



Identifying subtypes of type 2 diabetes mellitus with machine learning: development, internal validation, prognostic validation and medication burden in linked electronic health records in 420 448 individuals

Mehrdad A Mizani ^{1,2}, Ashkan Dashtban,¹ Laura Pasea,¹ Qingjia Zeng,^{1,3} Kamlesh Khunti,⁴ Jonathan Valabhji,^{5,6} Jil Billy Mamza,⁷ He Gao ⁸, Tamsin Morris,⁸ Amitava Banerjee ^{1,9}

To cite: Mizani MA, Dashtban A, Pasea L, *et al*. Identifying subtypes of type 2 diabetes mellitus with machine learning: development, internal validation, prognostic validation and medication burden in linked electronic health records in 420 448 individuals. *BMJ Open Diab Res Care* 2024;**12**:e004191. doi:10.1136/bmjdr-2024-004191

► Additional supplemental material is published online only. To view, please visit the journal online (<https://doi.org/10.1136/bmjdr-2024-004191>).

Received 20 March 2024
Accepted 22 May 2024



© Author(s) (or their employer(s)) 2024. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to
Professor Amitava Banerjee;
ami.banerjee@ucl.ac.uk

ABSTRACT

Introduction None of the studies of type 2 diabetes (T2D) subtyping to date have used linked population-level data for incident and prevalent T2D, incorporating a diverse set of variables, explainable methods for cluster characterization, or adhered to an established framework. We aimed to develop and validate machine learning (ML)-informed subtypes for type 2 diabetes mellitus (T2D) using nationally representative data.

Research design and methods In population-based electronic health records (2006–2020; Clinical Practice Research Datalink) in individuals ≥18 years with incident T2D (n=420 448), we included factors (n=3787), including demography, history, examination, biomarkers and medications. Using a published framework, we identified subtypes through nine unsupervised ML methods (K-means, K-means++, K-mode, K-prototype, mini-batch, agglomerative hierarchical clustering, Birch, Gaussian mixture models, and consensus clustering). We characterized clusters using intracluster distributions and explainable artificial intelligence (AI) techniques. We evaluated subtypes for (1) internal validity (within dataset; across methods); (2) prognostic validity (prediction for 5-year all-cause mortality, hospitalization and new chronic diseases); and (3) medication burden.

Results *Development:* We identified four T2D subtypes: metabolic, early onset, late onset and cardiometabolic. *Internal validity:* Subtypes were predicted with high accuracy (F1 score >0.98). *Prognostic validity:* 5-year all-cause mortality, hospitalization, new chronic disease incidence and medication burden differed across T2D subtypes. Compared with the metabolic subtype, 5-year risks of mortality and hospitalization in incident T2D were highest in late-onset subtype (HR 1.95, 1.85–2.05 and 1.66, 1.58–1.75) and lowest in early-onset subtype (1.18, 1.11–1.27 and 0.85, 0.80–0.90). Incidence of chronic diseases was highest in late-onset subtype and lowest in early-onset subtype. *Medications:* Compared with the metabolic subtype, after adjusting for age, sex, and pre-T2D medications, late-onset subtype (1.31, 1.28–1.35) and

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Studies of type 2 diabetes (T2D) subtyping have often used non-representative datasets, employed limited variables and comorbidities, lacked medication data as clustering features, and focused on a single time point.

WHAT THIS STUDY ADDS

⇒ At both T2D onset (incident T2D) and study exit (prevalent T2D), we identified four distinct clusters: (1) metabolic, (2) early onset, (3) late onset, and (4) cardiometabolic, which differed by baseline characteristics, medication usage, hospitalization and mortality.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Based on a replicable framework, explainable machine learning is an instrumental approach for subtyping diseases in electronic health record data, and these T2D subtypes, after external validation, have potential applications to standardize subtypes in research, inform clinical guidelines, improve T2D management, and optimize healthcare utilization.

early-onset subtype (0.83, 0.81–0.85) were most and least likely, respectively, to be prescribed medications within 5 years following T2D onset.

Conclusions In the largest study using ML to date in incident T2D, we identified four distinct subtypes, with potential future implications for etiology, therapeutics, and risk prediction.

INTRODUCTION

More than 537 million adults are affected by diabetes mellitus worldwide, which is predicted to rise to 643 million by 2030.¹

In the UK, approximately 7% of the population, corresponding to over 5 million individuals, are either diagnosed with or at risk of developing diabetes mellitus, with type 2 diabetes (T2D) accounting for over 90% of diagnosed cases.^{2,3} The mortality rate of individuals with T2D is almost twice as much as people without diabetes, making it the ninth leading cause of global mortality with over 1 million attributable deaths per year.^{4,5}

T2D is a multifactorial disease with complex and interconnected risk factors, comorbidities, and complications such as obesity, hypertension (HT), cardiovascular disease (CVD), chronic kidney disease (CKD), heart failure (HF), cancers, cognitive decline, mental health sequelae and age-related disability.⁶⁻⁸ Furthermore, prevalence of T2D and its outcomes and complications vary widely by risk factors such as genetic disposition, age, sex, ethnicity, sociodemographic status, lifestyle, family history, and medication.⁹⁻¹¹ Many T2D risk factors and complications are preventable and manageable with changes in pharmaceutical and lifestyle interventions (eg, diet, exercise, smoking cessation), education, or medication adherence.^{3-6,12,13} The complex phenotypic characteristics of T2D signal the need for an individualized approach for risk stratification,¹⁴ as emphasized in National Institute for Health and Care Excellence guidelines.¹⁵

Risk stratification to date has mostly relied on datasets which are unrepresentative of whole populations, limited availability of variables,^{16,17} focused on particular risk factors or biomarkers,¹⁸ or complications.¹⁹ Population-level, linked, longitudinal electronic health records (EHR) could facilitate more comprehensive investigation of T2D subtypes, as well as more advanced data analytics. We have developed a framework for use of machine learning (ML) in subtyping of long-term conditions, which has been used to develop and validate subtypes in HF²⁰ with external and genetic validation, and CKD,²¹ which may be useful to distinguish T2D subtypes.

In a linked, longitudinal, national EHR in England, in 420 448 individuals with T2D and 3787 factors, we used multiple ML methods following our published framework to (1) generate subtypes with clinical relevance throughout the course of T2D, and low risk of bias for individual selection and algorithms (*development*); (2) demonstrate internal (across methods) and prognostic validity (predictive accuracy for 5-year all-cause mortality and hospitalization) (*validation*); and (3) investigate distribution of medication classes at baseline and over time (*medications*).

