Article type: Letter

Title: Visualising the heterogeneity in phenotype, disease progression and drug response in a large population with type 2 diabetes

Authors: Anand Thakarakkattil Narayanan Nair[1], Agata Wesolowska-Andersen[2], Caroline Brorsson[3], Aravind Lathika Rajendrakumar[1], Simona Hapca[1], Sushrima Gan[1], Adem Y. Dawed[1], Louise A. Donnelly[1], Rory McCrimmon[1], Alex S F Doney[1], Colin N A Palmer[1], Mohan Viswanathan[4], Ranjit M Anjana[4], Andrew T Hattersley[5], John M Dennis[5], Ewan R Pearson[1].

Author Affiliations

[1] Division of Population Health and Genomics, School of Medicine, University of Dundee, Dundee, UK
[2] Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK
[3] Novo Nordisk Center for Protein Research, University of Copenhagen, Copenhagen, Denmark
[4] Madras Diabetes Research Foundation, Chennai, India
[5] Institute of Biomedical and Clinical Science, Royal Devon and Exeter Hospital, University of Exeter,UK

Address for Correspondence

Ewan Pearson
Professor of Diabetic Medicine
Population Health & Genomics
School of Medicine
University of Dundee
Dundee, DD1 9SY

Email e.z.pearson@dundee.ac.uk

Tel +44 1382 383387

Word count : 2185 (excluding introductory paragraph and methods)

Type 2 diabetes (T2D) is a complex chronic disease associated with substantial morbidity and mortality.[1] The consensus from genetic studies of T2D, reflected in the palette model, is that T2D development is influenced by multiple aetiological processes.[2] While some people may develop T2D because of extreme alteration in one or two such molecular processes; in most, multiple pathways are involved.[2,3] This results in considerable heterogeneity in the phenotype of T2D, that is largely ignored in individual patient management. In acknowledgement of this heterogeneity there have been recent attempts to place patients into discrete phenotypic clusters.[4,5] However, such an approach does not align with the emerging molecular understanding of disease architecture of T2D. We used a reverse graph embedding method to reduce multiple phenotypic characteristics into a non-linear tree structure. In an overall population of 23,137 people with newly diagnosed T2D, we show a continuum of phenotypic variation across the tree and show that this maps to different diabetes outcomes, in terms of variation in antihyperglycemic drug failure, risk of progression to insulin requirement and risk of micro and macrovascular complications. We validated our findings in two independent cohorts and developed an application to visualise the position of patients with new T2D within a two-dimensional space, with their 10 year risk of diabetes complications. We also find a distribution of T2D partitioned polygenic scores[6] across the tree consistent with variation in causal aetiological processes. Overall, our data highlights how the underlying phenotypic variation driving T2D onset impacts on subsequent diabetes outcomes and drug response, and the need to incorporate this into personalized treatment approaches for the management of type 2 diabetes.

We used electronic health record information of T2D cases from Tayside and Fife in Scotland[7] diagnosed between 1993-2017 to identify individuals who were not GAD antibody positive and diagnosed with T2D on or after the age of 35. To explore phenotypic heterogeneity, we used nine characteristics that are well captured in routine clinical care and reflect a range of measures known to be altered in T2D and associated with variation in risk of adverse diabetes outcomes: HbA1c, BMI, Total Cholesterol (TC), HDL-Cholesterol (HDL-C), Triglyceride (TG), Alanine Aminotransferase (ALT), Creatinine, and systolic and diastolic blood pressure (SBP & DBP). These were measured within one year from diagnosis closest to the point of diagnosis, normalised and residualised for age and sex. Patient selection is outlined in S*upplementary Figure 1*; 23,137 patients were included with complete data. The phenotypic characteristics of the study population are described in *Supplementary Table 1.*

To enable reduction of the phenotype data at diagnosis of T2D, we used DDRTree (Discriminative dimensionality reduction via learning a tree); which is a dimensionality reduction algorithm utilising reverse graph embedding.[8,9] This reduces the multi-dimensional data and projects into a low-dimensional space in the form of a minimal spanning tree structure. A principal tree trunk is at the centre of the data with the algorithm identifying a branch clustering structure of the data points in the reduced dimension. Each resulting branch will include individuals with an increasingly similar and distinctive pattern of phenotypes extending to the tip, while individuals located more proximally in the principal tree trunk will have mixed characteristics. Whilst data is continuous across the tree, consistent with the palette model[2], the patients with the most extreme distinct phenotypes are found at the end of each branch. For the internal validation process, we gave number to each tree branch (1-6) and excluded the individuals at the centre of data. We then tested the assignment of individuals to tree branches across multiple iterations of the DDRTree algorithm resulting in an adjusted rand index of 0.79, 0.74-0.84 (median, IQ range) which is indicative of stability of

the algorithm. To enable external validation of patient outcomes and to assess contribution of genetic risk scores we treated the tree derived from the Scottish data as a reference structure. We then developed a mapping function to map individuals with T2D from UK Biobank and the ADOPT clinical trial to the reference tree using the age of diagnosis, sex and the nine phenotypes used to define the reference tree. For validation, we used UK Biobank (UKBB) primary care data[10] of patients with newly diagnosed T2D (N=7332, details of sample selection are described in *Supplementary Figure 2,* and characteristics of the study population is given in *Supplementary Table 1)* and the ADOPT randomised controlled trial of monotherapy with metformin, rosiglitazone and glibenclamide[11] (N=4150, details of sample selection are described in *Supplementary Figure 3,* and characteristics of the study population are given in *Supplementary Table 1*).

To visualise and characterise the phenotypic variation present across the tree we overlaid the phenotypic data, as shown in *Figure 1A*. Other than the visual gradient, we used two metrics to assess the distribution of the phenotype across the tree: in *Figure 1B* we regress each phenotype against each of the two tree dimensions, and in *Figure 1C* we plot the Moran's I values for each phenotype representing the strength of spatial correlation across the tree. Based upon all these measures, of the phenotypic characteristics used to define the tree, HDL-C, SBP & DBP were most strongly distributed across the tree, followed by total cholesterol and triglycerides and then HbA1c and BMI. There was minimal variation in creatinine or ALT across the tree. The individuals located in left upper part of the tree had elevated HDL-C levels; those in the right upper tree had higher levels of blood pressure, cholesterol, and moderate levels hyperglycemia. The lower right part of the tree contained patients who were obese, hyperglycemic, with high triglycerides levels and low HDL-C. We saw no difference when we overlaid individuals whose BP measures or lipid measures were untreated vs treated. As expected, given we mapped individuals to the Scottish reference tree, the phenotype distribution was the same for the UK Biobank participants (*Supplementary Figure 4*) and the ADOPT clinical trial (*Supplementary Figure* 5). Of note, however, patients recruited to the ADOPT trial did not map to the tree areas with particularly high HbA1c or BMI as these patients were excluded from the trial.

A subset of Scottish individuals had measurements of C-peptide (n=3604), adiponectin (n=965) and leptin (n=742). While these measurements were not made at diagnosis, when these data are overlaid on the tree (S*upplementary Figure 6 A-D*) they show that the adiponectin largely positively correlated with the dimension 2 (similar to HDL-C), while C-peptide and leptin was positively associated with dimension 1 (similar to adiposity, hyperglycaemia, and dyslipidaemia) with the highest C-peptide and leptin concentrations seen in lower part of the tree. Fasting measures of beta-cell function and insulin sensitivity (HOMA B and HOMA IR) were available for the ADOPT trial at recruitment. When this data is overlayed upon the tree (S*upplementary Figure 7*) there was no visible gradient seen for beta-cell function, but insulin resistance showed a distribution opposite to HDL-Cholesterol (with high HDL-C and lower HOMA IR in the top left quadrant of the tree).

