Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables

Emma Ahlqvist, Petter Storm, Annemari Käräjämäki, Mats Martinell*, Mozhgan Dorkhan, Annelie Carlsson, Petter Vikman, Rashmi B Prasad, Dina Mansour Aly, Peter Almgren, Ylva Wessman, Nael Shaat, Peter Spégel, Hindrik Mulder, Eero Lindholm, Olle Melander, Ola Hansson, Ulf Malmqvist, Åke Lernmark, Kaj Lahti, Tom Forsén, Tiinamaija Tuomi, Anders H Rosengren, Leif Groop*

Summary

Background **Diabetes is presently classified into two main forms, type 1 and type 2 diabetes, but type 2 diabetes in particular is highly heterogeneous. A refined classification could provide a powerful tool to individualise treatment regimens and identify individuals with increased risk of complications at diagnosis.**

Methods **We did data-driven cluster analysis (k-means and hierarchical clustering) in patients with newly diagnosed diabetes (n=8980) from the Swedish All New Diabetics in Scania cohort. Clusters were based on six variables (glutamate decarboxylase antibodies, age at diagnosis, BMI, HbA_{1c}, and homoeostatic model assessment 2 estimates of β-cell function and insulin resistance), and were related to prospective data from patient records on development of complications and prescription of medication. Replication was done in three independent cohorts: the Scania Diabetes Registry (n=1466), All New Diabetics in Uppsala (n=844), and Diabetes Registry Vaasa (n=3485). Cox regression and logistic regression were used to compare time to medication, time to reaching the treatment goal, and risk of diabetic complications and genetic associations.**

Findings **We identified five replicable clusters of patients with diabetes, which had significantly different patient characteristics and risk of diabetic complications. In particular, individuals in cluster 3 (most resistant to insulin) had significantly higher risk of diabetic kidney disease than individuals in clusters 4 and 5, but had been prescribed similar diabetes treatment. Cluster 2 (insulin deficient) had the highest risk of retinopathy. In support of the clustering, genetic associations in the clusters differed from those seen in traditional type 2 diabetes.**

Interpretation **We stratified patients into five subgroups with differing disease progression and risk of diabetic complications. This new substratification might eventually help to tailor and target early treatment to patients who would benefit most, thereby representing a first step towards precision medicine in diabetes.**

Funding **Swedish Research Council, European Research Council, Vinnova, Academy of Finland, Novo Nordisk Foundation, Scania University Hospital, Sigrid Juselius Foundation, Innovative Medicines Initiative 2 Joint Undertaking, Vasa Hospital district, Jakobstadsnejden Heart Foundation, Folkhälsan Research Foundation, Ollqvist Foundation, and Swedish Foundation for Strategic Research.**

Introduction

Diabetes is the fastest increasing disease worldwide and a substantial threat to human health.[1] Existing treatment strategies have been unable to stop the progressive course of the disease and prevent development of chronic diabetic complications. One explanation for these shortcomings is that diagnosis of diabetes is based on measurement of only one metabolite, glucose, but the disease is heterogeneous with regard to clinical presentation and progression.

Diabetes classification into type 1 and type 2 diabetes relies primarily on the presence (type 1 diabetes) or absence (type 2 diabetes) of autoantibodies against pancreatic islet β-cell antigens and age at diagnosis (younger for type 1 diabetes). With this approach, 75–85% of patients are classified as having type 2 diabetes. A third subgroup, latent autoimmune diabetes in adults (LADA; affecting <10% of people with diabetes), defined by the presence of glutamic acid decarboxylase antibodies (GADA), is phenotypically indistinguishable from type 2 diabetes at diagnosis, but becomes increasingly similar to type 1 diabetes over time.[2] With the introduction of gene sequencing in clinical diagnostics, several rare monogenic forms of diabetes were described, including maturity-onset diabetes of the young and neonatal diabetes.[3,4]