METHODS

In this study, we used our framework for ML-informed subtype implementation, and internal and prognostic validation²⁰ to guide our methods. We have used the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis-Cluster checklist to report our methods.²¹

To generate subtypes (development)

Clinical relevance

Our methodology was focused on diagnostic and prognostic accuracy and risk stratification related to T2D. The primary objective was to improve patient outcomes and clinical benefits through better subtyping of T2D. We used individual-level, longitudinal data from Clinical Practice Research Datalink (CPRD) Gold, which links EHR, including primary care, Hospital Episode Statistics, and death registration from the Office for National Statistics (ONS). CPRD is representative of the UK population by sociodemographic variables.²¹

Patients

We included individuals aged ≥ 18 years with a diagnosis of T2D. The study entry date was the latest of the individual's 18th birthday, January 1, 1997, or the 'up-to-standard CPRD Gold date' when the primary care practice was deemed to qualify for research. T2D diagnosis was by Read code v2 and International Classification of Diseases Tenth Revision clinical vocabularies and validated CALIBER phenotype algorithms in the Health Data Research (HDR) UK phenotype library. Any patients with clinical codes of type 1 diabetes only were excluded. The study exit time was defined as the earliest of death date, April 30, 2020, or termination of CPRD recording.

Algorithm

We assessed combinations of variables, feature engineering methods, and clustering algorithms to identify the optimal clinically viable solution. A total of 3787 variables (online supplemental table S1), including socio-demographic factors ($n=3$), chronic diseases ($n=25$), medication ($n=3750$), behavioral factors ($n=2$), and biomarkers ($n=7$), were used to create the final set of 25 unitary and composite features for the analysis. Variables with missing categories, including ethnicity, Index of Multiple Deprivation (IMD), and behavioral factors, were used for cluster characterization only. All variables were determined at the time of T2D diagnosis for the incident and at the study exit for prevalent T2D. Biomarkers were assessed based on their mean values during the final year prior to T2D onset and at study exit.

For smoking and alcohol consumption, which may change over time, the final status before the index date was considered. Medications were grouped into five major categories to reduce feature dimensionality. We assessed factor analysis of multiple data and multiple correspondence analysis (MCA) to transform mixed and categorical variables, respectively, into numerical form, accounting for potential correlations and collinearity. We used MCA to transform variables to preserve $\geq 95\%$ of the variance of all features transformed into discrete variables. For biomarker variables, the categorization was based on reference ranges. Missing values were imputed with a default 'unknown' category. For clustering, we only considered variables without missing or unknown categories (ie, complete feature analysis), while we used

all the variables as shown in online supplemental table S1 for cluster characterization. To determine the optimal number of clusters, we used the Elbow, Davies-Bouldin (DB), and Calinski-Harabasz (CH) methods with the clinical expert review.

We compared the clustering results using nine algorithms (K-means with random and K-means++ initialization, K-mode, K-prototype, mini-batch, agglomerative hierarchical, Birch, Gaussian mixture models, and consensus clustering) on data subsamples to assess computational feasibility. Performance was measured using silhouette, DB, and CH cluster validity indices. For interpretation and characterization of clusters and assessing feature impact, we applied Light Gradient-Boosting Machine (LightGBM)²² classification based on cluster labels, followed by using SHapley Additive exPlanations (SHAP) values for explainable ML outputs.²³ SHAP values offer an improved alternative to pairwise correlation analysis by considering the relative importance of a variable in cluster membership, accounting for complex and non-linear interactions among all variables. Furthermore, SHAP values provide the direction of the feature importance, highlighting both the effect of high frequency as well as absence or rarity of a variable to cluster membership.²⁴ The LightGBM model was trained for imbalanced multiclass labels using multilog loss metric, gradient-boosted decision tree, and 10-fold cross-validation to avoid overfitting. Additionally, we assessed the prevalence of each variable per cluster, as well as the proportion of variables across all clusters, for detailed cluster characterization.

To demonstrate validity (validation) and medication profile (medications)

Internal and prognostic validity

We cross-validated cluster characteristics by SHAP values to validate the optimal cluster count. To evaluate the predictive validity of final cluster labels, we employed 10-fold cross-validation using three supervised classification algorithms, including LightGBM, support vector machine (SVM), and K-nearest neighbors (KNN). These classification models were assessed based on macro-level accuracy and F1 score. Our use of 'hard' rather than 'fuzzy' clustering methods meant that individuals were members of discrete clusters.

For prognostic validation, Kaplan-Meier survival analysis and Cox regression modeling were used to evaluate 5-year all-cause mortality, hospital admission, and medication prescription after diagnosis of T2D. The Cox regression model was adjusted for confounding factors of age and sex. For Cox regression of post-T2D medication prescription, we additionally adjusted the model for medications prescribed before the T2D onset. Furthermore, we assessed the statistical significance of the transition in cluster membership trajectories from T2D diagnosis (incident T2D) to study exit (prevalent T2D) using McNemar's test.

Ethical approval

This study is based in part on data from the CPRD obtained under license from the UK Medicines and Healthcare products Regulatory Agency. The data are provided by patients and collected by the National Health Service (NHS) as part of their care and support. The ONS is also acknowledged as the provider of the ONS Data contained within the CPRD Data. The interpretation and conclusions contained in this study are those of the author/s alone. Copyright (2020), reused with the permission of NHS Digital. All rights reserved. Anonymized data were used for analyses; therefore, informed consent was not necessary.

Funding

AstraZeneca UK, HDR UK.

RESULTS

Clinical relevance and patients

We included 420 448 individuals with incident T2D (mean age at T2D diagnosis 59.98 years (SD 13.86); male 56.21%) (online supplemental figures S1 and S2). By the end of the study, the mean age of individuals with prevalent T2D was 69.47 (SD 14.16), with 132 425 (30.59%) of participants exiting the study due to all-cause mortality (online supplemental figure S2 and online supplemental table S3).