We then investigated how the phenotypic variation at diagnosis translates to variation in four diabetes outcomes: time to insulin requirement; time to any diabetic retinopathy (DR); time to chronic kidney disease (CKD, eGFR<60); time to Major Adverse Cardiovascular Event (MACE). We used cox proportional hazard model and competing risk (Fine and Gray) model[12] from diagnosis of T2D, with death as competing risk and diabetes outcomes as events of interest and calculated the individual probabilities (event probabilities) of progression to each

of these outcomes (over 5/10 years) for each patient using the coordinates of each individual in the two-dimensional space (i.e. position in the tree). These probabilities for the discovery (Scottish, Tayside & Fife) cohort are overlaid on the tree in *Figure 2(A-D).* To assess the distribution of event probabilities we report the sub hazard ratio (sHR) from the competing risk model constructed with DDRTree dimensions (*Figure 2 E*) and the strength of spatial correlation (Moran's I) is shown in *Figure 2 F* showing strong spatial correlation of each outcome across *the tree*. The probabilities of time to insulin, MACE and CKD had a very similar pattern across the tree, with greatest risk observed in the most obese and dyslipidemic individuals (bottom right). By contrast, the risk of retinopathy showed a different pattern with risk largely driven by the combination of increased blood pressure and hyperglycemia (top right). Consistent with this, the UKDPS study found that blood pressure lowering from 159mmHg to 144mmHg on average reduced the progression of retinopathy[13], although this contrasts with the ACCORD study where lowering BP from 138mmHg to 114mmHg had little impact on retinopathy progression.[14] Scottish individuals in the top right of the tree had SBP of 160-180mmHg more in keeping with patients included in UKPDS. We saw a similar picture if we used referable retinopathy as the endpoint. In addition, we constructed the competing risk models with continuous variables (eg: HbA1c, BMI, HDL-c etc) as covariates and derived the event probabilities for each individual. These probabilities are overlaid on the tree (*Supplementary Figure 8 A-F)* with a similar pattern seen for those derived from models developed using the DDRTree dimensions. Competing risk model summaries for each outcome are provided in *Supplementary Table 2*. The performance for these competing risk models was assessed with the C-index, which is measure of discrimination with higher values indicating better models,[15,16] with internal validation and estimated the C-index for the four outcome models to range from 56.6-76.9 (*Supplementary Table3)*. For validation, we derived MACE, CKD, and insulin requirement outcomes for UK Biobank (*figure 3 A-C*) and CKD for ADOPT (*figure 4A and 4E*) and show a very similar risk distribution for these outcomes across the tree.

We then investigated how drug response varied with phenotypic variation. Here we focused on the ADOPT study, where patients were randomised soon after diagnosis to metformin, glibenclamide (Sulphonylurea) or rosiglitazone (Thiazolidinedione). We show a striking difference in monotherapy failure for these drugs, with metformin and sulphonylurea failure being faster in those obese hyperglycaemic individuals towards the bottom right of the tree, but with TZD failure quicker in those in the lower left of the tree (*Figure 4 B-D and 4 F-I*).

Having established a visual representation of the phenotypic variation at diagnosis and how this impacts on diabetes outcome and drug response, we moved to explore the heterogeneity in the genetic aetiology of diabetes to provide insight into causal processes underlying the phenotypic heterogeneity represented by the tree. For this purpose, we constructed partitioned polygenic scores (pPS) for beta cell dysfunction (with high proinsulin), proinsulin (beta-cell dysfunction with low proinsulin secretion), BMI, lipodystrophy, and liver/lipid, as described in *Udler et al.* We overlaid the pPS over the tree structure where we had genetic data (Scottish, Tayside & Fife=3512, UK Biobank,n=7145). Given the small genetic effect and small sample size there was no clear visible gradient of the pPS across the tree. However, when regressing against the tree dimensions (*Figure 5A*), dimension 1 (X-axis) was positively associated with obesity pPS and inversely associated with beta cell, proinsulin, and liver GRS, while the dimension 2 (Y-axis) was positively associated with beta cell, proinsulin and liver GRS and inversely associated with lipodystrophy pPS. *Figure 5B* shows that the beta-cell pPS

and obesity pPS were significantly spatially autocorrelated. As depicted in *figure 5C* the lower right of the tree had higher genetic obesity, the upper left of the tree had elevated genetic beta-cell dysfunction and increased diabetes risk mediated via liver/lipid mediated insulin resistance; the bottom left was characterized by genetic lipodystrophy with lower genetic obesity – interestingly this is the quadrant where thiazolidinediones were least durable.

In summary, we have applied a novel approach to data dimensionality reduction of health record data of 23,137 individuals with T2D with a complete set of clinically relevant phenotype data available at the time of diagnosis. The resulting tree structure provides a two-dimensional representation of the increasingly recognized complex phenotypic variation in patients with T2D and allows an overlay of clinically relevant phenotypes and complications. We have validated the outcomes in two independent data sets. In addition, this analysis provides insights into heterogeneity in diabetes at the genetic level through the significant association of pPS with the tree dimensions which imply that some of the observed heterogeneity is mediated via causal aetiological processes for type 2 diabetes. We have demonstrated how this approach enables an intuitive visual representation of how phenotypic variation translates to variation in glycaemic response to medication, glycaemic deterioration, and risk of micro and macrovascular disease. As a potential aid for assisting clinicians and their patients to demonstrate and visualize individual patient profiles at T2D diagnosis and how this indicates risks of disease progression and complications, we have developed an app (https://atn-uod2018.shinyapps.io/Prediction_diabetes_outcome_18082021/). Newly diagnosed patients can be placed in the diabetes continuum (tree structure) with their risk of micro- and macro-vascular complications predicted for a 10-year period, derived using age, sex, and the nine other continuous phenotypic measures used to define the tree.

Our approach differs from recent studies that have derived subtypes of T2D in a number of ways.[4,17–19] Firstly, the phenotypic data used in deriving the tree was adjusted for age and sex. We chose to do this as age, in particular, is a major driver of adverse outcomes – especially for risk of insulin initiation, MACE and CKD. Thus, the overlay of MACE and CKD risk on the tree is not mediated by an underlying distribution of age or sex. Secondly, whilst the DDRTree method groups together patients who are similar within the tree branches, it is clear from the figures, that the phenotype and the associated risks are distributed continuously along the whole tree. This highlights how, based upon the phenotypes included in the study, any discrete binning into subgroups will lose information as demonstrated by Dennis et al[20], and mis-represent the continuum that is T2D. A recent analysis of Swedish National Diabetes Register with 114,231 individuals with newly diagnosed T2D cases also reported the use of prediction models with simple clinical phenotypes outperform the models with cluster labels and they failed to identify T2D clusters while using nine continuous phenotypes at diagnosis.[21] We are not advocating the use of DDRTree to improve prediction, but rather to reduce a complex multi-dimensional disease into a simpler intuitively understandable two-dimensional model that can be readily visualized and used to enhance the therapeutic process between clinicians and individual patients, to see how their personal T2D profile compares to others of similar age and sex. The app enables the exploration of how lifestyle or pharmacological intervention to improve modifiable risk factors such as blood pressure, lipids, BMI and HbA1c might influence a patient's position in the tree and their subsequent risk of complications.