Existing treatment guidelines are limited by the fact they respond to poor metabolic control when it has developed, but do not have means to predict which patients will need intensified treatment. Evidence suggests that early treatment is crucial for prevention of life-shortening complications because target tissues seem to remember poor metabolic control decades later (so-called metabolic memory).[5,6]

A refined classification could provide a powerful tool to identify at diagnosis those at greatest risk of complications and enable individualised treatment regimens in the same way as genetic diagnosis of monogenic diabetes guides clinicians to optimal treatment.[7] With this aim, we present a novel diabetes classification based on unsupervised, data-driven cluster analysis of six commonly measured variables and compare it metabolically,

Lancet Diabetes Endocrinol 2018

Published Online
March 1, 2018
http://dx.doi.org/10.1016/S2213-8587(18)30051-2

See Online/Comment
http://dx.doi.org/10.1016/S2213-8587(18)30070-6

*Contributed equally

Lund University Diabetes Centre, Department of Clinical Sciences, Lund University, Skåne University Hospital, Malmö, Sweden (E Ahlqvist PhD, P Storm PhD, M Dorkhan PhD, P Vikman PhD, R B Prasad PhD, D M Aly MSc, P Almgren MSc, Y Wessman MSc, N Shaat PhD, P Spégel PhD, Prof H Mulder PhD, E Lindholm PhD, Prof O Melander PhD, O Hansson PhD, Prof Å Lernmark PhD, A H Rosengren PhD, Prof L Groop PhD)**; Department of Primary Health Care, Vaasa Central Hospital, Vaasa, Finland** (A Käräjämäki MD, K Lahti MD)**; Diabetes Center, Vaasa Health Care Center, Vaasa, Finland** (A Käräjämäki, K Lahti)**; Department of Public Health and Caring Sciences, Uppsala University, Uppsala, Sweden** (M Martinell MD)**; Lund University Diabetes Centre, Department of Clinical Sciences, Skåne University Hospital** (A Carlsson PhD)**, and Department of Chemistry, Centre for Analysis and Synthesis** (P Spégel)**, Lund University, Lund, Sweden; Clinical Research and Trial Center, Lund University Hospital, Sweden** (U Malmqvist PhD)**; Folkhälsan Research Center, Helsinki, Finland** (T Forsén PhD, T Tuomi PhD)**; Abdominal Center, Endocrinology, Helsinki University Central Hospital, Research Program for Diabetes and Obesity** (T Tuomi)**, and Finnish Institute for Molecular Medicine** (Prof L Groop, T Tuomi)**, University of**

Helsinki, Helsinki, Finland; and Department of Neuroscience and Physiology, Wallenberg Center for Molecular and Translational Medicine, University of Gothenburg, Gothenburg, Sweden
(A H Rosengren)

Correspondence to:
Prof Leif Groop, Lund University Diabetes Centre, Malmö 21428, Sweden
leif.groop@med.lu.se

Research in context

Evidence before this study

National guidelines maintain information about diabetes classification, but this classification has not been much updated during the past 20 years, and very few attempts have been made to explore heterogeneity of type 2 diabetes. We searched PubMed up to Jan 1, 2017, using the Medical Subject Heading terms "diabetes mellitus", "type 2", and "classification". We identified several calls from expert groups for a revised classification, but few efforts to subgroup type 2 diabetes, none of which have been implemented in the clinic.

Added value of this study

In this study, a data-driven cluster analysis of six simple variables measured at diagnosis in adult patients with newly diagnosed diabetes (n=14 755) identified five replicable clusters of patients with significantly different characteristics and risk of diabetic complications. These included a cluster of very insulin-resistant individuals with significantly higher risk of diabetic kidney disease than the other clusters, a cluster of relatively young insulin-deficient individuals with poor metabolic control (high HbA_{1c}), and a large group of elderly patients with the most benign disease course.

