer FX, Eckel RH. Obesity and Cardiovascular Disease: Pathophysiology, Evaluation, and Effect of Weight Loss: An Update of the 1997 American Heart Association Scientific Statement on Obesity and Heart Disease From the Obesity Committee of the Council on Nutrition, Physical Activity, and Metabolism. *Circulation* 2006;**113**(6):898-918.

217. Ni Mhurchu C, Rodgers A, Pan WH, Gu DF, Woodward M, Asia Pacific Cohort Studies Collaboration. Body mass index and cardiovascular disease in the Asia-Pacific Region: an overview of 33 cohorts involving 310 000 participants. *International Journal of Epidemiology* 2004;**33**(4):751-8.

218. Zaccardi F, Dhalwani NN, Papamargaritis D, Webb DR, Murphy GJ, Davies MJ, Khunti K. Nonlinear association of BMI with all-cause and cardiovascular mortality in type 2 diabetes mellitus: a systematic review and meta-analysis of 414,587 participants in prospective studies. *Diabetologia* 2017;**60**(2):240-248.

219. The Look AHEAD Research Group. Cardiovascular Effects of Intensive Lifestyle Intervention in Type 2 Diabetes. *New England Journal of Medicine* 2013;**369**(2):145-154.

220. Turner DA, Paul S, Stone MA, Juarez-Garcia A, Squire I, Khunti K. Cost-effectiveness of a disease management programme for secondary prevention of coronary heart disease and heart failure in primary care. *Heart* 2008;**94**(12):1601-6.

221. Yusuf S, Hawken S, Ôunpuu S, Bautista L, Franzosi MG, Commerford P, Lang CC, Rumboldt Z, Onen CL, Lisheng L, Tanomsup S, Wangai P, Razak F, Sharma AM, Anand SS. Obesity and the risk of myocardial infarction in 27 000 participants from 52 countries: a case-control study. *The Lancet* 2005;**366**(9497):1640-1649.
222. Goyal A, Nimmakayala KR, Zonszein J. Is There a Paradox in Obesity? *Cardiology in Review* 2014;**22**(4):163–170.
223. de Fine Olivarius N, Siersma VD, Koster-Rasmussen R, Heitmann BL, Waldorff FB. Weight changes following the diagnosis of type 2 diabetes: the impact of recent and past weight history before diagnosis. results from the Danish Diabetes Care in General Practice (DCGP) study. *PloS ONE* 2015;**10**(4).
224. Heianza Y, Arase Y, Kodama S, Tsuji H, Tanaka S, Saito K, Hara S, Sone H. Trajectory of body mass index before the development of type 2 diabetes in Japanese men: Toranomon Hospital Health Management Center Study 15. *Journal of Diabetes Investigation* 2015;**6**(3):289-294.
225. Looker HC, Knowler WC, Hanson RL. Changes in BMI and weight before and after the development of type 2 diabetes. *Diabetes Care* 2001;**24**(11):1917-22.
226. Vistisen D, Witte DR, Tabák AG, Herder C, Brunner EJ, Kivimäki M, Færch K. Patterns of obesity development before the diagnosis of type 2 diabetes: The Whitehall II Cohort Study. *PLoS Medicine* 2014;**11**(2):e1001602.
227. Wannamethee SG, Shaper AG. Weight change and duration of overweight and obesity in the incidence of type 2 diabetes. *Diabetes Care* 1999;**22**(8):1266-1272.
228. Owusu Adjah ES, Montvida O, Agbeve J, Paul SK. Data Mining Approach to Identify Disease Cohorts from Primary Care Electronic Medical Records: A Case of Diabetes Mellitus. *The Open Bioinformatics Journal* 2017;**10**:16-27.
229. Aucott LS, Philip S, Avenell A, Afolabi E, Sattar N, Wild S. Patterns of weight change after the diagnosis of type 2 diabetes in Scotland and their relationship with glycaemic control, mortality and cardiovascular outcomes: a retrospective cohort study. *BMJ Open* 2016;**6**(7):e010836.
230. Pi-Sunyer FX. Weight loss in type 2 diabetic patients. *Diabetes Care* 2005;**28**(6):1526-1527.
231. Carnethon MR, Rasmussen-Torvik LJ, Palaniappan L. The obesity paradox in diabetes. *Current Cardiology Reports* 2014;**16**(2):446.
232. Banack HR, Kaufman JS. The "Obesity Paradox" explained. *Epidemiology* 2013;**24**(3):461-462.
233. Batterham M, Tapsell LC, Charlton KE. Baseline characteristics associated with different BMI trajectories in weight loss trials: a case for better targeting of interventions. *European Journal of Clinical Nutrition* 2016;**70**(2):207-211.
234. Owusu Adjah ES, Samanta M, Shaw JE, Majeed A, Khunti K, Paul SK. Weight loss and mortality risk in patients with different adiposity at diagnosis of type 2 diabetes: a longitudinal cohort study. *Nutrition & diabetes* 2018;**8**(1):37.
235. George J, Mathur R, Shah AD, Pujades-Rodriguez M, Denaxas S, Smeeth L, Timmis A, Hemingway H. Ethnicity and the first diagnosis of a wide range of cardiovascular diseases: Associations in a linked electronic health record cohort of 1 million patients. *PLoS ONE* 2017;**12**(6):e0178945.
236. Lavie CJ, De Schutter A, Patel D, Artham SM, Milani RV. Body Composition and Coronary Heart Disease Mortality—An Obesity or a Lean Paradox? *Mayo Clinic Proceedings* 2011;**86**(9):857-864.
237. Lavie CJ, Milani RV, Ventura HO. Obesity and the "Obesity Paradox" in cardiovascular diseases. *Clinical Pharmacology & Therapeutics* 2011;**90**(1):23-25.
238. Chen Y, Copeland WK, Vedanthan R, Grant E, Lee JE, Gu D, Gupta PC, Ramadas K, Inoue M, Tsugane S, Tamakoshi A, Gao Y-T, Yuan J-M, Shu X-O, Ozasa K, Tsuji I, Kakizaki M, Tanaka H, Nishino Y, Chen C-J, Wang R, Yoo K-Y, Ahn Y-O, Ahsan H, Pan W-H, Chen C-S, Pednekar MS, Sauvaget C, Sasazuki S, Yang G, Koh W-P, Xiang Y-B, Ohishi W,

Watanabe T, Sugawara Y, Matsuo K, You S-L, Park SK, Kim D-H, Parvez F, Chuang S-Y, Ge W, Rolland B, McLerran D, Sinha R, Thornquist M, Kang D, Feng Z, Boffetta P, Zheng W, He J, Potter JD. Association between body mass index and cardiovascular disease mortality in east Asians and south Asians: pooled analysis of prospective data from the Asia Cohort Consortium. *BMJ* 2013;**347**.