Current clinical guidelines for the management of T2D generally do not consider individual patient phenotype when considering what is the optimal treatment, or what are the risks of

progression to insulin or microvascular disease, yet we highlight how outcome and drug response varies considerably across the complex phenotypic spectrum. While demonstration of the clinical value of such information will require validation in a prospective trial, our data supports the concept that the management of individual patients with T2D should be informed by their specific phenotypic profile e.g., with respect to retinal screening intervals, monitoring of HbA1c, and consideration of optimal diabetes treatment and patient lifestyle choices. Incorporation of individual phenotypic variation into clinical practice has clear potential to make a significant contribution to a precision approach to the management of T2D.

Methods

The cohort and variable definitions

We used three datasets for this analysis (i) Scottish (Tayside & Fife); a subset of this cohort had consented for genetic analysis as part of the Genetics of Diabetes Audit and Research in Tayside and Fife (GoDARTS) study (ii) UKBB primary care data (iii) A Diabetes Outcome Progression Trial (ADOPT) data.

Scottish cohort

We used electronic health record information of type 2 diabetes cases from Tayside and Fife in Scotland diagnosed between 1993-2017. The Scottish Care Information -Diabetes Collaboration (Tayside & Fife) has longitudinal data on biochemical investigations and prescriptions.[7] We excluded those who were known to be GAD positive and those with age of T2D diagnosis before 35 years of age. We used eleven phenotypes, which includes age of diagnosis, sex, HbA1c, BMI, HDL-Cholesterol (HDL-C), Triglyceride (TG), Total Cholesterol (TC), ALT, Creatinine, Systolic and diastolic blood pressure (SBP & DBP) at diagnosis. All covariates were measured within one-year from the date of diagnosis and if multiple recordings were available the measurement closest to date of diagnosis was used. For a sub population C-peptide, adiponectin and leptin measurements were available, these measurements at recruitment rather than diagnosis of diabetes. A subset had also consented for genetic analysis (genome wide genotyping) as part of the GoDARTS Study (1998-2015)[22] and this was used for derivation of partitioned polygenic scores.

UKBB Primary care data

The UK (United Kingdom) Biobank (UKBB) is a large prospective epidemiological resource with consented data of 500,000 individuals recruited during the period of 2006-2010. UKBB contains information from electronic health records as well as from interviews and questionnaire and a proportion of individuals consented were genotyped also. The UKBB resource was designed with an objective to improve the prevention, diagnosis and treatment of non-communicable disease including cardiovascular diseases, diabetes, and cancer.[10,23] Currently, UKBB resources are also utilized for COVID-19 research.[24] Longitudinal primary care data of around 45% of the UKBB cohort has been made available to facilitate disease progression-based analysis.[25]

We identified T2D cases from the longitudinal and cross-sectional data based on their diagnosis labels and antihyperglycemic prescriptions and collected genetic data for these T2D cases. The phenotype at diagnosis was defined as recorded values in the primary care data set one year before or after diagnosis of T2D. The genetic data available on all UKBB participants was used to derive partitioned polygenic scores(pPS).

ADOPT trial data.

ADOPT was a multicentre randomised controlled trial of rosiglitazone, glibenclamide (glyburide) and metformin in recently diagnosed cases with T2D.[11] This trial had a multiethnic study population with majority 88.3% being Caucasian. The trial follow up period was four years (2002-2006) and inclusion criteria for the trial was age of T2D diagnosis between 30-75 years with fasting glucose levels 7-13 mmol/L. During the follow up period HbA1c, change in beta cell function and Insulin sensitivity and other diabetes related phenotypes were assessed at regular time points.

Statistical Analysis

DDRTree

DDRTree is a dimensionality reduction approach with a reverse graph structure embedding algorithm and a mapping function. It was initially developed for the analysis of high dimensional single cell transcriptomics data with an aim of understanding the role of gene regulation in cell fate.[9,26] We used the implementation of DDRTree from Monocle 2.[8] In this analysis data bifurcation was done with DDRTree and this method showed better accuracy and normalized mutual information (NMI) compared to local linear embedding (LLE) and principal component analysis (PCA) with different datasets.[9]

We excluded outliers from data based on 5 SD values and data was transformed with rank normalization. In a second step, we residualised each phenotype for age and sex using linear regression analysis. This age and sex residualised matrix of phenotypes was entered to the Monocle 2 algorithm without any initial dimensionality reduction. The DDRTree algorithm reduces high dimensional data (residualised nine phenotypes) and projects it into a two-dimensional space. Later a smoothed graph structure is constructed from the reduced data and then similar data points are grouped together to obtain a tree structure with tree branches and group of individuals located in tree trunk is considered as center of the data.

To validate DDRTree, we gave a label to tree ends (1-6) and internal validation was conducted by estimating the Adjusted Rand Index (ARI) which is a measure of agreement for positioning the individual in same tree branch at different execution of the algorithm. We considered an adjusted rand index of values above 0.75 as an indication of stability of the method. We ran the DDRTree algorithm on 10000 randomly selected individual's phenotype data and re-ran it for 500 times excluding 10% of individuals randomly in each run. While estimating the adjusted rand index, we excluded the contribution from individuals located in the principal tree trunk.

We used R version 3.5.2 for all data management and statistical analysis,[27] and 'monocle2' package from Bioconductor for implementing DDRTree.

Endpoint definitions

Insulin Initiation: This was defined as sustained use of insulin for more than 6 months or a clinical requirement for insulin, indicated as two or more HbA1c reading >=8.5% more than three months apart while taking two or more oral antihyperglycaemic agents.[28]

Diabetic retinopathy (DR): We used the Scottish Diabetic Retinopathy Grading Scheme and any grade R1, R2, R3 and R4 (background retinopathy and above) is considered as incidence of diabetic retinopathy.[29] As a sensitivity analysis we used referable retinopathy as an endpoint – being either R3 or R4.

Major Adverse Cardiac Events (MACE): We identified MACE from Scottish Morbidity Records (SMR, for Scottish cohort) and Hospital episode statistics (HES, for UKBB) using ICD 10 and ICD 9 codes (ICD 10- codes I20-I25 and I60-I69, ICD 9- codes 410-414 and 430-438).

Chronic Kidney Disease (CKD): CKD was identified from the electronic health records based upon an estimated Glomerular Filtration Rate (eGFR) of <= 60 ml/min/1.73m^2 on at least 2 reading which were 90 days apart.[30] eGFR was calculated from serum creatinine using the CKD-EPI equation.[31]

Drug failure: In the ADOPT trial, monotherapy failure was indicated by Fasting Plasma Glucose (FPG) >180 mg/dl (>10mmol/L) after at least 6 weeks of initiation of treatment with a tolerated dose of antihyperglycaemic drugs.

Modelling of risk of glycemic deterioration, micro and macrovascular complications.

We assessed the risk of diabetes progression using a competing risk model (Fine and Gray model) and derived sub hazard ratios with death as a competing event in the Scottish cohort and UKBB.[32] For this purpose, we developed five competing risk models, one for each diabetes related outcome and in all models death was considered as a 'competing event' which hinders the incidence of the 'event of interest'. Major diabetes progression end points were insulin initiation, diabetic retinopathy (DR), Major Adverse Cardiac Events (MACE) and chronic kidney disease (CKD). In each model we considered T2D diagnosis as the baseline and we excluded individuals who did not have follow up data or individuals who had already experienced the event of interest (Insulin Initiation/DR/MACE/CKD) from the corresponding competing risk models. Later we constructed Fine and Gray models by using the tree dimensions from DDRTree, or continuous variables as covariates in the model.

To obtain the individual probability for developing each diabetes outcome (eg: CKD or MACE) for each study participant over a 10-year follow up period, we used previously constructed competing risk models with dimensions from DDRTree as covariates. For example, to estimate the probability of incidence of CKD for a study participant, we used the CKD competing risk model (event of interest: CKD, competing event: death) with DDRTree dimensions as covariates and predicted event probability for that individual. Similarly, we calculated individual level probability for the incidence of insulin initiation, MACE and DR using corresponding competing risk models. These event probabilities were overlaid on the tree diagram to visualize the heterogeneity in diabetes progression. A similar approach was followed in ADOPT data while using cox proportional hazard model for deriving event probabilities.