Algorithm

Variables (online supplemental table S1) included comorbidities, family history of diabetes and CVD, and medication categories (including glucose-lowering therapies, diuretic, cardiac other, lipid lowering, renin angiotensin aldosterone inhibitors), comprising 3750 individual medications. Variables with unknown or missing categories, including smoking, alcohol consumption, ethnicity, and IMD, were excluded from clustering. Biomarkers were based on the mean of all available test results 1 year prior to T2D diagnosis and at study exit. Biomarkers were also excluded from clustering. However, all variables were used for characterization of final clusters. After performing MCA on categorized variables, 25 MCA features were retained, effectively preserving 97% and 95% of the variation in original datasets at the time of T2D onset and study exit, respectively. We selected the K-means clustering algorithm with K-means++ centroid initialization on the output of MCA as the final clustering algorithm based on its performance, the trade-off between cardinality and magnitude, and computational complexity as tested on subsampled data. We evaluated cluster numbers for K in the range of 2–9 based on Elbow, DB, and CH indices. Optimal cluster numbers were 3–4 at the time of T2D diagnosis and 4 at the study exit (online supplemental figure S3).

Internal validity

There were four clusters identified at T2D onset: metabolic, early onset, late onset, and cardiometabolic, and

four clusters for prevalent T2D at study exit labeled as Pr-metabolic, and Pr-cardiometabolic based on SHAP values (figure 1). The metabolic subtype was the most common (47.05%) with mean age of 62.54 (SD 7.93) years, representing generic T2D with no distinct pattern of comorbidity or medication usage. Those in the early-onset subtype were younger (mean age 42.23, SD 8.87), with more frequent family history of diabetes, high prevalence of depression, high prevalence and proportion of smoking and heavy alcohol consumption. Individuals in the late-onset cluster were older (mean age 82.86, SD 6.47), more likely to be female, with CKD and other multiple comorbidities, more likely to be a non-smoker, non-drinker and of white ethnic background. The cardiometabolic subtype had mean age of 62.61 (SD 9.68), with family history of CVD, HT, dyslipidemia, and high medication burden in all categories (figure 1, online supplemental table S2, figure 2, online supplemental figure S4). At study end, the Pr-metabolic cluster had mean age of 66.94 (SD 8.69) with no specific pattern of comorbidities or medication usage. The Pr-early onset, with mean age of 44.68 (SD 8.39), was characterized by high prevalence of depression, smoking and alcohol consumption. The Pr-late-onset subtype was oldest (mean age 86.53, SD 5.97), more likely to be female, and higher prevalence of cardiac medications and comorbidities, particularly dementia, macular degeneration, nephrotic syndrome, and neuropathy, was common. The Pr-cardiometabolic subtype (mean age 71.53, SD 10.05) was characterized by HT, dyslipidemia, and glucose disorder, elevated body mass index (BMI), highest medication burden, and higher mortality (online supplemental figures S5–S7, online supplemental table S3). Cluster labels of subtypes at T2D onset and study exit showed predictive validity using LightGBM, SVM, and KNN classification algorithms with 10-fold cross-validation (F1 scores all >0.98; online supplemental table S4).

Prognostic validity

All subtypes at T2D onset predominantly progressed to the equivalent prevalent subtype at study end. Those with cardiometabolic and late onset had the highest proportions remaining in the same category (87.24% and 82.01%, respectively) (figure 3, online supplemental table S5). Early onset commonly progressed to metabolic (43.98%), followed by Pr-cardiometabolic (6.40%) and Pr-late onset (3.11%). T2D generic commonly progressed to Pr-late onset (22.19%) and Pr-cardiometabolic (10.43%). Late onset mainly progressed to Pr-cardiometabolic (15.80%), and rarely to metabolic and early onset (2.18%). Those in the cardiometabolic subtype progressed mainly to the Pr-late-onset subtype (6.29%).

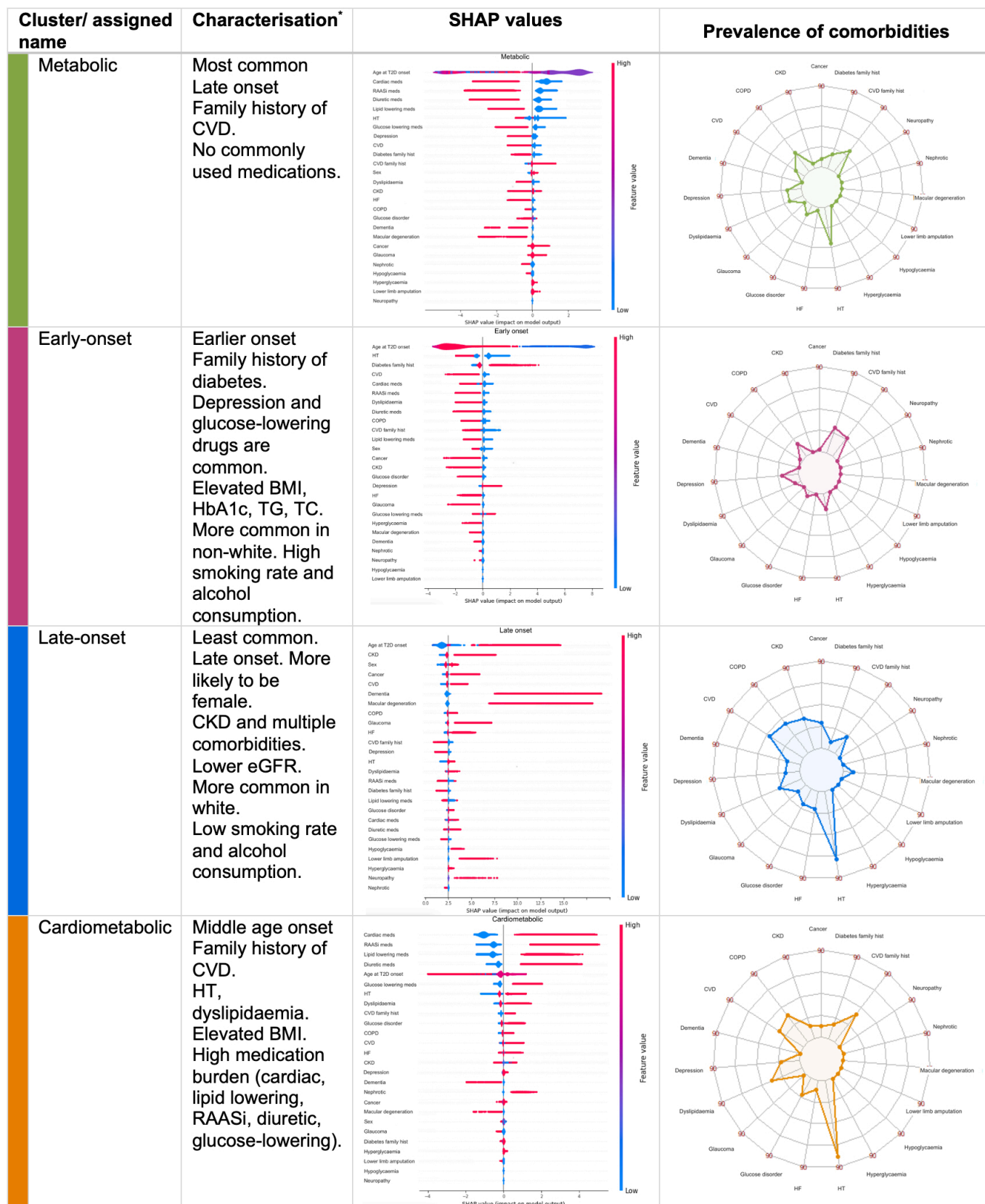
The 1-year and 5-year all-cause mortality rates were 1.05% and 3.89%, 0.81% and 2.87%, 1.06% and 4.41%, and 0.83% and 2.83% in metabolic, early-onset, cardiometabolic and late-onset subtypes (figure 4). The 1-year and 5-year all-cause hospitalization rates were 3.04% and 7.75%, 1.03% and 2.76%, 1.35% and