Implications of all the available evidence

This new substratification could change the way we think about type 2 diabetes and help to tailor and target early treatment to patients who would benefit most, thereby representing a first step towards precision medicine in diabetes.

genetically, and clinically to the current classification in four separate populations from Sweden and Finland.

Methods

Study populations

We used data from five cohorts: All New Diabetics in Scania (ANDIS), the Scania Diabetes Registry (SDR), All New Diabetics in Uppsala (ANDIU), Diabetes Registry Vaasa (DIREVA), and Malmö Diet and Cancer CardioVascular Arm (MDC-CVA).

For more on the **ANDIS project** see http://andis.ludc.med.lu.se

The ANDIS project aims to recruit all incident cases of diabetes within Scania County in Sweden (about 1 200 000 inhabitants). All health-care providers in Scania were invited; the current registration covered the period from Jan 1, 2008, to Nov 3, 2016, during which 177 clinics registered 14 625 patients (>90% of eligible patients) aged 0–96 years within a median of 40 days (IQR 12–99) after diagnosis. Median follow-up for this cohort was 4·01 years (IQR 2·02–6·00).

Between 1996 and 2009, SDR recruited more than 7400 individuals with diabetes of all types from Scania County, 1466 of whom were recruited within 2 years after diagnosis and had all data necessary for clustering.[8] Median follow-up for this cohort was 11·05 years (IQR 8·33–14·56).

For more on the **ANDIU project** see http://www.andiu.se

Of the remaining three cohorts, ANDIU is a project similar to ANDIS in the Uppsala region (about 300 000 inhabitants) in Sweden and provided complete data on all clustering variables for 844 patients; DIREVA is from western Finland (roughly 170 000 inhabitants) and includes 5107 individuals with diabetes recruited from 2009 to 2014; and MDC-CVA includes 3300 individuals randomly selected from the larger Malmö Diet and Cancer study, to which all men and women born between 1923 and 1950 from the city of Malmö, southern Sweden, were invited to participate.[9]

The ANDIS and SDR study protocols were approved by the regional ethics review committee in Lund (ANDIS: 584/2006 and 2012/676; SDR: LU 35-99), DIREVA was approved by the ethics committee in Vasa (6/2007), and ANDIU was approved by the regional ethics review committee in Uppsala (2011/155). All participants gave written informed consent.

Measurements

In ANDIS, blood samples were drawn at registration, and fasting plasma glucose was analysed after overnight fasting with the HemoCue Glucose System (HemoCue AB, Ängelholm, Sweden). C-peptide concentrations were measured with an electro-chemiluminescence immunoassay on Cobas e411 (Roche Diagnostics, Mannheim, Germany) or a radioimmunoassay (Human C-peptide RIA; Linco, St Charles, MO, USA; or Peninsula Laboratories, Belmont, CA, USA). In ANDIS and SDR, GADA was measured with an ELISA (reference <11 U/mL[10]) or with radiobinding assays using ^{35}S-labelled protein[11] (positive cutoff: 5 relative units or 32 IU/mL). The radiobinding assays had 62–88% sensitivity and 91–99% specificity, and the ELISA assay had 72% sensitivity and 99% specificity (Combinatorial Autoantibody or Diabetes/Islet Autoantibody Standardization Programs 1998–2013). In ANDIU, GADA was measured at Laboratory Medicine in Uppsala (ref <5 U/mL). In DIREVA, GADA was measured with an ELISA (RSR, Cardiff, UK; positive cutoff 10 IU/mL). Zinc transporter 8 autoantibodies (ZnT8A) were measured with a radiobinding assay, as previously described.[12] HbA_{1c} was measured at diagnosis with the Variant II Turbo HbA1c Kit 2.0 (Bio-Rad Laboratories, Copenhagen, Denmark). Measurements of HbA_{1c}, alanine aminotransferase, ketones, and serum creatinine over time were obtained from the Clinical Chemistry database.