239. Wright AK, Kontopantelis E, Emsley R, Buchan I, Sattar N, Rutter MK, Ashcroft DM. Life Expectancy and Cause-Specific Mortality in Type 2 Diabetes: A Population-Based Cohort Study Quantifying Relationships in Ethnic Subgroups. *Diabetes Care* 2016.
240. Ho AK, Bartels CM, Thorpe CT, Pandhi N, Smith MA, Johnson HM. Achieving Weight Loss and Hypertension Control Among Obese Adults: A US Multidisciplinary Group Practice Observational Study. *American Journal of Hypertension* 2016;**29**(8):984-991.
241. Lavie CJ, McAuley PA, Church TS, Milani RV, Blair SN. Obesity and Cardiovascular Diseases. *Journal of the American College of Cardiology* 2014;**63**(14):1345-1354.
242. Zheng Y, Song M, Manson JE, Giovannucci EL, Hu FB. Group-Based Trajectory of Body Shape From Ages 5 to 55 Years and Cardiometabolic Disease Risk in 2 US Cohorts. *American Journal of Epidemiology* 2017;**186**(11):1246-1255.
243. Canoy D, Wareham N, Luben R, Welch A, Bingham S, Day N, Khaw KT. Cigarette smoking and fat distribution in 21,828 British men and women: a population-based study. *Obesity Research* 2005;**13**(8):1466-75.
244. Jerez JM, Molina I, García-Laencina PJ, Alba E, Ribelles N, Martín M, Franco L. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine* 2010;**50**(2):105-115.
245. Biering K, Hjollund NH, Frydenberg M. Using multiple imputation to deal with missing data and attrition in longitudinal studies with repeated measures of patient-reported outcomes. *Clinical Epidemiology* 2015;**7**:91-106.
246. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009;**338**:b2393.
247. Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, Frangakis C, Hogan JW, Molenberghs G, Murphy SA, Neaton JD, Rotnitzky A, Scharfstein D, Shih WJ, Siegel JP, Stern H. The Prevention and Treatment of Missing Data in Clinical Trials. *New England Journal of Medicine* 2012;**367**(14):1355-1360.
248. Wells BJ, Chagin KM, Nowacki AS, Kattan MW. Strategies for handling missing data in electronic health record derived data. *EGEMS (Washington, DC)* 2013;**1**(3):1035-1035.
249. Lu CY, Madden JM, Lakoma MD, Soumerai SB, Rusinak D. Missing clinical and behavioral health data in a large electronic health record (EHR) system. *Journal of the American Medical Informatics Association* 2016;**23**(6):1143-1149.
250. Mackinnon A. The use and reporting of multiple imputation in medical research – a review. *Journal of Internal Medicine* 2010;**268**(6):586-593.

Appendices

Appendix A

Appendix Table 1: T2DM Read codes used in extracting cohort of patients with T2DM

Read Code	Description	Read Code	Description
C100100	Diabetes mellitus, adult onset, no mention of complication	C109411	Type II diabetes mellitus with ulcer
C100111	Maturity onset diabetes	C109412	Type 2 diabetes mellitus with ulcer
C100112	Non-insulin dependent diabetes mellitus	C109500	Non-insulin dependent diabetes mellitus with gangrene
C101100	Diabetes mellitus, adult onset, with ketoacidosis	C109511	Type II diabetes mellitus with gangrene
C102100	Diabetes mellitus, adult onset, with hyperosmolar coma	C109512	Type 2 diabetes mellitus with gangrene
C103100	Diabetes mellitus, adult onset, with ketoacidotic coma	C109600	Non-insulin-dependent diabetes mellitus with retinopathy
C104100	Diabetes mellitus, adult onset, with renal manifestation	C109611	Type II diabetes mellitus with retinopathy
C105100	Diabetes mellitus, adult onset, + ophthalmic manifestation	C109612	Type 2 diabetes mellitus with retinopathy
C106100	Diabetes mellitus, adult onset, + neurological manifestation	C109700	Non-insulin dependent diabetes mellitus - poor control
C107400	NIDDM with peripheral circulatory disorder	C109711	Type II diabetes mellitus - poor control
C109.00	Non-insulin dependent diabetes mellitus	C109712	Type 2 diabetes mellitus - poor control
C109.11	NIDDM - Non-insulin dependent diabetes mellitus	C109900	Non-insulin-dependent diabetes mellitus without complication
C109.12	Type 2 diabetes mellitus	C109911	Type II diabetes mellitus without complication
C109.13	Type II diabetes mellitus	C109912	Type 2 diabetes mellitus without complication
C109000	Non-insulin-dependent diabetes mellitus with renal comps	C109A00	Non-insulin dependent diabetes mellitus with mononeuropathy
C109011	Type II diabetes mellitus with renal complications	C109A11	Type II diabetes mellitus with mononeuropathy
C109012	Type 2 diabetes mellitus with renal complications	C109A12	Type 2 diabetes mellitus with mononeuropathy
C109100	Non-insulin-dependent diabetes mellitus with ophthalm comps	C109B00	Non-insulin dependent diabetes mellitus with polyneuropathy
C109111	Type II diabetes mellitus with ophthalmic complications	C109B11	Type II diabetes mellitus with polyneuropathy
C109112	Type 2 diabetes mellitus with ophthalmic complications	C109B12	Type 2 diabetes mellitus with polyneuropathy
C109200	Non-insulin-dependent diabetes mellitus with neuro comps	C109C00	Non-insulin dependent diabetes mellitus with nephropathy
C109211	Type II diabetes mellitus with neurological complications	C109C11	Type II diabetes mellitus with nephropathy