3.00%, and 0.96% and 1.85% in metabolic, early-onset, cardiometabolic and late-onset subtypes (figure 4). Compared with the metabolic subtype (online supplemental figure S8), age and sex-adjusted 5-year mortality risks in descending order were late onset (HR 1.95, 1.85–2.05, $p<0.0001$), cardiometabolic (1.50, 1.46–1.53, $p<0.0001$) and early onset (1.18, 1.11–1.27, $p<0.0001$). The 18% higher mortality risk in the early-onset subtype compared with the metabolic subtype underscores the importance of further investigation into the severity of outcomes in the early-onset cluster. Age and sex-adjusted 5-year hospitalization risks in descending order were late onset (1.66, 1.58–1.75, $p<0.0001$), cardiometabolic (1.37, 1.34–1.40, $p<0.0001$) and early onset (0.85, 0.80–0.90, $p<0.0001$), respectively. After age and sex adjustment, individuals with early onset were less likely to develop chronic diseases after T2D onset than those with metabolic. Late-onset and cardiometabolic subtypes had a higher risk of new chronic diseases compared with metabolic, particularly in the late-onset subtype for lower limb amputation (2.51, 1.51–4.17, $p<0.0001$), neuropathy (2.06, 1.63–2.59, $p<0.0001$), and nephrotic syndrome (HR 2.05, 1.76–2.40, $p<0.0001$); and in cardiometabolic for lower limb amputation (1.81, 1.51–2.16, $p<0.0001$), nephrotic syndrome (1.75, 1.67–1.82, $p<0.0001$), and CKD (1.59, 1.56–1.62, $p<0.0001$), dementia (1.53, 1.48–1.59, $p<0.0001$), macular degeneration (1.49, 1.42–1.56, $p<0.0001$), and chronic obstructive pulmonary disease (1.48, 1.45–1.51, $p<0.0001$) (online supplemental table S6).

Medications

Individuals in cardiometabolic and late-onset subtypes were respectively 2.60 (HR 2.60, 2.58–2.63, $p<0.0005$) and 1.69 times (HR 1.69, 1.65–1.74, $p<0.0001$) more likely to receive medication within 5 years after T2D onset compared with the metabolic subtype. Compared with metabolic, age and sex-adjusted medication prescription rate in early onset was 19% less. Adjusting for pre-T2D medication prescription, risk of new prescription was reduced in late onset (1.31, 1.28–1.35, $p<0.001$) and cardiometabolic (1.03, 1.01–1.05, $p<0.0005$), suggesting associations between medication prescription before and after T2D.

Individuals in the early-onset subtype were overall less likely to be prescribed with medication after T2D onset compared with those in the metabolic subtype, particularly for mineralocorticoid receptor antagonist (MRA) (HR 0.50, 0.45–0.54, $p<0.0001$), warfarin (HR 0.56, 0.50–0.62, $p<0.0001$), potassium-sparing diuretic (0.57, 0.39–0.84, $p<0.0001$), and thiazide diuretic (0.58, 0.56–0.61, $p<0.0001$). The likelihood of medication prescription was higher in late onset and cardiometabolic compared with metabolic. Increases in likelihood of medication prescription compared with metabolic in late onset were MRA (HR 2.06, 1.87–2.27, $p<0.0001$), warfarin (HR 2.03, 1.89–2.18, $p<0.0001$), GLP1 (HR 1.89, 1.67–2.13, $p<0.0001$),



*Based on SHAP values and prevalence and proportion of variables across clusters. These results show the overall characteristics based on variable importance in clustering.

Figure 1 Cluster-specific characteristics at type 2 diabetes (T2D) onset. BMI, body mass index; CKD, chronic kidney disease; COPD, chronic obstructive pulmonary disease; CVD, cardiovascular disease; eGFR, estimated glomerular filtration rate; HF, heart failure; HT, hypertension; RAASi, renin angiotensin aldosterone inhibitor; SHAP, SHapley Additive exPlanations; TC, total cholesterol; TG, triglyceride.