Genotyping

Genotyping of ANDIS participants was done on frozen DNA samples prepared from blood with Gentra Puregene Blood Kits (Qiagen, Hilden, Germany) using iPlex (Sequenom, San Diego, CA, USA) or TaqMan

assays (Thermo Fisher Scientific, Carlsbad, CA, USA) at the Clinical Research Center in Malmö, Sweden. In ANDIS, 5625 of the clustered individuals were genotyped, of whom 1714 were excluded because of non-Swedish origin and 164 were excluded because they had a call rate of less than 90%. MDC-CVA samples (controls) were genotyped at the Broad genotyping facility with the Infinium OmniExpressExome-8 version 1.0 BeadChip array (Illumina, San Diego, CA, USA). Quality control was done as previously described.[13] All single-nucleotide polymorphisms (SNPs) were in Hardy–Weinberg equilibrium in the controls.

Definitions of diabetes and diabetic complications

Type 1 diabetes was defined as GADA positive and C-peptide concentrations of less than 0·3 nmol/L. LADA was defined as GADA positive and C-peptide concentrations of 0·3 nmol/L or higher.

Estimated glomerular filtration rate (eGFR) was calculated with the Modification of Diet in Renal Disease formula.[14] Chronic kidney disease was defined as an eGFR of less than 60 (stage 3A) or less than 45 (stage 3B) for more than 90 days (onset of chronic kidney disease was set as the start of this period). End-stage renal disease was defined as at least one eGFR below 15 mL/min per 1·73 m².

Macroalbuminuria was defined as at least two of three consecutive visits with an albumin excretion rate of 200 µg/min or higher, an albumin excretion rate of 300 mg per day or higher, or an albumin to creatinine ratio of 25 mg/mmol or higher for men and 35 mg/mmol or higher for women.

Diabetic retinopathy was diagnosed by an ophthalmologist on the basis of fundus photographs.[15] Coronary events were defined by International Classification of Diseases (ICD)-10 codes I20-21, I24, I251, and I253-I259. Stroke was defined by ICD-10 codes I60-I61 and I63-I64. Individuals with known previous events were excluded.

Cluster analysis

Model variables were selected on the premise that patients develop diabetes when they can no longer increase their insulin secretion (whatever the reason) to meet the increased demands imposed by obesity and insulin resistance, and because they were easily obtainable from different clinical settings without interpretation and included the minimum number of laboratory tests. We chose BMI, age at onset of diabetes, and homoeostasis model assessment (HOMA) 2 estimates of β-cell function (HOMA2-B) and insulin resistance (HOMA2-IR) based on C-peptide concentrations (which performs better than insulin in patients with diabetes) calculated with the HOMA calculator (University of Oxford, Oxford, UK).[16] Presence or absence of GADA was included as a binary variable. Cluster analysis was done on values centred to a mean value of 0 and an SD of 1. In ANDIS, men and women were clustered separately to avoid stratification due to sex-dependent differences in the cluster variables and to provide separate cohorts for validation of results. Patients with secondary diabetes (n=162) and extreme outliers (>5 SDs from the mean; n=42) were excluded. TwoStep clustering, in which the first step estimates the optimal number of clusters on the basis of silhouette width and