External validation of analysis

Development of a mapping function: To map individuals with newly diagnosed T2D to the Scottish tree, we constructed a 'mapping function'. This considers the 11 diabetes related phenotypes (age of T2D diagnosis, Sex, HbA1c, BMI, HDL-C, TC, TG, SBP, DBP, ALT and Creatinine) and assigns each individual a position in the Scottish tree. The mapping function is built with two components (i) two generalised additive models (GAM) with smooth terms fitted with cubic regression splines which will predict the DDRTree dimensions (dimension 1 and dimension2, the coordinates in 2D space) with 11 phenotypes as covariates (*Supplementary Table 4*). (ii) a distance estimating algorithm, which will estimate the Euclidean distance between two points in 2D space. So, when all phenotypes are given for a newly diagnosed T2D, the trained spline model will predict the dimension 1 and dimension 2 for the individual, which is a provisional position for the individual. In second step, a distance estimating algorithm was used to calculate the distance between the provisional position of the individual and all the positions in reference tree. Then the mapping function will assign a new position to the individual in reference tree, which will have shortest distance to the provisional point. In this way we identify the individual (from 23,137 people) who is most similar to the new individual being mapped. We used adj R^2 of the model as an indicator of model performance, the model used to predict the dimension1 had an adj R^2 0.76 and an adj R^2 0.86 for the model used to predict dimension 2.

We used the UKBB primary care data for the external validation, by using the age of diagnosis, sex and other nine phenotype we mapped the individuals in UKBB data to reference tree. Later, we overlayed the nine phenotypes over this tree and assessed the gradient of phenotype across tree and compared it with reference tree. To assess diabetes progression, we used three end points available in UKBB data, (i) time to insulin initiation (ii) time to CKD and (iii) time to MACE. We followed the similar definitions for the diabetes outcomes as described in earlier. We constructed competing risk models with death as competing event and used Fine and Gray models as before. The predicted probability of each event was overlayed on the tree.

We applied the 'mapping function' for ADOPT trial participants and followed the same steps as for UKPDS data. We assessed the distribution of beta cell function and insulin resistance over the ADOPT tree. In ADOPT we then assessed time to CKD, and time to metformin, sulphonylurea and thiazolidinedione failure using cox proportional hazard models and plotted the predicted event probabilities from these models over the tree.

Statistical evaluation of phenotypes or outcome distribution across the tree

Whilst most phenotypes and outcomes could clearly be seen to vary across the tree, we quantified this variation in two ways. First, for phenotype distribution, we regressed each phenotype against the tree dimensions and plot the regression coefficient and 95% CI. To assess distribution for each diabetes outcome probability, we used the sub hazard ratios (sHR) from the competing risk model. From each competing risk model, we derived sHR with 95% CI and these sHR indicated how the risk of diabetes outcomes (probability of Insulin initiation/DR/MACE/CKD) varies across the two-dimensional space (X and Y axis). Secondly, we undertook spatial autocorrelation analysis using Moran's I statistic. For this we considered each individuals location (with their coordinate in space) and assessed its relationship with

values associated with these locations.[33] A positive value of Moran's I suggests similar values (phenotype/GRS/event probabilities) are located together and close while a negative value indicates dissimilar values are located close.

Genetic analysis

In order to deconstruct the heterogeneity of type 2 diabetes *Udler et al* conducted a soft clustering analysis of 94 established type 2 diabetes genetic loci and 47 diabetes related traits using Bayesian nonnegative matrix factorization (bNMF)[34] and identified five clusters. Out of five clusters, two of them were related to beta cell function but differed by proinsulin levels and three of them were related to insulin resistance mediated through obesity, lipids, and lipodystrophy. Based on this we constructed five partitioned polygenic scores (pPS) in Scottish (GoDARTS) patients and UKBB patients as previously defined and labelled as beta cell dysfunction, proinsulin, body mass index (BMI), lipodystrophy and liver/lipid.

We assessed the relation between these five pPS and DDRTree dimensions (dimension 1 and dimension 2) with linear regression models in both GoDARTS and UKBB data. The significant positive relation with pPS and dimension 1 (X- axis) indicate as the individuals located in the origin of X-axis have lesser or higher risk of corresponding pPS compared to the individual located at the extremes of axis. In GoDARTS data we assessed the spatial distribution of pPS using spatial autocorrelation (Moran's I).

Development of the R Shiny app and prediction models for micro and macro vascular complications

As we described above, we use the 'mapping function' to find the position of a newly diagnosed T2D patient in the tree. To predict the risk of micro and macro vascular complications we used the Fine-Gray models constructed with all continuous variables (age of diagnosis, sex, HbA1c, HDL-C, BMI, TG, TC, ALT, creatinine, and systolic and diastolic blood pressure) for each outcome. For each model, we assessed the shape of the relationship between phenotype and outcome and undertook transformations as require. Similarly, the time varying effects of the phenotype were assessed and those with time varying effects were included in the models. In addition to the phenotype, we assessed the use of statins or antihypertensive at the time of diabetes diagnosis and used in the model building process. Then we predicted the event probabilities for our event of interest at each of multiple time points from diagnosis through to 10 years after T2D diagnosis.

References

1. Tuomi T, Santoro N, Caprio S, Cai M, Weng J, Groop L. The many faces of diabetes: A disease with increasing heterogeneity. Vol. 383, The Lancet. Lancet Publishing Group; 2014. p. 1084–94.

2. McCarthy MI. Painting a new picture of personalised medicine for diabetes. Diabetologia. 2017 May 1;60(5):793–9.

3. Pearson ER. Type 2 diabetes: a multifaceted disease. Diabetologia. 2019 Jul 1;62(7):1107–12.

4. Ahlqvist E, Storm P, Käräjämäki A, Martinell M, Dorkhan M, Carlsson A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. Lancet Diabetes Endocrinol. 2018 May 1;6(5):361–9.

5. Zaharia OP, Strassburger K, Strom A, Bönhof GJ, Karusheva Y, Antoniou S, et al. Risk of diabetes-associated diseases in subgroups of patients with recent-onset diabetes: a 5-year follow-up study. Lancet Diabetes Endocrinol. 2019 Sep 1;7(9):684–94.

6. Udler MS, Kim J, von Grotthuss M, Bonàs-Guarch S, Cole JB, Chiou J, et al. Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. Langenberg C, editor. PLOS Med. 2018 Sep 21;15(9):e1002654.

7. Scottish Care Information - Diabetes Collaboration (SCI-Diabetes) [Internet]. [cited 2021 Feb 13]. Available from: https://www.sci-diabetes.scot.nhs.uk/

8. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, et al. Reversed graph embedding resolves complex single-cell trajectories. Nat Methods. 2017;14(10):979–82.

9. Mao Q, Wang L, Goodison S, Sun Y. Dimensionality reduction via graph structure learning. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery; 2015. p. 765–74.

10. UK Biobank: a large scale prospective epidemiological resource - Health Research Authority [Internet]. [cited 2021 Mar 1]. Available from: https://www.hra.nhs.uk/planning-and-improving-research/application-summaries/research-summaries/uk-biobank-a-large-scale-prospective-epidemiological-resource/

11. Viberti G, Kahn SE, Greene DA, Herman WH, Zinman B, Holman RR, et al. A Diabetes Outcome Progression Trial (ADOPT): An international multicenter study of the comparative efficacy of rosiglitazone, glyburide, and metformin in recently diagnosed type 2 diabetes. Diabetes Care. 2002;25(10):1737–43.