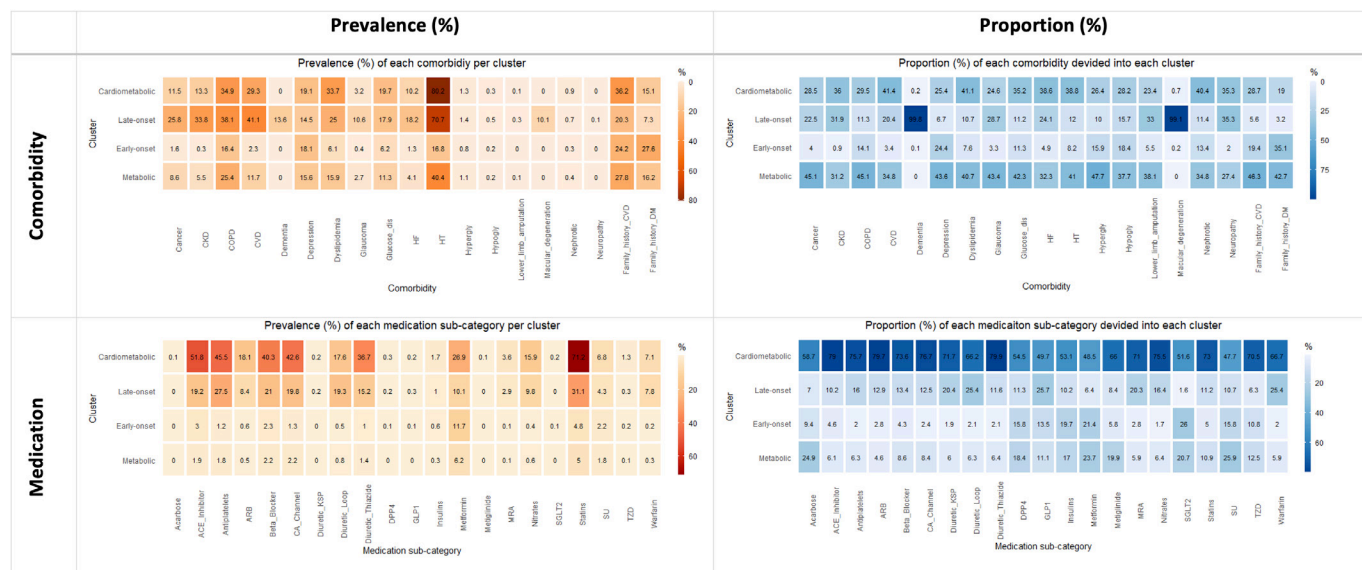


Figure 2 Prevalence and proportion of comorbidities and medication subcategories in the four clusters at type 2 diabetes (T2D) onset. CKD, chronic kidney disease; COPD, chronic obstructive pulmonary disease; CVD, cardiovascular disease; HF, heart failure; HT, hypertension.

and insulins (HR 1.81, 1.65–1.99, $p < 0.0001$); and in cardiometabolic were thiazide diuretics (1.49, 1.44–1.53, $p < 0.0001$), GLP1 (HR 1.45, 1.41–1.50, $p < 0.0001$), and insulins (1.43, 1.40–1.47, $p < 0.0001$) (online supplemental table S7).

DISCUSSION

In the largest study of ML-informed subtypes in T2D to date, we had three major findings. First, we identified four clinically distinguishable subtypes across incident and prevalent T2D: metabolic, early onset, late onset and cardiometabolic, with thorough internal validation. Second, there were distinct trajectories of these T2D subtypes, whether by subtype at the end of the study period, all-cause hospitalization or mortality. Third, we confirmed major differences in new and existing medication usage across novel T2D subtypes.

A recent systematic review included 62 studies of ‘complex’ or ML approaches to T2D subclassification in a total of 793 291 participants.²⁵ However, efforts to understand subtypes in T2D have neither used nationally representative data, nor used generalizable, reproducible methods, nor been validated, nor been associated with a wide range of outcomes.²⁵ Using nationally representative EHR data in 420 448 individuals, we have used multiple, explainable methods in incident T2D with a larger number of variables and a longer follow-up than prior studies. The same systematic review speculated ‘whether subclassification approaches at diagnosis alone are enough’ and therefore, our longitudinal follow-up of individuals and subtype classification at the end of the study period (‘prevalent’) is an advance in methodology. However, prior to clinical implementation, external validation of the subtypes observed in this study in other datasets,²⁰ as well as genetic validation,²⁶ is necessary for

cluster replication, and to better understand the overlaps and differences, compared with other proposed T2D subtypes. As Misra and colleagues note, clinical utility of T2D subtypes depends on the ability to use easily accessible data. Therefore, our use of routine EHR and simple variables is likely to increase generalizability and applicability of our T2D subtypes.

In 114 231 individuals in the Swedish National Diabetes Register, a recent study derived five subtypes in T2D using K-means clustering: older onset, severe hyperglycemia, severe obesity, younger onset, and insulin use.²⁷ Another study proposed five clusters based on only six clinical variables, including BMI and HbA1c, in All New Diabetics In Scania, a purposively sampled diabetes dataset.^{16 28} Our subtypes are plausible and consistent with these subtypes and other studies but need to be validated externally prior to clinical implementation or evaluation. The identified subtypes are likely to be more clinically generalizable than those identified in other studies using smaller samples of research cohorts or registry-based populations, which may not represent clinical practice or the general population. Interestingly, our T2D subtypes are also similar to the subtypes which we identified in HF (early onset, late onset, atrial fibrillation related, metabolic, cardiometabolic) and CKD (early onset, late onset, cancer related, metabolic, cardiometabolic).^{20 21} Increasingly, links between T2D, CVD (particularly HF) and CKD are recognized from epidemiology to clinical practice, and there are also overlaps with other diseases such as non-alcoholic fatty liver disease and obesity. Approaches to precision medicine and subtyping across diseases have been based on incident diseases ‘one-at-a-time’ and depending on which disease occurs first.^{20 21 25} However, as lifetime risk and multiple long-term conditions are increasingly investigated, the similarity of subtypes across