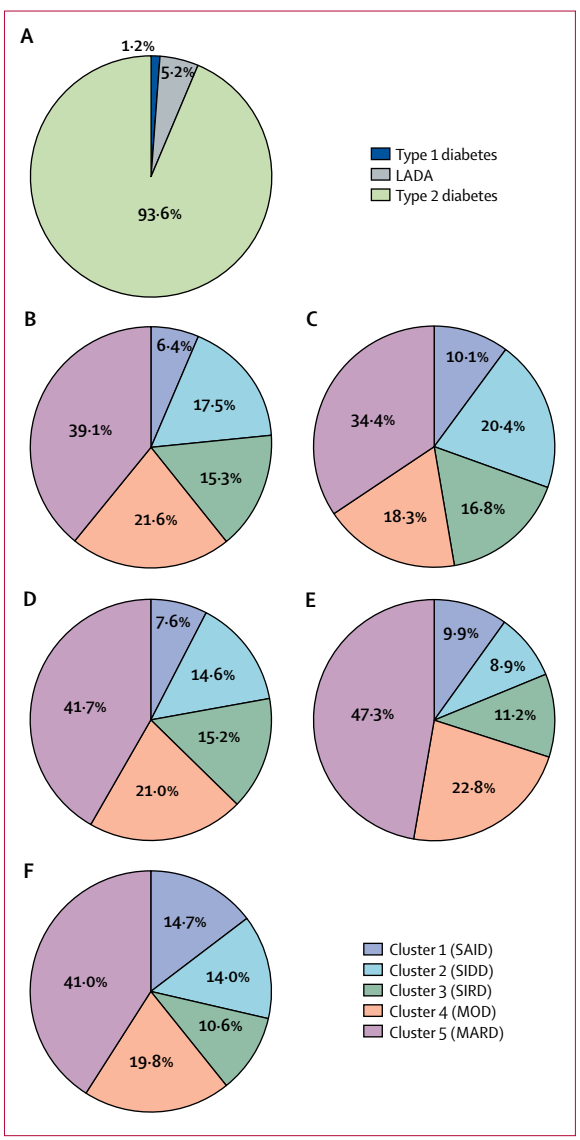

Figure 1: **Patient distribution according to method of classification**
(A) Distribution of ANDIS patients (n=8980) according to traditional classification. (B) Distribution of ANDIS patients (n=8980) according to k-means clustering. (C) Distribution of patients in the Scania Diabetes Registry (n=1466) according to k-means clustering. (D) Distribution of patients in the All New Diabetics in Uppsala cohort (n=844) according to k-means clustering. (E) Distribution of DIREVA patients with newly diagnosed diabetes (n=878) according to k-means clustering. (F) Distribution of DIREVA patients with longer-term diabetes (n=2607) according to k-means clustering. LADA=latent autoimmune diabetes in adults. SAID=severe autoimmune diabetes. SIDD=severe insulin-deficient diabetes. SIRD=severe insulin-resistant diabetes. MOD=mild obesity-related diabetes. MARD=mild age-related diabetes. ANDIS=All New Diabetics in Scania. DIREVA=Diabetes Registry Vaasa.