12. Austin PC, Lee DS, Fine JP. Introduction to the Analysis of Survival Data in the Presence of Competing Risks. Circulation. 2016 Feb 9;133(6):601–9.

13. Turner R, Holman R, Stratton I, Cull C, Frighi V, Manley S, et al. Tight blood pressure control and risk of macrovascular and microvascular complications in type 2 diabetes: UKPDS 38. Br Med J. 1998;317(7160):703–13.

14. EY C, WT A, MD D, RP D, S G, CM G, et al. Effects of medical therapies on retinopathy

progression in type 2 diabetes. N Engl J Med. 2010 Jul 15;363(3):233–44.

15. Jr HF, KL L, DB M. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996 Feb 1;15(4):361–87.

16. M W, MT K, JC W, EW S. Prognostic models with competing risks: methods and application to coronary risk prediction. Epidemiology. 2009;20(4):555–61.

17. Zou X, Zhou X, Zhu Z, Ji L. Novel subgroups of patients with adult-onset diabetes in Chinese and US populations. Vol. 7, The Lancet Diabetes and Endocrinology. Lancet Publishing Group; 2019. p. 9–11.

18. Li L, Cheng WY, Glicksberg BS, Gottesman O, Tamler R, Chen R, et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. Sci Transl Med. 2015;7(311).

19. Slieker RC, Donnelly LA, Fitipaldi H, Bouland GA, Giordano GN, Åkerlund M, et al. Replication and cross-validation of type 2 diabetes subtypes based on clinical variables: an IMI-RHAPSODY study. Diabetologia. 2021 Jun 10;1–8.

20. Dennis JM, Shields BM, Henley WE, Jones AG, Hattersley AT. Disease progression and treatment response in data-driven subgroups of type 2 diabetes compared with models based on simple clinical features: an analysis using clinical trial data. Lancet Diabetes Endocrinol. 2019 Jun 1;7(6):442–51.

21. Lugner M, Gudbjörnsdottir S, Sattar N, Svensson AM, Miftaraj M, Eeg-Olofsson K, 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 May 31;1–9.

22. Hébert HL, Shepherd B, Milburn K, Veluchamy A, Meng W, Carr F, et al. Cohort Profile: Genetics of Diabetes Audit and Research in Tayside Scotland (GoDARTS). Int J Epidemiol. 2018 Apr 1;47(2):380-381j.

23. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Med. 2015 Mar 31;12(3):e1001779.

24. New primary care data added from medical records to aid vital COVID-19 research [Internet]. [cited 2021 Mar 1]. Available from: https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/news/new-primary-care-data-added-from-medical-records-to-aid-vital-covid-19-research

25. UK Biobank Primary Care Linked Data. 2019.

26. Mao Q, Wang L, Tsang IW, Sun Y. Principal Graph and Structure Learning Based on Reversed Graph Embedding. IEEE Trans Pattern Anal Mach Intell. 2017 Nov 1;39(11):2227–41.

27. Core.Team R. R: A Language and Environment for Statistical Computing. www.R-project.org. R Foundation for Statistical Computing; 2020.

28. Zhou K, Donnelly LA, Morris AD, Franks PW, Jennison C, Palmer CNA, et al. Clinical

and genetic determinants of progression of type 2 diabetes: A direct study. Diabetes Care. 2014 Mar;37(3):718–24.

29. Zachariah S, Wykes W, Yorston D. Grading diabetic retinopathy (DR) using the Scottish grading protocol. Community Eye Heal J. 2015;28(92):72–3.

30. KDIGO. Clinical practice guideline for the evaluation and management of chronic kidney disease. Kidney Int Supplments. 2013;84(3):622–3.

31. Levey AS, Stevens LA, Schmid CH, Zhang Y, Castro AF, Feldman HI, et al. A new equation to estimate glomerular filtration rate. Ann Intern Med. 2009 May 5;150(9):604–12.

32. Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. J Am Stat Assoc. 1999 Jun;94(446):496.

33. Ord JK, Getis A. Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. Geogr Anal. 1995 Oct 1;27(4):286–306.

34. Tan VYF, Févotte C. Automatic relevance determination in nonnegative matrix factorization with the (β)-divergence. IEEE Trans Pattern Anal Mach Intell. 2013;35(7):1592–605.