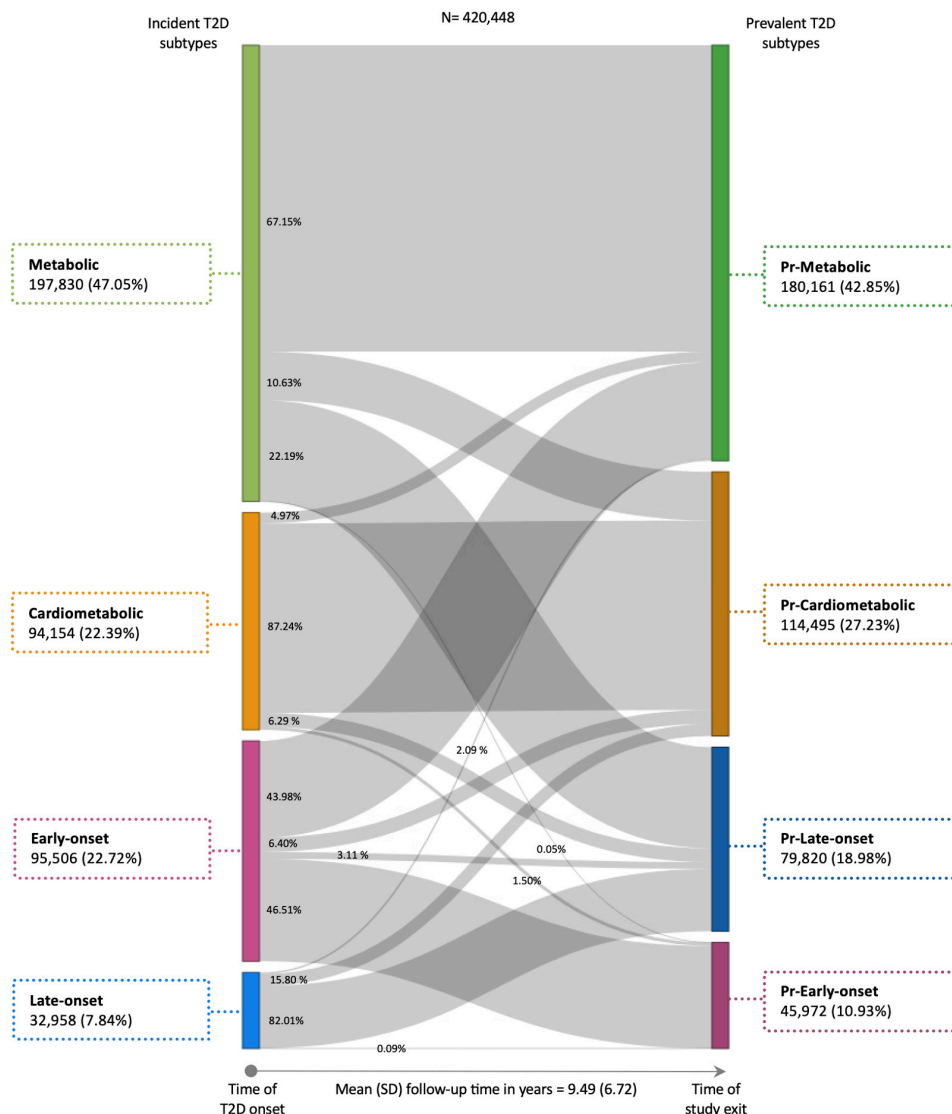


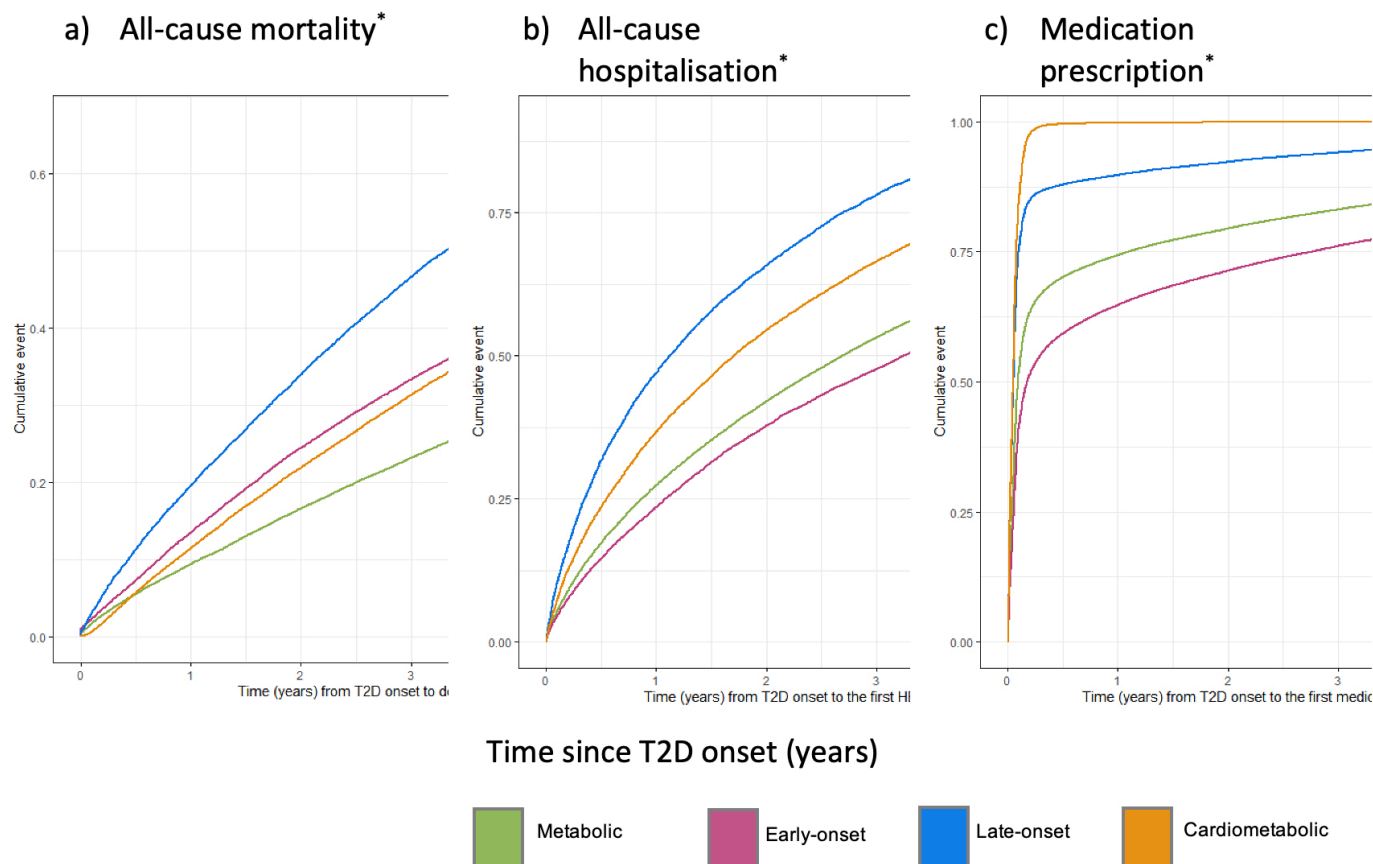
Figure 3 Trajectory of clusters from type 2 diabetes (T2D) onset to study exit with an average of 9.49 (SD 6.72) years of follow-up time.

diseases suggests that it may be more appropriate to use clustering approaches in combinations of diseases and over the life course, with implications for potential etiologies and mechanisms of disease subtypes.