C109212	Type 2 diabetes mellitus with neurological complications	C109C12	Type 2 diabetes mellitus with nephropathy
C109300	Non-insulin-dependent diabetes mellitus with multiple comps	C109D00	Non-insulin dependent diabetes mellitus with hypoglyca coma
C109311	Type II diabetes mellitus with multiple complications	C109D11	Type II diabetes mellitus with hypoglycaemic coma
C109312	Type 2 diabetes mellitus with multiple complications	C109D12	Type 2 diabetes mellitus with hypoglycaemic coma
C109400	Non-insulin dependent diabetes mellitus with ulcer	C109E00	Non-insulin depend diabetes mellitus with diabetic cataract
C109E11	Type II diabetes mellitus with diabetic cataract	C10F711	Type II diabetes mellitus - poor control
C109E12	Type 2 diabetes mellitus with diabetic cataract	C10F900	Type 2 diabetes mellitus without complication
C109F00	Non-insulin-dependent d m with peripheral angiopath	C10F911	Type II diabetes mellitus without complication
C109F11	Type II diabetes mellitus with peripheral angiopathy	C10FA00	Type 2 diabetes mellitus with mononeuropathy
C109F12	Type 2 diabetes mellitus with peripheral angiopathy	C10FA11	Type II diabetes mellitus with mononeuropathy
C109G00	Non-insulin dependent diabetes mellitus with arthropathy	C10FB00	Type 2 diabetes mellitus with polyneuropathy
C109G11	Type II diabetes mellitus with arthropathy	C10FB11	Type II diabetes mellitus with polyneuropathy
C109G12	Type 2 diabetes mellitus with arthropathy	C10FC00	Type 2 diabetes mellitus with nephropathy
C109H00	Non-insulin dependent d m with neuropathic arthropathy	C10FC11	Type II diabetes mellitus with nephropathy
C109H11	Type II diabetes mellitus with neuropathic arthropathy	C10FD00	Type 2 diabetes mellitus with hypoglycaemic coma
C109H12	Type 2 diabetes mellitus with neuropathic arthropathy	C10FD11	Type II diabetes mellitus with hypoglycaemic coma
C109J00	Insulin treated Type 2 diabetes mellitus	C10FE00	Type 2 diabetes mellitus with diabetic cataract
C109J11	Insulin treated non-insulin dependent diabetes mellitus	C10FE11	Type II diabetes mellitus with diabetic cataract
C109J12	Insulin treated Type II diabetes mellitus	C10FF00	Type 2 diabetes mellitus with peripheral angiopathy
C109K00	Hyperosmolar non-ketotic state in type 2 diabetes mellitus	C10FF11	Type II diabetes mellitus with peripheral angiopathy
C10F.00	Type 2 diabetes mellitus	C10FG00	Type 2 diabetes mellitus with arthropathy
C10F.11	Type II diabetes mellitus	C10FG11	Type II diabetes mellitus with arthropathy
C10F000	Type 2 diabetes mellitus with renal complications	C10FH00	Type 2 diabetes mellitus with neuropathic arthropathy
C10F011	Type II diabetes mellitus with renal complications	C10FH11	Type II diabetes mellitus with neuropathic arthropathy
C10F100	Type 2 diabetes mellitus with ophthalmic complications	C10FJ00	Insulin treated Type 2 diabetes mellitus
C10F111	Type II diabetes mellitus with ophthalmic complications	C10FJ11	Insulin treated Type II diabetes mellitus
C10F200	Type 2 diabetes mellitus with neurological complications	C10FK00	Hyperosmolar non-ketotic state in type 2 diabetes mellitus
C10F211	Type II diabetes mellitus with neurological complications	C10FK11	Hyperosmolar non-ketotic state in type II diabetes mellitus
C10F300	Type 2 diabetes mellitus with multiple complications	C10FL00	Type 2 diabetes mellitus with persistent proteinuria
C10F311	Type II diabetes mellitus with multiple complications	C10FL11	Type II diabetes mellitus with persistent proteinuria
C10F400	Type 2 diabetes mellitus with ulcer	C10FM00	Type 2 diabetes mellitus with persistent microalbuminuria
C10F411	Type II diabetes mellitus with ulcer	C10FM11	Type II diabetes mellitus with persistent microalbuminuria

C10F500	Type 2 diabetes mellitus with gangrene	C10FN00	Type 2 diabetes mellitus with ketoacidosis
C10F511	Type II diabetes mellitus with gangrene	C10FN11	Type II diabetes mellitus with ketoacidosis
C10F600	Type 2 diabetes mellitus with retinopathy	C10FP00	Type 2 diabetes mellitus with ketoacidotic coma
C10F611	Type II diabetes mellitus with retinopathy	C10FP11	Type II diabetes mellitus with ketoacidotic coma
C10F700	Type 2 diabetes mellitus - poor control	C10FQ00	Type 2 diabetes mellitus with exudative maculopathy
C10FQ11	Type II diabetes mellitus with exudative maculopathy	C10P100	Type II diabetes mellitus in remission
C10FR00	Type 2 diabetes mellitus with gastroparesis	C10P111	Type 2 diabetes mellitus in remission
C10FR11	Type II diabetes mellitus with gastroparesis	C10y100	Diabetes mellitus, adult, + other specified manifestation
C10z100	Diabetes mellitus, adult onset, + unspecified complication		

Appendix Table 2: T1DM Read codes used in extracting cohort of patients with T1DM

Read Code	Description	Read Code	Description
C100000	Diabetes mellitus, juvenile type, no mention of complication	C108H00	Insulin dependent diabetes mellitus with arthropathy
C100011	Insulin dependent diabetes mellitus	C108H11	Type I diabetes mellitus with arthropathy
C101000	Diabetes mellitus, juvenile type, with ketoacidosis	C108H12	Type 1 diabetes mellitus with arthropathy
C102000	Diabetes mellitus, juvenile type, with hyperosmolar coma	C108J00	Insulin dependent diab mell with neuropathic arthropathy
C103000	Diabetes mellitus, juvenile type, with ketoacidotic coma	C108J11	Type I diabetes mellitus with neuropathic arthropathy
C104000	Diabetes mellitus, juvenile type, with renal manifestation	C108J12	Type 1 diabetes mellitus with neuropathic arthropathy
C105000	Diabetes mellitus, juvenile type, + ophthalmic manifestation	C10E.00	Type 1 diabetes mellitus
C106000	Diabetes mellitus, juvenile, + neurological manifestation	C10E.11	Type I diabetes mellitus
C107000	Diabetes mellitus, juvenile +peripheral circulatory disorder	C10E.12	Insulin dependent diabetes mellitus
C107300	IDDM with peripheral circulatory disorder	C10E000	Type 1 diabetes mellitus with renal complications
C108.00	Insulin dependent diabetes mellitus	C10E011	Type I diabetes mellitus with renal complications
C108.11	IDDM-Insulin dependent diabetes mellitus	C10E012	Insulin-dependent diabetes mellitus with renal complications
C108.12	Type 1 diabetes mellitus	C10E100	Type 1 diabetes mellitus with ophthalmic complications
C108.13	Type I diabetes mellitus	C10E111	Type I diabetes mellitus with ophthalmic complications
C108000	Insulin-dependent diabetes mellitus with renal complications	C10E112	Insulin-dependent diabetes mellitus with ophthalmic comps
C108011	Type I diabetes mellitus with renal complications	C10E200	Type 1 diabetes mellitus with neurological complications
C108012	Type 1 diabetes mellitus with renal complications	C10E211	Type I diabetes mellitus with neurological complications
C108100	Insulin-dependent diabetes mellitus with ophthalmic comps	C10E212	Insulin-dependent diabetes mellitus with neurological comps
C108111	Type I diabetes mellitus with ophthalmic complications	C10E300	Type 1 diabetes mellitus with multiple complications
C108112	Type 1 diabetes mellitus with ophthalmic complications	C10E311	Type I diabetes mellitus with multiple complications
C108200	Insulin-dependent diabetes mellitus with neurological comps	C10E312	Insulin dependent diabetes mellitus with multiple complicat
C108211	Type I diabetes mellitus with neurological complications	C10E400	Unstable type 1 diabetes mellitus
C108212	Type 1 diabetes mellitus with neurological complications	C10E411	Unstable type I diabetes mellitus
C108300	Insulin dependent diabetes mellitus with multiple complicatn	C10E412	Unstable insulin dependent diabetes mellitus
C108311	Type I diabetes mellitus with multiple complications	C10E500	Type 1 diabetes mellitus with ulcer
C108312	Type 1 diabetes mellitus with multiple complications	C10E511	Type I diabetes mellitus with ulcer
C108400	Unstable insulin dependent diabetes mellitus	C10E512	Insulin dependent diabetes mellitus with ulcer
C108411	Unstable type I diabetes mellitus	C10E600	Type 1 diabetes mellitus with gangrene