Figure 1

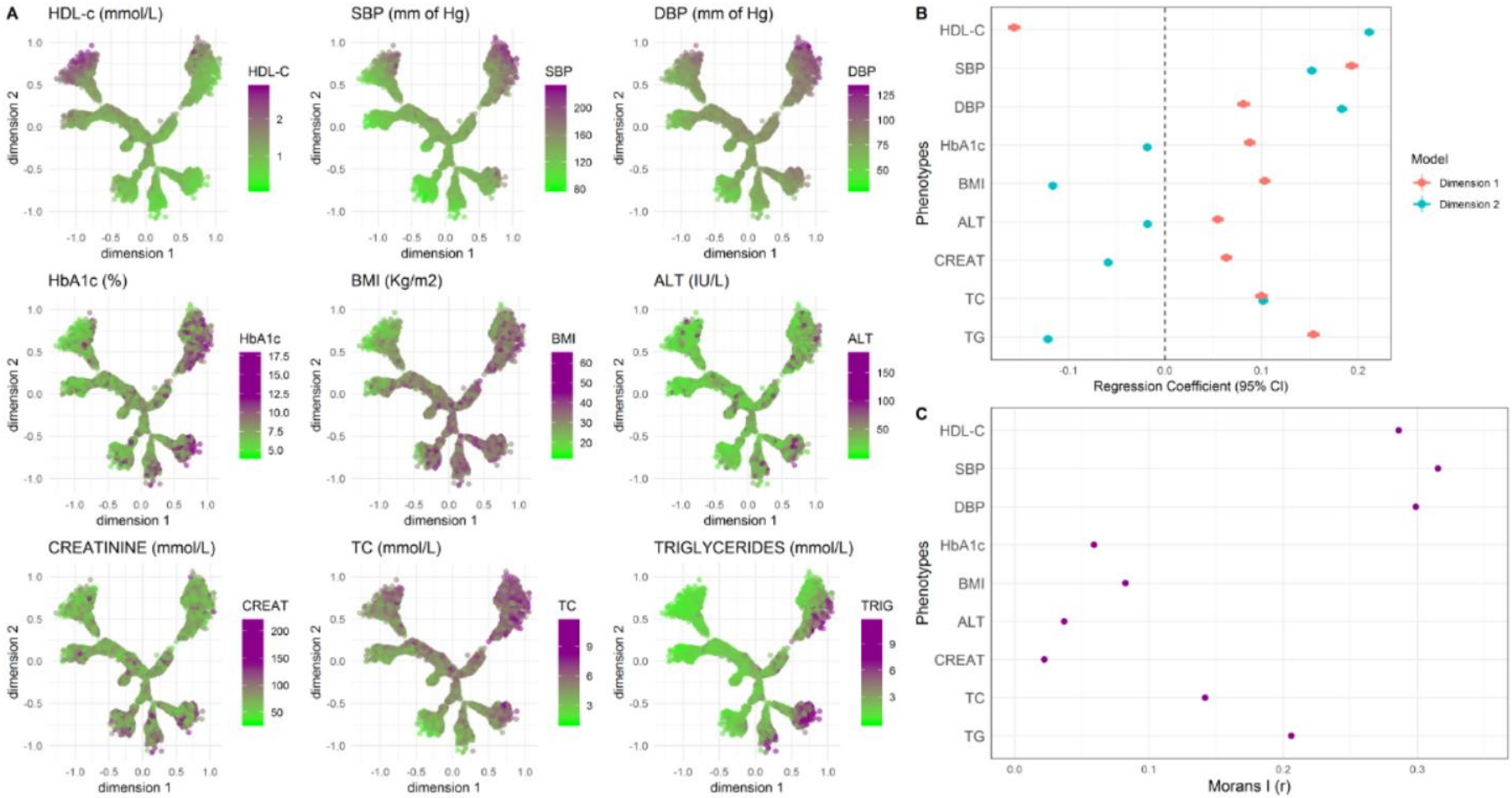

Fig 1| A visual representation of the phenotypic characteristics of 23,137 patients at diagnosis of T2D. **A.** DDRTree was used to reduce the 9 phenotypic variables (HbA1c, BMI, HDL-c, TC, TG, ALT, creatinine, and blood pressure) residualised for age and sex into a non-linear tree structure. The phenotype values are overlayed on the tree structure to visualise the distribution of nine phenotypes (HbA1c, BMI, HDL-c, TC, TG, ALT, creatinine, and SBP and DBP) over the reduced tree structure. Each point in the figure represents one individual. The magenta colour of the point indicates a higher value of the phenotypic variable for that individual and the green colour indicates lower values. **B.** Linear regression estimates (with 95% CI) between the DDRTree dimensions and the nine phenotypes showing the association between phenotypes and dimensions **C.** Spatial autocorrelation (n=23137) of the nine phenotypes; The Moran's I statistic is shown on the X-axis, with higher values representing phenotypes that are more strongly autocorrelated; all values were at p<0.0001.

Figure 2

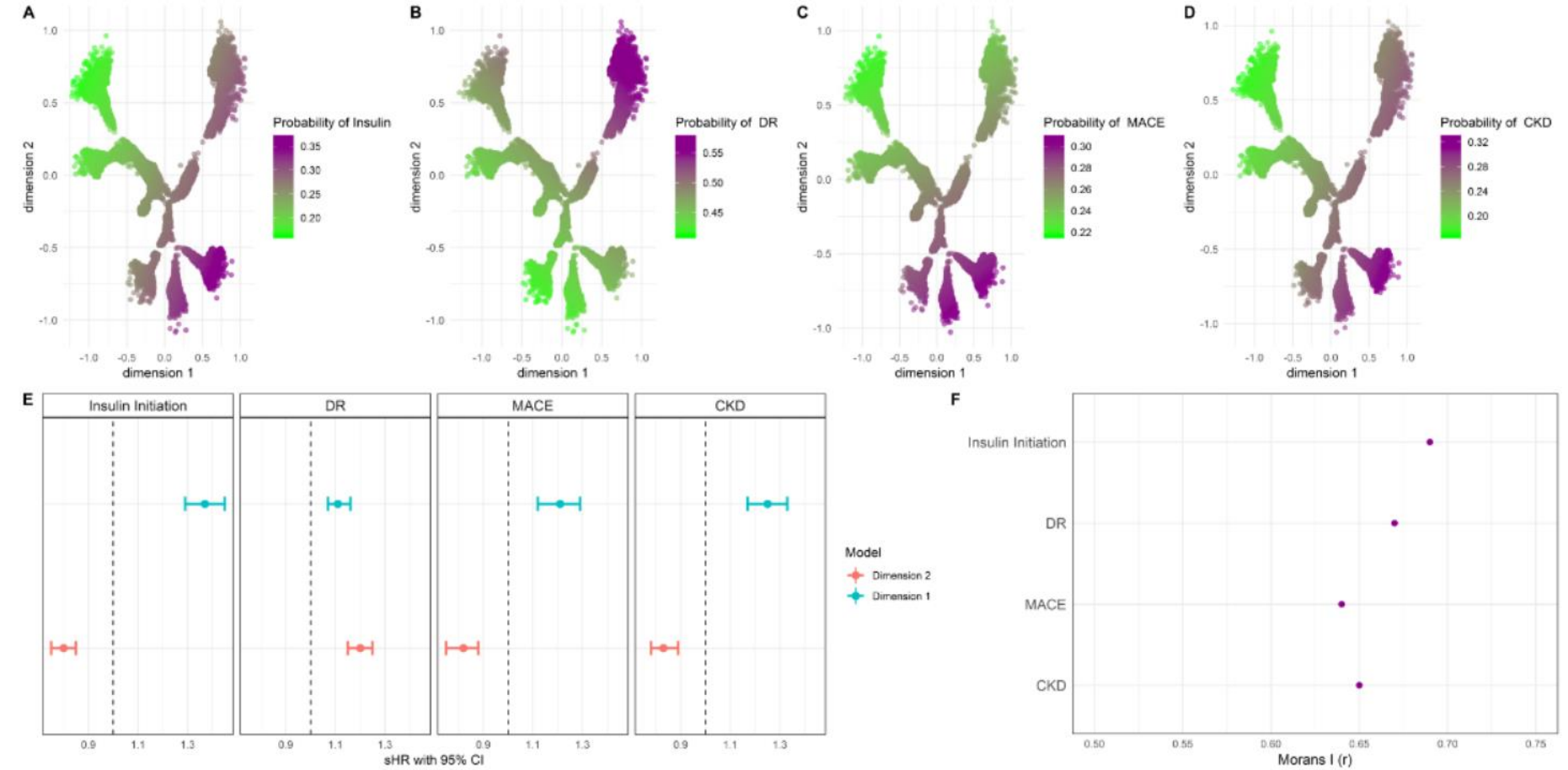

Fig 2| Visualising the heterogeneity in diabetes progression in Scottish patients with T2DM. All predictions are from models with DDRTree dimensions. **A.** Predicted probability of insulin initiation (use of insulin for more than 6 months or a clinical requirement for insulin, indicated as two or more HbA1c reading >=8.5% more than three months apart while taking two or more oral antihyperglycaemic agents) over 10 year period from the diagnosis of T2D. **B.** Probability of any incident diabetic retinopathy over a 10-year period. **C**. Probability of incident major adverse cardiac events (identified from SMR based on ICD 9 and ICD 10 codes) over a 10-year period. **D.** Probability of incident chronic kidney disease (eGFR<= 60 ml/min/1.73m^2 on at least 2 reading which were 90 days apart) over 10-year period. For all outcomes (A-D) probabilities were generated from a competing risk model constructed with DDRTree dimensions. **E.** Hazard ratios (95% CI) of DDRTree dimensions for each outcome from competing risk models **F**. Spatial autocorrelation of four diabetes progression outcomes The Moran's I statistic is shown on the X-axis, with higher values representing phenotypes that are more strongly autocorrelated; all values were at p<0.0001.