T2D is associated with high morbidity, mortality and healthcare utilization and costs, particularly in the context of multiple long-term conditions.²⁹ We confirm the high burden of disease and high hospitalization rates, consistent with previous studies. There are clear differences in prognosis across the four subtypes, whether by all-cause mortality, hospital admissions or new diseases. For mortality, the worst prognosis is in the late-onset subtype and the best prognosis in the metabolic subtype, whereas for hospital admissions, the best prognosis is in the early-onset subtype. The high risk of developing severe complications, including over double the risk of lower limb amputation (HR 2.51, 1.51–4.17, $p < 0.0001$), neuropathy (HR 2.06, 1.63–2.59, $p < 0.0001$), and nephrotic syndrome (HR 2.05, 1.76–2.40, $p < 0.0001$) in the late-onset versus the metabolic subtype, illustrates

the major burden of healthcare need associated with T2D. The fact that the majority of individuals with T2D remained in the same subtype throughout the study period suggests that subtyping at time of diagnosis is likely to be clinically useful but requires external validation. There was a significant transition from metabolic to late-onset subtype, and from late-onset to cardiometabolic subtype over time. Both trajectories highlight the importance of preventing progression to more morbid or high-burden subtypes in people with T2D.

Investigation of differences in medication across subtypes of T2D may be instructive both in terms of understanding healthcare utilization and trajectory of disease, but also in informing future preventive and therapeutic strategies and clinical trials. Patients in the early-onset subtype exhibited a lower likelihood of medication prescription within 5 years after T2D onset than other clusters. Patients with T2D diagnosed at ages 19–40 are known to be under-represented in pharmaceutical studies,³⁰ and further investigation in other datasets is



*Log-rank $p < 0.0001$

Figure 4 Cumulative incidence rates over 5 years from type 2 diabetes (T2D) diagnosis to the first (A) all-cause mortality, (B) all-cause hospitalization, and (C) medication prescription.

required to assess this pattern within the early-onset cluster. Individuals in the late-onset subtype had the highest likelihood of medication prescription within 5 years after T2D diagnosis, adjusted for sex, age, and pre-T2D medication. In the cardiometabolic subtype, there was a high medication burden prior to diagnosis of T2D. There is evidence of a significant impact of adjusting for pre-T2D medication on the trajectory of medication in the cardiometabolic subtype, emphasizing the importance of considering both pre-T2D and post-T2D medication prescription in analysis of subtypes.

Strengths and limitations

Our study introduces T2D subtyping in the largest study population in a nationally representative dataset encompassing the most comprehensive range of variables. We have implemented and validated ML-based subtype definition using an established framework. Additionally, we have employed an explainable AI technique for subtype characterization, capturing non-linear and complex interconnection among variables and highlighting the positive and negative impacts of variables on cluster membership. Cluster characterization and internal validation were conducted through rigorous cross-validated supervised learning. Compared with simple, statistical subtyping based on single variables, we were able to the

assess a wide range of variables on cluster membership in more realistic models, and the representativeness of the study population makes our subtypes highly likely to be generalizable, unlike prior smaller, less representative studies.

There are several limitations to our study. First, we did not have access to biomarker variables in the clustering algorithm. While CPRD is a comprehensive, linked dataset, representative of the UK population, it is a generic EHR rather than a diabetes-specific dataset, leading to incomplete biomarker values at T2D onset or study exit. Therefore, our T2D phenotype relied primarily on clinical coding rather than biomarkers. Without complete biomarker variables at T2D onset in the data, there may be a minority of type 1 diabetes records in the cohort due to miscoding at the point of care. To minimize the risk of misclassification, we have used validated phenotype definitions exclusively based on explicit diagnostic codes for T2D.³¹ The interpretation of our findings must account for this potential misclassification of diabetes common to code-based phenotyping in EHR data. Second, additional incomplete variables were ethnicity, IMD, and behavioral factors for all individuals. Based on the internal validation findings, we used a complete feature rather than a complete case approach to mitigate these limitations and minimize the effects of incomplete variables.

Third, we only considered all-cause rather than cause-specific mortality and hospitalization, and average follow-up was 9.49 years. Fourth, while we have labeled the clusters as metabolic, early onset, late onset and cardiometabolic, they represent the overall attributes of cluster membership rather than strict clustering rules. For example, 'early onset' and 'late onset' are simplified labels to name the subtypes but should not be interpreted as definitive rules or cut-offs for categorizing patients into subtypes based on age alone. Incorporating a probabilistic soft clustering approach in the future could enhance subtype characterization. Fifth, as already stated, we have not yet externally validated the subtypes in other datasets for phenotypic or genetic replicability. Sixth, we only had data about medication prescription, not adherence. Finally, we have not explored the clinical utility or cost-effectiveness of the defined subtypes, which is needed prior to implementation.

CONCLUSION

We have developed four novel subtypes in T2D with potential application to research and clinical practice, which have clearly defined differences in baseline characteristics, outcomes and medications. Although these subtypes require external validation and assessment of clinical utility, they may have potential to improve design and implementation of prevention and treatment in people with T2D.

Author affiliations

¹University College London, London, UK

²British Heart Foundation Data Science Centre, Health Data Research UK, London, UK

³Peking Union Medical College Hospital, Beijing, China

⁴Diabetes Research Department, University of Leicester, Leicester, UK

⁵NHS England and NHS Improvement London, London, UK

⁶Imperial College Healthcare NHS Trust, London, UK

⁷AstraZeneca Cambridge Biomedical Campus, Cambridge, UK

⁸AstraZeneca, Cambridge, UK

⁹Barts Health NHS Trust, London, UK

Acknowledgements KK is supported by the National Institute for Health Research (NIHR) Applied Research Collaboration East Midlands (ARC EM), NIHR Global Research Centre for Multiple Long Term Conditions and the NIHR Leicester Biomedical Research Centre (BRC).

Contributors AB conceived the research question. AB, JBM, and TM obtained funding. MAM, LP, AB, and AD wrote the study protocol. MAM designed and led the analysis. AB and AD verified the data. AB and MAM drafted the initial and final versions of the manuscript. QZ contributed to the literature review. All authors critically reviewed the early and final versions of the manuscript. AB is the guarantor.

Funding AstraZeneca UK, Health Data Research UK.

Competing interests JBM, HG and TM are employed by AstraZeneca UK, a biopharmaceutical company, and declare stock and stock options. AB has received research funding from AstraZeneca. KK has acted as a consultant, speaker or received grants for investigator-initiated studies for AstraZeneca, Bayer, Novartis, Novo Nordisk, Sanofi-Aventis, Lilly and Merck Sharp & Dohme, Boehringer Ingelheim, Oramed Pharmaceuticals, Roche and Applied Therapeutics.

Patient consent for publication Not applicable.