C108412	Unstable type 1 diabetes mellitus	C10E611	Type I diabetes mellitus with gangrene
C108500	Insulin dependent diabetes mellitus with ulcer	C10E612	Insulin dependent diabetes mellitus with gangrene
C108511	Type I diabetes mellitus with ulcer	C10E700	Type 1 diabetes mellitus with retinopathy
C108512	Type 1 diabetes mellitus with ulcer	C10E711	Type I diabetes mellitus with retinopathy
C108600	Insulin dependent diabetes mellitus with gangrene	C10E712	Insulin dependent diabetes mellitus with retinopathy
C108611	Type I diabetes mellitus with gangrene	C10E800	Type 1 diabetes mellitus - poor control
C108612	Type 1 diabetes mellitus with gangrene	C10E811	Type I diabetes mellitus - poor control
C108700	Insulin dependent diabetes mellitus with retinopathy	C10E812	Insulin dependent diabetes mellitus - poor control
C108711	Type I diabetes mellitus with retinopathy	C10E900	Type 1 diabetes mellitus maturity onset
C108712	Type 1 diabetes mellitus with retinopathy	C10E911	Type I diabetes mellitus maturity onset
C108800	Insulin dependent diabetes mellitus - poor control	C10E912	Insulin dependent diabetes maturity onset
C108811	Type I diabetes mellitus - poor control	C10EA00	Type 1 diabetes mellitus without complication
C108812	Type 1 diabetes mellitus - poor control	C10EA11	Type I diabetes mellitus without complication
C108900	Insulin dependent diabetes maturity onset	C10EA12	Insulin-dependent diabetes without complication
C108911	Type I diabetes mellitus maturity onset	C10EB00	Type 1 diabetes mellitus with mononeuropathy
C108912	Type 1 diabetes mellitus maturity onset	C10EB11	Type I diabetes mellitus with mononeuropathy
C108A00	Insulin-dependent diabetes without complication	C10EB12	Insulin dependent diabetes mellitus with mononeuropathy
C108A11	Type I diabetes mellitus without complication	C10EC00	Type 1 diabetes mellitus with polyneuropathy
C108A12	Type 1 diabetes mellitus without complication	C10EC11	Type I diabetes mellitus with polyneuropathy
C108B00	Insulin dependent diabetes mellitus with mononeuropathy	C10EC12	Insulin dependent diabetes mellitus with polyneuropathy
C108B11	Type I diabetes mellitus with mononeuropathy	C10ED00	Type 1 diabetes mellitus with nephropathy
C108B12	Type 1 diabetes mellitus with mononeuropathy	C10ED11	Type I diabetes mellitus with nephropathy
C108C00	Insulin dependent diabetes mellitus with polyneuropathy	C10ED12	Insulin dependent diabetes mellitus with nephropathy
C108C11	Type I diabetes mellitus with polyneuropathy	C10EE00	Type 1 diabetes mellitus with hypoglycaemic coma
C108C12	Type 1 diabetes mellitus with polyneuropathy	C10EE11	Type I diabetes mellitus with hypoglycaemic coma
C108D00	Insulin dependent diabetes mellitus with nephropathy	C10EE12	Insulin dependent diabetes mellitus with hypoglycaemic coma
C108D11	Type I diabetes mellitus with nephropathy	C10EF00	Type 1 diabetes mellitus with diabetic cataract
C108D12	Type 1 diabetes mellitus with nephropathy	C10EF11	Type I diabetes mellitus with diabetic cataract
C108E00	Insulin dependent diabetes mellitus with hypoglycaemic coma	C10EF12	Insulin dependent diabetes mellitus with diabetic cataract
C108E11	Type I diabetes mellitus with hypoglycaemic coma	C10EG00	Type 1 diabetes mellitus with peripheral angiopathy
C108E12	Type 1 diabetes mellitus with hypoglycaemic coma	C10EG11	Type I diabetes mellitus with peripheral angiopathy
C108F00	Insulin dependent diabetes mellitus with diabetic cataract	C10EG12	Insulin dependent diab mell with peripheral angiopathy

C108F11	Type I diabetes mellitus with diabetic cataract	C10EH00	Type 1 diabetes mellitus with arthropathy
C108F12	Type 1 diabetes mellitus with diabetic cataract	C10EH11	Type I diabetes mellitus with arthropathy
C108G00	Insulin dependent diab mell with peripheral angiopathy	C10EH12	Insulin dependent diabetes mellitus with arthropathy
C108G11	Type I diabetes mellitus with peripheral angiopathy	C10EJ00	Type 1 diabetes mellitus with neuropathic arthropathy
C10EJ11	Type I diabetes mellitus with neuropathic arthropathy	C10P000	Type I diabetes mellitus in remission
C10EJ12	Insulin dependent diab mell with neuropathic arthropathy	C10P011	Type 1 diabetes mellitus in remission
C10EK00	Type 1 diabetes mellitus with persistent proteinuria	C10y000	Diabetes mellitus, juvenile, + other specified manifestation
C10EK11	Type I diabetes mellitus with persistent proteinuria	C10z000	Diabetes mellitus, juvenile type, + unspecified complication
C10EL00	Type 1 diabetes mellitus with persistent microalbuminuria	C10EP00	Type 1 diabetes mellitus with exudative maculopathy
C10EL11	Type I diabetes mellitus with persistent microalbuminuria	C10EP11	Type I diabetes mellitus with exudative maculopathy
C10EM00	Type 1 diabetes mellitus with ketoacidosis	C10EQ00	Type 1 diabetes mellitus with gastroparesis
C10EM11	Type I diabetes mellitus with ketoacidosis	C10EQ11	Type I diabetes mellitus with gastroparesis
C10EN00	Type 1 diabetes mellitus with ketoacidotic coma	C10EN11	Type I diabetes mellitus with ketoacidotic coma

Appendix Table 3: Gestational Diabetes Read codes used in extracting cohort of patients with gestational diabetes