Figure 3

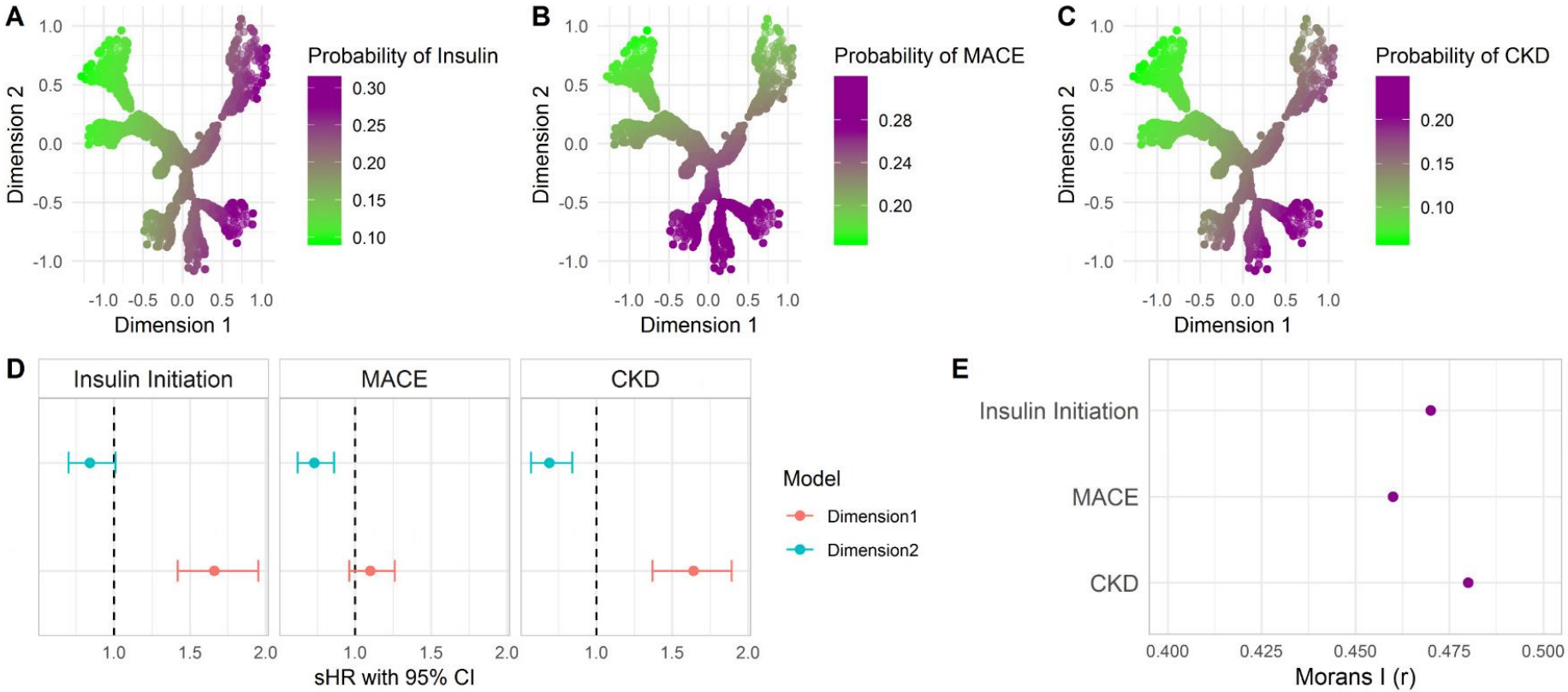

Fig 3| Visualising the heterogeneity in diabetes progression in UKBB data. All predictions are from models with DDRTree dimensions. **A.** Predicted probability of insulin initiation (use of insulin for more than 6 months or a clinical requirement for insulin, indicated as two or more HbA1c reading >=8.5% more than three months apart while taking two or more oral antihyperglycemic agents) over 10-year period from the diagnosis of T2D. **B.** Probability of incident major adverse cardiac events (identified from SMR based on ICD 9 and ICD 10 codes) over a 10-year period. **C**. Probability of incident chronic kidney disease (eGFR<= 60 ml/min/1.73m^2 on at least 2 reading which were 90 days apart) over 10-year period. For all outcomes (A-C) probabilities were generated from a competing risk model constructed with DDRTree dimensions**. D.**sub Hazard ratios (95% CI) of DDRTree dimensions for each outcome from competing risk models. **E.** Spatial autocorrelation of three diabetes progression outcomes. The Moran's I statistic is shown on the X-axis, with higher values representing phenotypes that are more strongly autocorrelated; all values were at p<0.0001.