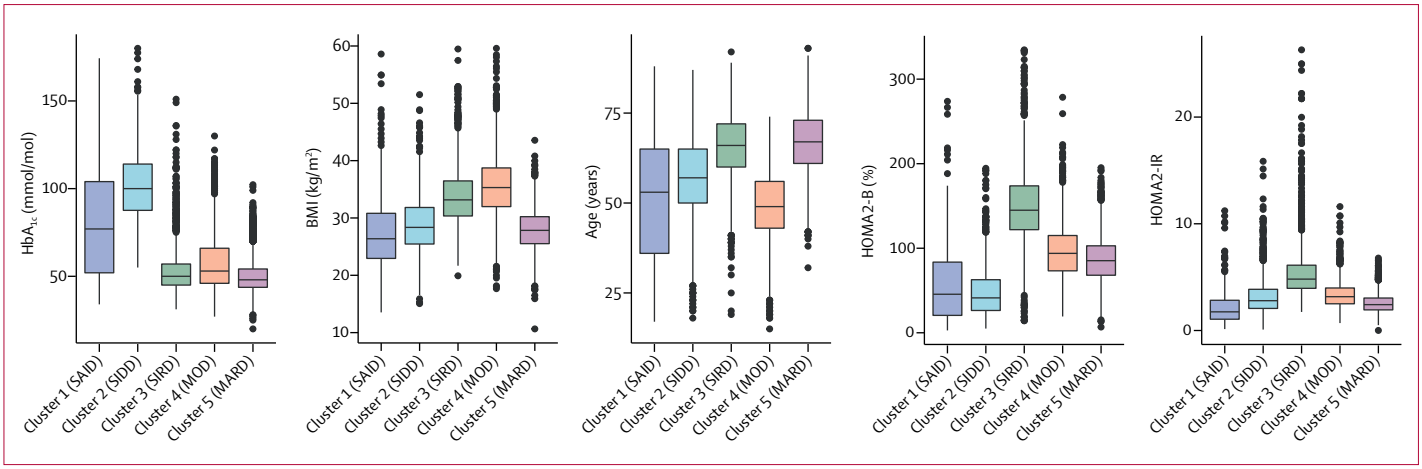

Figure 2: **Cluster characteristics in the ANDIS cohort**
Distributions of HbA_{1c} and age at diagnosis, and BMI, HOMA2-B, and HOMA2-IR at registration, in the ANDIS cohort for each cluster. k-means clustering was done separately for men and women; pooled data are shown here for clusters 2–5. SAID=severe autoimmune diabetes. SIDD=severe insulin-deficient diabetes. SIRD=severe insulin-resistant diabetes. MOD=mild obesity-related diabetes. MARD=mild age-related diabetes. HOMA2-B=homoeostatic model assessment 2 estimates of β-cell function. HOMA2-IR=homoeostatic model assessment 2 estimates of insulin resistance. ANDIS=All New Diabetics in Scania.

the second step does hierarchical clustering, was done in SPSS version 23 for two to 15 clusters using log-likelihood as a distance measure and Schwarz's Bayesian criterion for clustering. k-means clustering was done with a k value of 4 using the kmeansruns function (runs=100) in the fpc package in R version 3.3.1. Only individuals negative for GADA were included because the k-means method does not accommodate binary variables and all individuals who were GADA positive were clustered together with the TwoStep method. Cluster-centre coordinates in ANDIS are shown in the appendix.

See Online for appendix

Clusterwise stability was assessed through resampling the dataset 2000 times and computing the Jaccard similarities to the original cluster.[17] Generally, stable clusters should yield a Jaccard similarity of greater than 0·75.[17] Cluster labels were assigned by examining cluster variable means.

Statistical analysis

We calculated the risk of complications using Cox regression in SPSS version 23, including covariates. Post-hoc comparisons of effects across clusters were tested in Stata version 13.1.

Associations between clusters and genotypes were calculated with the maximum likelihood estimation method in SNPtest2 version 2.5.2.[18] A p value of less than 0·010 was considered significant in the genetic-association analyses. The equality of odds ratios (ORs) across strata was tested with seemingly unrelated estimation in Stata version 13.1. Bonferroni correction was used to determine significance for multiple tests. Genetic risk scores were calculated on the basis of the number of risk alleles weighed by their effect sizes reported in previous genome-wide association studies. Logistic regression was done for each cluster against the controls in SPSS version 23.

Role of the funding source

The funding sources had no part in study design, data collection, data analysis, data interpretation, or writing of the report. EA and LG had full access to all data and were responsible for the decision to submit for publication.

Results

We first analysed the ANDIS cohort, consisting of 14652 patients with newly diagnosed diabetes from Sweden, 932 (6·4%) of whom were registered before age 18 years and were not included in our analysis of adult diabetes. Of the 13720 adult patients, 204 (1·5%) had type 1 diabetes, 723 (5·3%) had LADA, 162 (1·2%) had secondary diabetes (coexisting pancreatic disease), and 519 (3·8%) were unclassifiable because of missing data. The remaining 12112 (88·3%) patients were considered to have type 2 diabetes (appendix).