Read Code	Description
L180.00	Diabetes mellitus during pregnancy/childbirth/puerperium
L180000	Diabetes mellitus - unspec whether in pregnancy/puerperium
L180100	Diabetes mellitus during pregnancy - baby delivered
L180200	Diabetes mellitus in puerperium - baby delivered
L180300	Diabetes mellitus during pregnancy - baby not yet delivered
L180400	Diabetes mellitus in pueperium - baby previously delivered
L180800	Diabetes mellitus arising in pregnancy
L180811	Gestational diabetes mellitus
L180900	Gestational diabetes mellitus
L180z00	Diabetes mellitus in pregnancy/childbirth/puerperium NOS
ZV13F00	[V]Personal history of gestational diabetes mellitus

Appendix B

Appendix Table 4: Anti-diabetic drug (ADD) generic, brand names in THIN, and SAS codes for their extraction

generic names	brand names	SAS SQL CODE		
INSULIN				
insulin		upcase (genericname)	like	'%INSULIN%'
Insulin aspart	Novolog	upcase (genericname)	like	'%INSULIN ASPART%'
Insulin glulisine	Apidra	upcase (genericname)	like	'%INSULIN GLULISINE%'
Insulin lispro	Humalog	upcase (genericname)	like	'%INSULIN LISPRO%'
Insulin human	Afrezza Inhalation Powder	upcase (genericname)	like	'%INSULIN HUMAN%'
Regular insulin	Humulin R, Novolin R	upcase (genericname)	like	'%REGULAR INSULIN%'
Insulin NPH	Hagedorn NPH, Humulin N, Novolin N	upcase (genericname)	like	'%INSULIN NPH%'
Insulin detemir	Levemir	upcase (genericname)	like	'%INSULIN DETEMIR%'
Insulin glargine	Lantus	upcase (genericname)	like	'%INSULIN GLARGINE%'
Insulin aspart protamine	NovoLog 50/50, NovoLog 70/30	upcase (genericname)	like	'%INSULIN ASPART PROTAMINE%'
insulin aspart	NovoLog 50/50, NovoLog 70/30	upcase (genericname)	like	'%INSULIN ASPART%'
Insulin lispro protamine	Humalog 50/50, Humalog 75/25	upcase (genericname)	like	'%INSULIN LISPRO PROTAMINE%'
insulin lispro	Humalog 50/50, Humalog 75/25	upcase (genericname)	like	'%INSULIN LISPRO%'
BIGUADINES				
metformin	Glucophage, Glucophage XR, Glumetza, Riomet, Fortamet	upcase (genericname)	like	'%METFORMIN%'
Phenformin	Glucophage, Glucophage XR, Glumetza, Riomet, Fortamet	upcase (genericname)	like	'%PHENFORMIN%'
Buformin	Glucophage, Glucophage XR, Glumetza, Riomet, Fortamet	upcase (genericname)	like	'%BUFORMIN%'
SULPHONYLUREAS				
Acetohexamide	Dymelor	upcase (genericname)	like	'%ACETOHEXAMIDE%'
butanamide		upcase (genericname)	like	'%BUTANAMIDE%'
Daonil		upcase (genericname)	like	'%DAONIL%'
Chlorpropamide	Diabinese	upcase (genericname)	like	'%CHLORPROPAMIDE%'

Tolazamide	Tolinase	upcase (genericname)	like	'%TOLAZAMIDE%'
Tolbutamide	Orinase	upcase (genericname)	like	'%TOLBUTAMIDE%'
Glipizide	Glucotrol, Minidiab, Glibenese	upcase (genericname)	like	'%GLIPIZIDE%'
Glyburide	Diabeta, Micronase, Glynase, Daonil, Euglycon	upcase (genericname)	like	'%GLYBURIDE %'
glibenclamide	Diabeta, Micronase, Glynase, Daonil, Euglycon	upcase (genericname)	like	'%GLIBENCLAMIDE%'
Glimepiride	Amaryl	upcase (genericname)	like	'%GLIMEPIRIDE%'
Gliclazide	Uni Diamicron	upcase (genericname)	like	'%GLICLAZIDE%'
Glyclopyramide	Deamelin-S	upcase (genericname)	like	'%GLYCLOPYRAMIDE%'
Gliquidone	Glurenorm	upcase (genericname)	like	'%GLIQUIDONE%'
TZDs		upcase (genericname)	like	'%ROSIGLITAZONE%'
Rosiglitazone	Avandia	upcase (genericname)	like	'%PIOGLITAZONE%'
Pioglitazone	Actos	upcase (genericname)	like	'%TROGLITAZONE%'
Troglitazone	Tolinase			
ALPHA GLUCOSIDASE		upcase (genericname)	like	'%ACARBOSE%'
Acarbose	Precose, Glucobay	upcase (genericname)	like	'%MIGLITOL%'
Miglitol	Glyset	upcase (genericname)	like	'%VOGLIBOSE%'
Voglibose	Basen			
GLP1-RA		upcase (genericname)	like	'%EXENATIDE%'
Exenatide	Byetta	upcase (genericname)	like	'%LIXISENATIDE%'
Lixisenatide	Lyxumia	upcase (genericname)	like	'%LIRAGLUTIDE%'
Liraglutide	Victoza	upcase (genericname)	like	'%ALBIGLUTIDE%'
Albiglutide	Tanzeum	upcase (genericname)	like	'%DULAGLUTIDE%'
Dulaglutide	Trulicity	upcase (genericname)	like	'%EXENATIDE ONCE WEEKLY%'
Exenatide once weekly	Bydureon			
DPP4 INHIBITORS		upcase (genericname)	like	'%ALOGLIPTIN%'
Alogliptin	Nesina, Vipidia	upcase (genericname)	like	'%ANAGLIPTIN%'
Anagliptin	Suiny	upcase (genericname)	like	'%LINAGLIPTIN%'

Linagliptin	Trajenta	upcase (genericname)	like	'%SAXAGLIPTIN%'
Saxagliptin	Onglyza	upcase (genericname)	like	'%SITAGLIPTIN%'
Sitagliptin	Januvia	upcase (genericname)	like	'%TENELIGLIPTIN%'
Teneligliptin	Tenelia			
AMYLIN ANALOGUES		upcase (genericname)	like	'%PRAMLINTIDE%'
Pramlintide	Symlin			
SGLT2		upcase (genericname)	like	'%CANAGLIFLOZIN%'
Canagliflozin	Invokana	upcase (genericname)	like	'%DAPAGLIFLOZIN%'
Dapagliflozin	Forxiga, Farxiga	upcase (genericname)	like	'%EMPAGLIFLOZIN%'
Empagliflozin	Jardiance			
METGLINIDE				
Nateglinide	Starlix	upcase (genericname)	like	'%NATEGLINIDE%'
Repaglinide	Prandin, NovoNorm	upcase (genericname)	like	'%REPAGLINIDE%'
OTHER ADDS				
Bromocriptine	Parlodel, Cycloset	upcase (genericname)	like	'%BROMOCRIPTINE%'
Colesevelam	Welchol, Cholestagel, Lodalis	upcase (genericname)	like	'%COLESEVELAM%'