Ethics approval The study protocol was approved by the Independent Scientific Advisory Committee (19_245).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available. The Clinical Practice Research Datalink (CPRD) data are not publicly accessible. All derived analysis and aggregated data are included in the manuscript and supplementary materials.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Mehrdad A Mizani <http://orcid.org/0000-0002-2441-895X>

He Gao <http://orcid.org/0000-0001-7764-198X>

Amitava Banerjee <http://orcid.org/0000-0001-8741-3411>

REFERENCES

- Magliano DJ, Boyko EJ, Balkau B, *et al*. IDF Diabetes Atlas | Tenth Edition, Available: https://diabetesatlas.org/idfawp/resource-files/2021/07/IDF_Atlas_10th_Edition_2021.pdf
- Reed J, Bain S, Kanamarlapudi V. A review of current trends with type 2 diabetes epidemiology, aetiology, pathogenesis, treatments and future perspectives. *Diabetes Metab Syndr Obes* 2021;14:3567–602.
- Zheng Y, Ley SH, Hu FB. Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. *Nat Rev Endocrinol* 2018;14:88–98.
- Mulnier HE, Seaman HE, Raleigh VS, *et al*. Mortality in people with type 2 diabetes in the UK. *Diabet Med* 2006;23:516–21.
- Khan MAB, Hashim MJ, King JK, *et al*. Epidemiology of type 2 diabetes – global burden of disease and forecasted trends. *J Epidemiol Glob Health* 2020;10:107–11.
- Dworzynski P, Aasbrenn M, Rostgaard K, *et al*. Nationwide prediction of type 2 diabetes Comorbidities. *Sci Rep* 2020;10:1776.
- Khan A, Uddin S, Srinivasan U. Comorbidity network for chronic disease: a novel approach to understand type 2 diabetes progression. *Int J Med Inform* 2018;115:1–9.
- Gregg EW, Sattar N, Ali MK. The changing face of diabetes complications. *Lancet Diabetes Endocrinol* 2016;4:537–47.
- Laaksonen MA, Knekt P, Rissanen H, *et al*. The relative importance of Modifiable potential risk factors of type 2 diabetes: a meta-analysis of two cohorts. *Eur J Epidemiol* 2010;25:115–24.
- Galicia-Garcia U, Benito-Vicente A, Jebbari S, *et al*. Pathophysiology of type 2 diabetes mellitus. *IJMS* 2020;21:6275.
- Wu Y, Ding Y, Tanaka Y, *et al*. Risk factors contributing to type 2 diabetes and recent advances in the treatment and prevention. *Int J Med Sci* 2014;11:1185–200.
- Tahrani AA, Bailey CJ, Del Prato S, *et al*. Management of type 2 diabetes: new and future developments in treatment. *The Lancet* 2011;378:182–97.
- Crandall JP, Knowler WC, Kahn SE, *et al*. The prevention of type 2 diabetes. *Nat Clin Pract Endocrinol Metab* 2008;4:382–93.
- Chan JCN, Lim L-L, Wareham NJ, *et al*. The lancet Commission on diabetes: using data to transform diabetes care and patient lives. *The Lancet* 2020;396:2019–82.
- NICE. Type 2 diabetes in adults: management. 2015. Available: <https://www.nice.org.uk/guidance/ng28/chapter/Recommendations#individualised-care>
- Ahlqvist E, Prasad RB, Groop L. Subtypes of type 2 diabetes determined from clinical parameters. *Diabetes* 2020;69:2086–93.
- Hwang Y-C, Ahn H-Y, Jun JE, *et al*. Subtypes of type 2 diabetes and their association with outcomes in Korean adults – a cluster analysis of community-based prospective cohort. *Metabolism* 2023;141:155514.
- Slieker RC, Donnelly LA, Fitipaldi H, *et al*. Replication and cross-validation of type 2 diabetes subtypes based on clinical variables: an IMI-RHAPSODY study. *Diabetologia* 2021;64:1982–9.

- 19 Wang Y, Katzmarzyk PT, Horswell R, *et al.* Comparison of the heart failure risk stratification performance of the CKD-EPI equation and the MDRD equation for estimated glomerular filtration rate in patients with type 2 diabetes. *Diabet Med* 2016;33:609–20.
- 20 Banerjee A, Chen S, Fatemifar G, *et al.* Machine learning for subtype definition and risk prediction in heart failure, acute coronary syndromes and atrial fibrillation: systematic review of validity and clinical utility. *BMC Med* 2021;19:85.
- 21 Dashtban A, Mizani MA, Pasea L, *et al.* Identifying subtypes of chronic kidney disease with machine learning: development, internal validation and Prognostic validation using linked electronic health records in 350,067 individuals. *EBioMedicine* 2023;89:104489.
- 22 Ke G, Meng Q, Finley T, *et al.* Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems* (Curran Associates, Inc); 2017 Available: <https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html>
- 23 Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* (Curran Associates, Inc); 2017 Available: <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>
- 24 Rodríguez-Pérez R, Bajorath J. Interpretation of machine learning models using Shapley values: application to compound potency and multi-target activity predictions. *J Comput Aided Mol Des* 2020;34:1013–26.
- 25 Misra S, Wagner R, Ozkan B, *et al.* Precision subclassification of type 2 diabetes: a systematic review. *Commun Med (Lond)* 2023;3:138.
- 26 Leslie RD, Ma RCW, Franks PW, *et al.* Understanding diabetes heterogeneity: key steps towards precision medicine in diabetes. *Lancet Diabetes Endocrinol* 2023;11:848–60.
- 27 Lugner M, Gudbjörnsdóttir S, Sattar N, *et al.* Comparison between data-driven clusters and models based on clinical features to predict outcomes in type 2 diabetes: nationwide observational study. *Diabetologia* 2021;64:1973–81.
- 28 Groop L, SND, Swedish National Data Service. All new diabetics in Scania - ANDIS. 2008. Available: <https://snd.gu.se/en/catalogue/dataset/ext0057-1>
- 29 Pasea L, Dashtban A, Mizani M, *et al.* Risk factors, outcomes and healthcare utilisation in individuals with multimorbidity including heart failure, chronic kidney disease and type 2 diabetes mellitus: a national electronic health record study. *Open Heart* 2023;10:e002332.
- 30 Misra S, Ke C, Srinivasan S, *et al.* Current insights and emerging trends in early-onset type 2 diabetes. *Lancet Diabetes Endocrinol* 2023;11:768–82.
- 31 Tate AR, Dungey S, Glew S, *et al.* Quality of recording of diabetes in the UK: how does the GP's method of coding clinical data affect incidence estimates? cross-sectional study using the CPRD database. *BMJ Open* 2017;7:e012905.