To classify patients into novel diabetes subgroups, first we used the TwoStep clustering method in 8980 patients in the ANDIS cohort with complete data available for the clustering variables. The minimum silhouette width was found for five clusters in both men (n=5334) and women (n=3646) in the ANDIS cohort, and patient distributions and characteristics were similar in men and women (appendix). We verified the results using k-means clustering in GADA-negative patients, resulting in similar cluster distributions to TwoStep, with the same overall cluster characteristics in both sexes (figures 1, 2; appendix). Cluster stability was estimated as Jaccard means,[17] which were greater than 0·8 for all clusters, regardless of sex.

Cluster 1, including 577 (6·4%) of the 8980 clustered patients, was characterised by early-onset disease, relatively low BMI, poor metabolic control, insulin deficiency, and presence of GADA (appendix), and was labelled as severe autoimmune diabetes (SAID). Cluster 2, including 1575

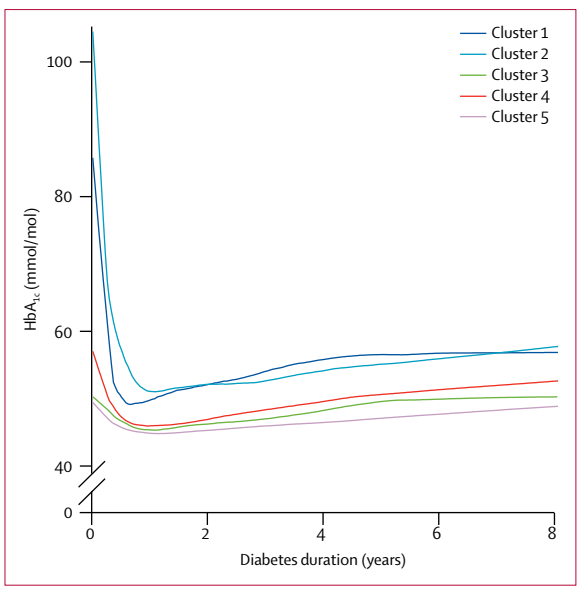

Figure 3: **Mean HbA_{1c} over time in the All New Diabetics in Scania cohort**

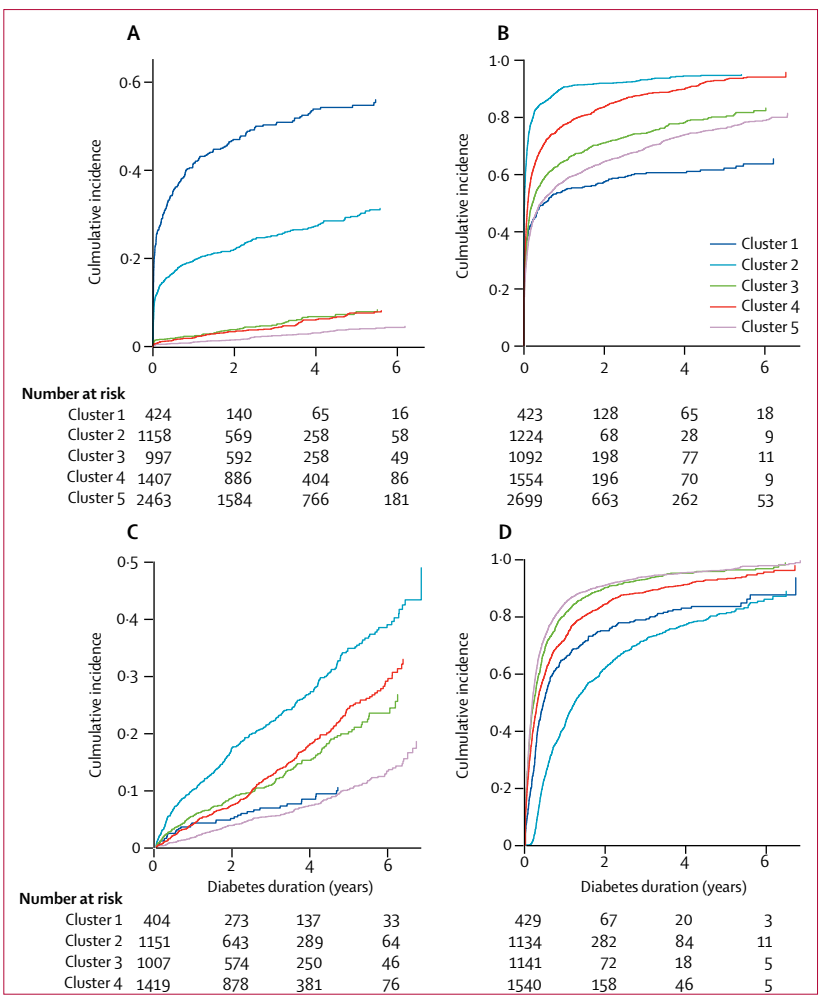

Figure 4: **Antidiabetic therapy in All New Diabetics in Scania cohort during follow-up**
(A) Time to sustained insulin use. (B) Time to metformin treatment. (C) Time to treatment with oral medication other than metformin. (D) Time to reach treatment goal (HbA_{1c}<6·9% [52 mmol/mol]).

(17·5%) patients and labelled as severe insulin-deficient diabetes (SIDD), was GADA negative but otherwise similar to cluster 1: low age at onset, relatively low BMI, low insulin secretion (low HOMA2-B index), and poor metabolic control. Cluster 3, labelled as severe insulin-resistant diabetes (SIRD) and including 1373 (15·3%) patients, was characterised by insulin resistance (high HOMA2-IR index) and high BMI. Cluster 4, including 1942 (21·6%) patients, was also characterised by obesity but not by insulin resistance, and was labelled as mild obesity-related diabetes (MOD). The 3513 (39·1%) patients in cluster 5 (labelled as mild age-related diabetes [MARD]) were older than patients in other clusters, but showed, similar to cluster 4, only modest metabolic derangements.