Appendix C

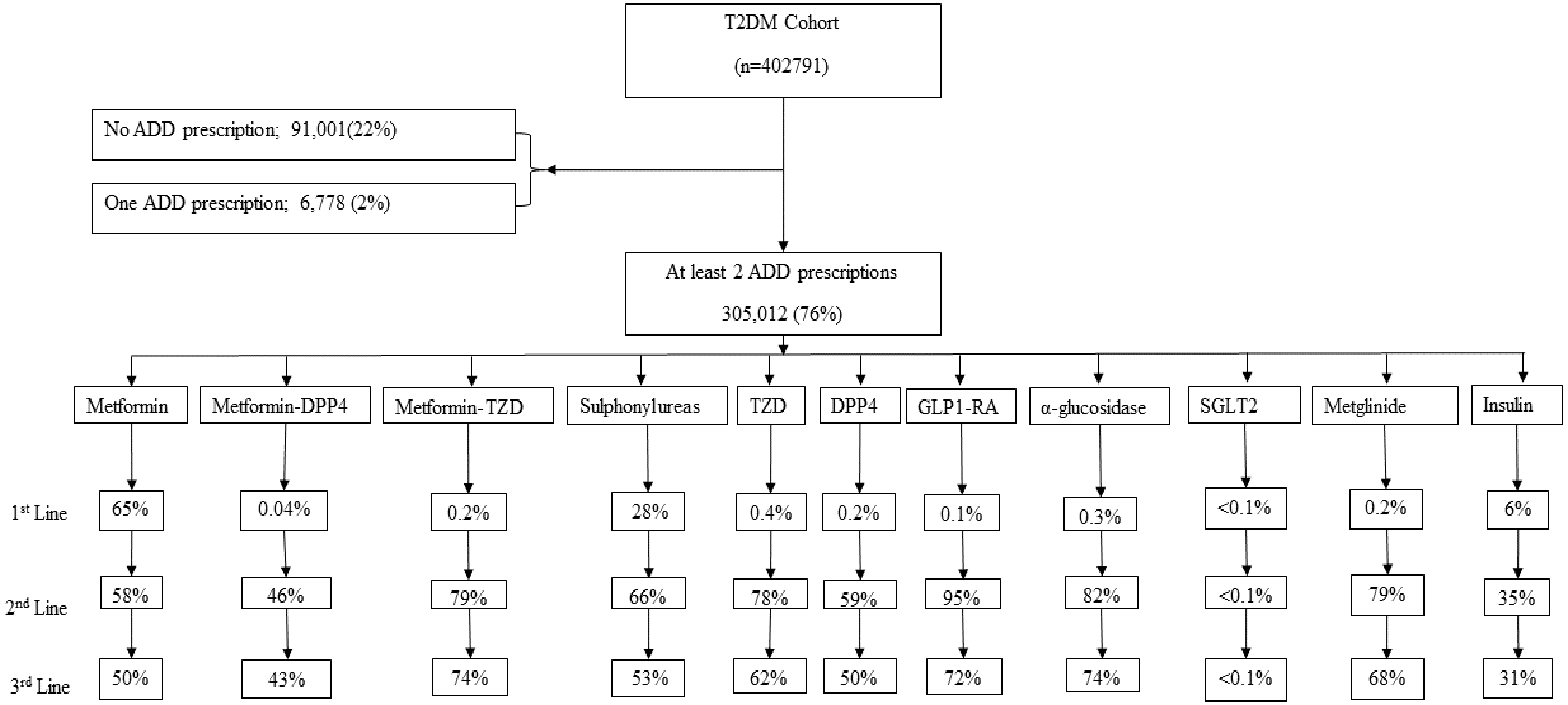
Appendix Table 5: Number (N) and proportion (%) of patients with T2DM that have ever been prescribed an anti-hyperglycaemic drug.

	All	1 prescription	≥2 prescriptions	time to first ADD	
	N (%)	N (%)	N (%)	Mean (SD)	median (Q1,Q3)
Insulin	74,626(19)	347(<0.1)	74,279(18)	8.6 (7.3)	7.3 (3.0,12.3)
Metformin only	260,705(65)	4,251 (1)	256,454(64)	3.3 (5.0)	1.1 (0.0,4.8)
Metformin - TZD Combination	10,584(3)	14(<0.1)	10,570(3)	6.3 (5.4)	5.0 (2.4,8.8)
Metformin - DPP4 Combination	2,934(1)	11(<0.1)	2,923(1)	7.6 (5.7)	6.8 (3.3,10.6)
Metformin - SGLT2 Combination		-	-	-	-
Sulphonylureas	184,146(46)	2,070 (1)	182,076(45)	4.1 (5.2)	2.3 (0.2,6.2)
TZD Only	47,019(12)	19(<0.1)	47,000(12)	6.8 (5.6)	5.6 (2.7,9.5)
DPP4 Only	34,463(9)	27(<0.1)	34,436(9)	8.4 (6.1)	7.5 (3.9,11.5)
GLP1RA	11,472(3)	3(<0.1)	11,469(3)	9.1(5.6)	8.3(5.0,12.1)
Alpha Glucosidase	8,789(2)	26(<0.1)	8,763(2)	7.3 (6.3)	6.0 (2.6,10.4)
SGLT2 Only	2,072(1)	-	2,072(1)	9.7 (5.8)	9.0 (5.4, 13.2)
Metglinide	4,493(1)	10(<0.1)	4,483(1)	6.4 (6.0)	5.0 (1.9,9.3)

Appendix Table 6: Prescription patterns of anti-hyperglycaemic drugs among patients with type 2 diabetes