Figure 4

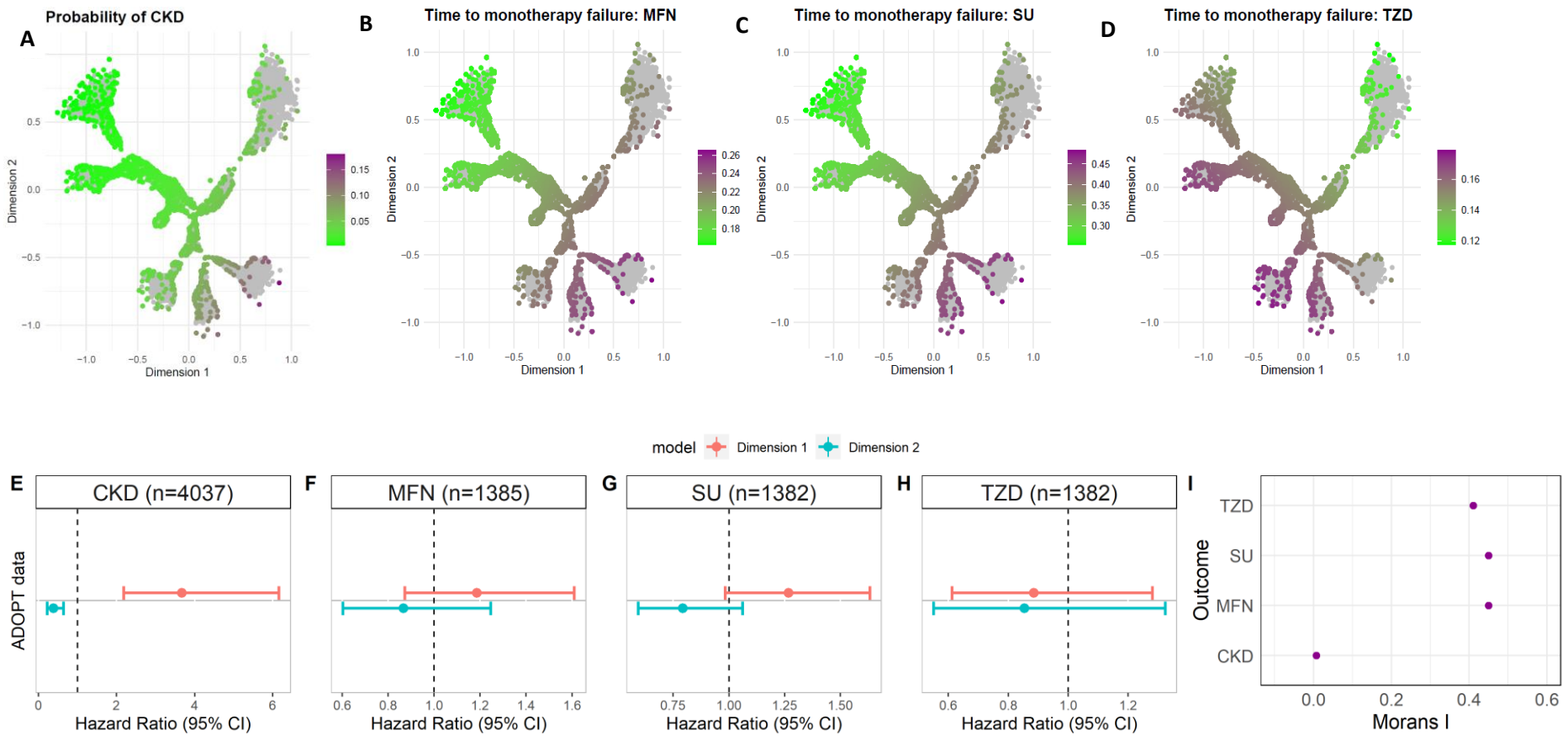

Fig 4| Visualising the heterogeneity in antidiabetic drug response using ADOPT trial data. **A**. Probability of incident chronic kidney disease (eGFR<= 60 ml/min/1.73m^2 on at least 2 reading which were 90 days apart) over 5-year period **B.** Predicted probability of metformin drug failure indicated by Fasting Plasma Glucose (FPG) >180 mg/dl (>10mmol/L) derived from cox proportional hazard model with DDRTree dimensions as covariates over period 5 years. **C**. Predicted probability of sulphonyl urea drug failure indicated by Fasting Plasma Glucose (FPG) >180 mg/dl (>10mmol/L) derived from cox proportional hazard model with DDRTree dimensions as covariates over period 5 years. **D**. Predicted probability of thiazolidinedione drug failure indicated by Fasting Plasma Glucose (FPG) >180 mg/dl (>10mmol/L) derived from cox proportional hazard model with DDRTree dimensions as covariates over period 5 years. **E.** Hazard ratio with 95% CI of DDRTree dimensions (Dimesnion1 and Dimension 2) for CKD, **F.**Hazard ratio with 95% CI of DDRTree dimensions (Dimesnion1 and Dimension 2) for Metformin drug failure, **G.** Hazard ratio with 95% CI of DDRTree dimensions (Dimesnion1 and Dimension 2) for sulphonyl urea drug failure. **H.** Hazard ratio with 95% CI of DDRTree dimensions (Dimesnion1 and Dimension 2) for thiazolidinedione drug failure from cox proportional hazard model**.I.** Spatial autocorrelation of CKD and drug failure probabilities. The Moran's I statistic is shown on the X-axis, with higher values representing phenotypes that are more strongly autocorrelated; all values were at p<0.0001.

Figure 5

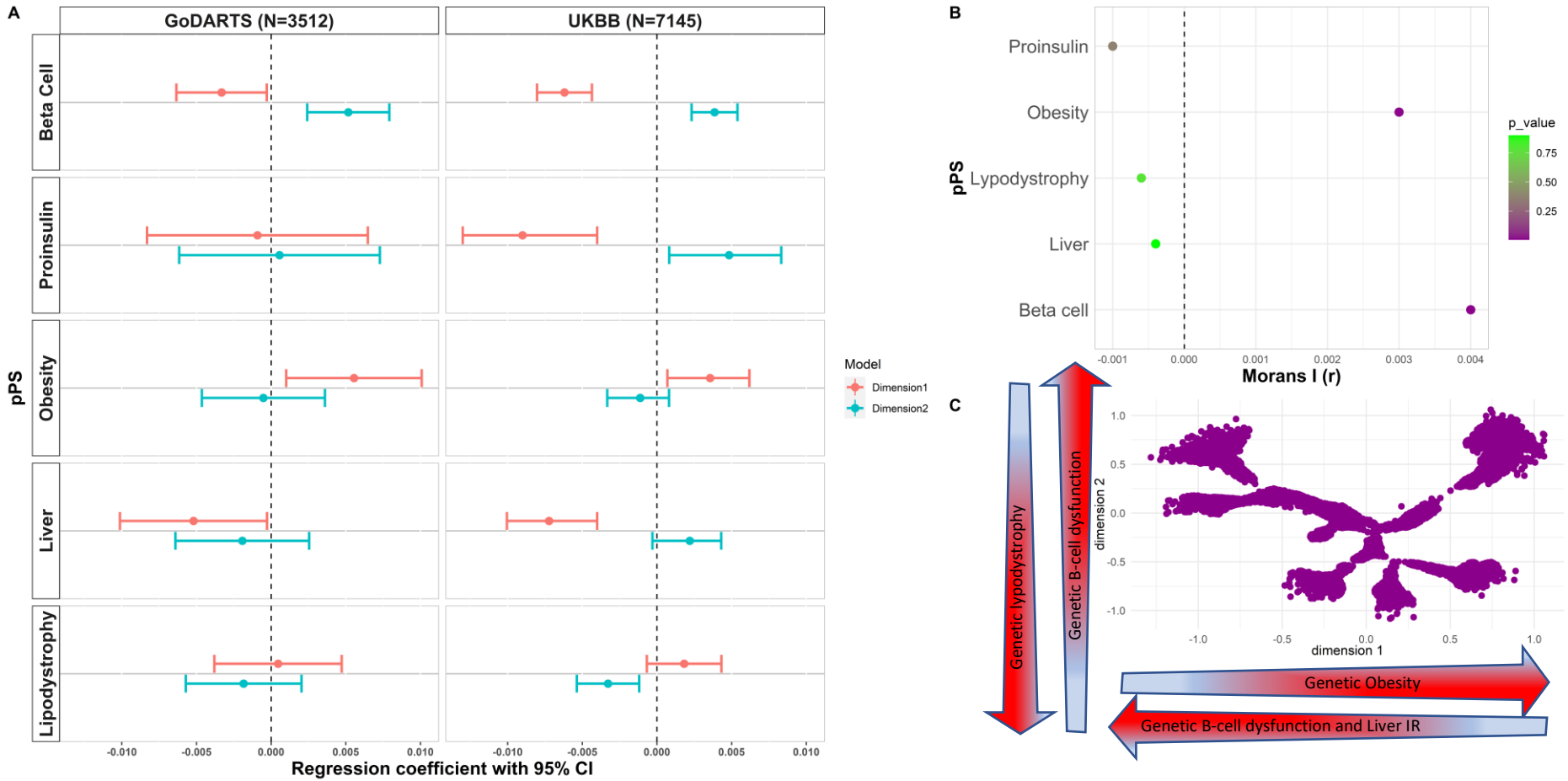

Fig 5| Distibrution of T2D partitioned polygenic scores across the phenotypic tree. **A**. Association between five pPS and dimension 1 and dimension 2 in GoDARTS (n=3512) and UKBB (n=7145). **B**. Spatial autocorrelation in GoDARTS (n=3512) of five pPS. X-axis shows Moran's I and Y axis the five pPS with green colour indicating lower p-value and magenta with higher values of p values. **C.** A schematic showing how genetically determined obesity (increases on X-axis), beta cell dysfunction and liver/lipid mediated insulin resistance (decreases on X axis), and beta cell dysfunction and lipid/liver mediated insulin resistance (increases on the Y-axis) are distributed.

Ethics declarations

Data provision and linkage was carried by the University of Dundee Health Informatics Centre (HIC) with analysis of anonymized data performed in an ISO27001 and Scottish Government accredited secure safehaven. HIC Standard Operating Procedures have been reviewed and approved by the NHS East of Scotland Research Ethics Service and consent for this study was obtained from the NHS Fife Caldicott Guardian. This population-level database contains longitudinal data including demographical, clinical and biochemistry linked by a unique identifier (PROCHI) from SCI-Diabetes (electronic repository with details of all patients with diabetes in Scotland).

Contributions

ERP, AWA, CB and ATN conceived and designed the study. ALR, SH, LD, JMD and ATN were involved in data preparation and analysis of the data. ERP, LD and ATN interpreted the results and wrote the manuscript. AD, SG, AWA, RMA, VM, CNAP, RMC, ASFD, JMD, ATH and CB provided critical inputs to the revision of the manuscript.

Acknowledgements

We thank all personnel at Health Informatics Centre (HIC) for linking different data sets, maintaining all statistical packages, and providing the data.

Grant Support

The research was supported by the National Institute for Health Research using Official Development Assistance (ODA) funding [INSPIRED 16/136/102] and Health Data Research UK which receives its funding from HDR UK Ltd (HDR-5012) funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation (BHF) and the Wellcome Trust. (The views expressed in this publication are those of the author(s) and not necessarily those of the NIHR or the UK Department of Health and Social Care). ERP holds Wellcome Trust New Investigator Award (102820/Z/13/Z).