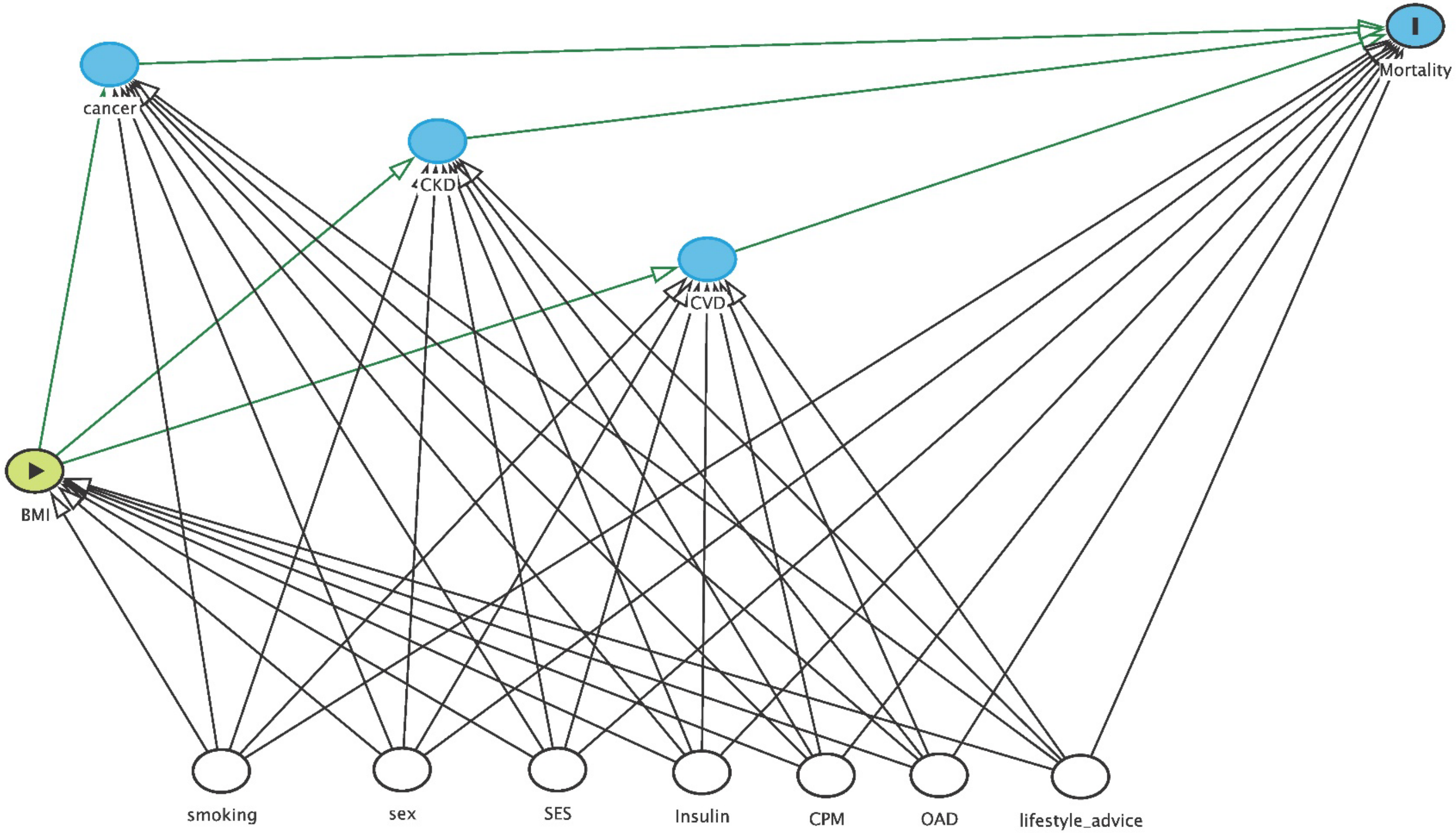
	Metformin	**Sulphonylureas**	**TZD**	**Insulin**
1st line therapy	198,504 (65)	84,518 (28)	1,276 (0.4)	17,909 (6)
remained on 1st line ADD	83,152 (42)	28,369 (34)	277 (22)	11684 (65)
2nd line therapy	115,252 (58)	56,149 (66)	999 (78)	6225(35)
Added on 2nd ADD	104,460 (98)	54,135 (96)	914(91)	5501(88)
Switched to 2nd ADD	2,844 (2)	2014 (4)	85 (9)	724 (12)
2nd line therapy name				
Metformin only	-	45,672 (81)	542 (54)	4862 (78)
Metformin-DPP4 combination	724 (0.5)	78 (0.1)	4 (0.4)	7 (0.1)
Metformin-TZD combination	3,346 (3)	327 (0.6)	25 (2.5)	8 (0.1)
Sulphonylureas	81,709(71)	-	301 (30)	1014 (16)
TZD only	12,853 (11)	2,898(5)	-	79(1)
DPP4 only	8,073 (7)	570(1)	30 (3)	87 (1)
GLP1RA	706 (0.6)	27 (<0.1)	6 (0.6)	69 (1)
α-glucosidase	800 (0.7)	1,441 (3)	2 (0.2)	70 (1)
SGLT2 only	158 (0.1)	1 (<0.1)	-	2(<0.1)
Metglinide	1,118 (1)	258 (0.5)	23 (2)	27(0.4)
Insulin	5,865 (5)	4,877(9)	66 (7)	-
min 1yr on 2nd line therapy	53,118 (46)	37,151 (66)	322 (32)	3032 (49)
Time to 2nd line from 1st line, months	9.4 (0.5,31.5)	23.7 (6.6,50.2)	2.6 (0.7,19.6)	11 (1.6,46.4)
remained on 2nd line ADD	56,239 (50)	26,277 (47)	380 (38)	4303 (69)
3rd line therapy	56,269 (50)	29,872 (53)	619 (62)	1922(31)
Added on 3rd ADD	50,592 (90)	26,442 (89)	589 (95)	1,744 (91)
Switched to 3rd ADD	5,677 (10)	3,430 (11)	30 (5)	178 (9)
3rd line therapy name				
Metformin only	-	3675 (12)	214 (35)	597 (31)
Metformin-DPP4 combination	772 (1)	141 (0.5)	8 (1)	8 (0.4)
Metformin-TZD combination	2,619 (5)	1202 (4)	32 (5)	32 (2)
Sulphonylureas	9,646 (17)	-	181 (29)	596 (31)
TZD only	14,439 (26)	9340 (31)	-	216 (11)

DPP4 only	11,286 (20)	2647 (9)	84 (14)	190 (10)
GLP1RA	2,156 (4)	176 (0.6)	21 (3)	202 (11)
α-glucosidase	2,232 (4)	2414 (8)	7 (1)	50 (3)
SGLT2	328(0.6)	20 (0.1)	-	10 (0.5)
Metglinide	792 (1)	670 (2)	4 (0.7)	21 (1)
Insulin	1,248 (21)	9587 (32)	68 (11)	-



Appendix Figure 1: Flow chart illustrating the proportion of patients who progressed to 2nd and 3rd line therapy from 1st line therapy

Appendix D



Appendix Figure 2: Direct Acyclic Graph (DAG) showing the relationship between BMI (exposure), potential confounders, and mortality (outcome).

The model below is a mathematical representation of the variables depicted in Appendix Figure 1. The exposure is BMI and the outcome is all-cause mortality. Potential confounders (white circle) include smoking status, sex, socio-economic status (SES), use of insulin, use of oral antidiabetic medication, use of cardio-protective medications, and receipt of lifestyle advice post-diagnosis. Clinically diagnosed cancer, CKD, and any cardiovascular disease are in the pathway to mortality. This framework was used in the statistical modelling during the mortality risk assessment. There is no adjustment for CVD, cancer, and CKD in equation 1, whereas, in equation 2, there is an adjustment for these variables. We show that adjusting for cancer, CKD and cancer does not introduce bias into our estimates.

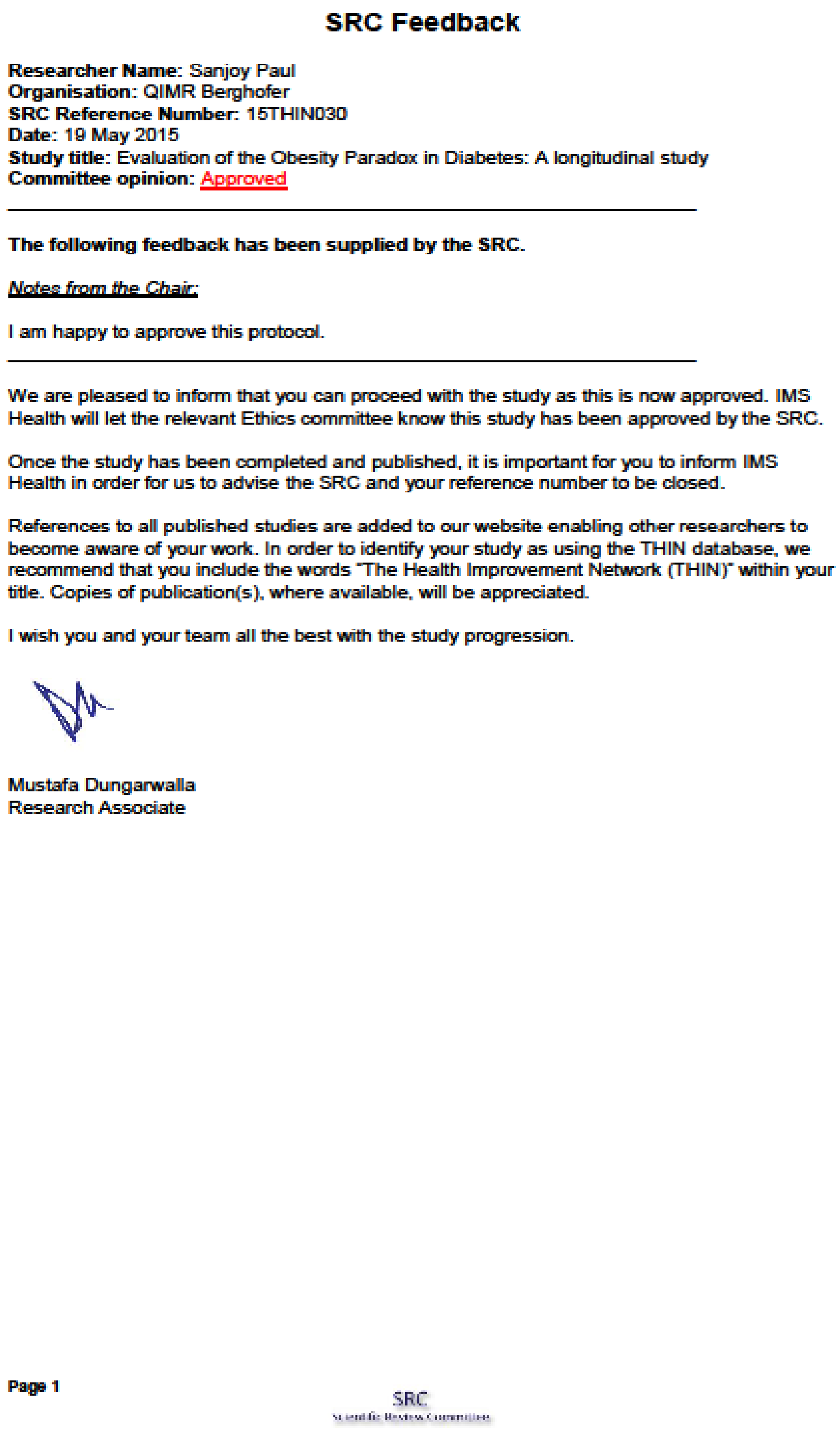
Model 1: $\Pr(dead = 1) = \beta_1(BMI) + \beta_2(Smoke) + \beta_2(SES) + \beta_3(Sex) + \beta_4(Insulin) + \beta_5(OAD) + \beta_6(CPM) + \beta_7(Lifestyle\ advice) + \beta_8(HbA1c) + \beta_9(LDL) + \beta_{10}(HDL)$ **Equation 1**

Model 2: $\Pr(dead = 1) = All\ Covariates\ of\ Model\ 1 + \left(\beta_{12}(CVD) + \beta_{13}(CKD) + \beta_{14}(Cancer)\right)$ **Equation 2**

Appendix Table 7: Mortality risk by BMI category at the time of diabetes diagnosis using two models 1 and 2.

	All		Lost body weight		No weight loss	
	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
Normal weight	1.29 (1.13,1.48)	1.31 (1.13,1.51)	0.89 (0.62,1.26)	0.92 (0.64,1.32)	1.43 (1.23,1.67)	1.44 (1.22,1.70)
Overweight	1.05 (0.96,1.14)	1.05 (0.96,1.15)	0.98 (0.79,1.21)	0.99 (0.80,1.24)	1.05 (0.96,1.15)	1.05 (0.95,1.16)
Grade 1 Obese	Reference	Reference	Reference	Reference	Reference	Reference
Grade 2 Obese	1.00 (0.90,1.11)	1.04 (0.93,1.15)	1.05 (0.80,1.36)	1.04 (0.79,1.37)	1.00 (0.89,1.11)	1.04 (0.93,1.17)
Grade 3 Obese	1.06 (0.90,1.25)	1.11 (0.93,1.33)	1.39 (0.89,2.17)	1.27 (0.79,2.04)	1.00 (0.84,1.21)	1.10 (0.90,1.33)

Appendix E



SRC Feedback

Researcher Name: Sanjoy Paul
Organisation: QIMR Berghofer
SRC Reference Number: 15THIN030
Date: 19 May 2015
Study title: Evaluation of the Obesity Paradox in Diabetes: A longitudinal study
Committee opinion: Approved

The following feedback has been supplied by the SRC.

Notes from the Chair:

I am happy to approve this protocol.

We are pleased to inform that you can proceed with the study as this is now approved. IMS Health will let the relevant Ethics committee know this study has been approved by the SRC.

Once the study has been completed and published, it is important for you to inform IMS Health in order for us to advise the SRC and your reference number to be closed.

References to all published studies are added to our website enabling other researchers to become aware of your work. In order to identify your study as using the THIN database, we recommend that you include the words "The Health Improvement Network (THIN)" within your title. Copies of publication(s), where available, will be appreciated.

I wish you and your team all the best with the study progression.

Mustafa Dungarwalla
Research Associate

Page 1

SRC
Scientific Review Committee