We used three independent cohorts to replicate the clustering: SDR (n=1466), ANDIU (n=844), and DIREVA (n=3485). In SDR, the optimal number of clusters was also estimated to be five, and k-means (k=4) and TwoStep clustering yielded similar results (92·4% clustered identically). Patient distributions and cluster characteristics were similar to ANDIS (figure 1; appendix). Jaccard bootstrap means were greater than 0·8 for all clusters. k-means clustering in ANDIU also replicated the results from ANDIS (figure 1; appendix). In the DIREVA cohort, we found that clustering gave similar results in 2607 patients with longer diabetes duration (mean 10·15 years [SD 10·34]) as in 878 patients with newly-diagnosed diabetes (diabetes duration <2 years; figure 1; appendix).

To be clinically useful, patients would need to be assigned to clusters without de-novo clustering of a full cohort. Therefore, we assigned patients in replication cohorts to clusters on the basis of which cluster they were most similar to, calculated as their Euclidian distance from the nearest cluster centre derived from ANDIS coordinates, and found similar distributions (appendix). Sensitivity and specificity were highest in ANDIU and DIREVA patients recruited soon after diagnosis (appendix), probably reflecting how and when clustering variables were obtained.

We then compared disease progression, treatment, and development of diabetic complications between clusters in ANDIS. Clusters 1 and 2 had substantially higher HbA_{1c} at diagnosis than the other clusters, a difference persisting throughout follow-up (figure 3). Ketoacidosis at diagnosis was most frequent in cluster 1 (31% [124/406]) and cluster 2 (25% [259/1033]; *vs* <5% in other clusters; appendix). HbA_{1c} was the strongest predictor of ketoacidosis at diagnosis (OR 2·73, 95% CI 2·47–3·03; p<0·0001, per 1 SD change; appendix). Cluster 3 had the highest prevalence of non-alcoholic fatty liver disease (appendix). ZnT8A autoantibodies were primarily seen in patients with SAID (27% [79/289] *vs* <2% in other clusters; appendix).

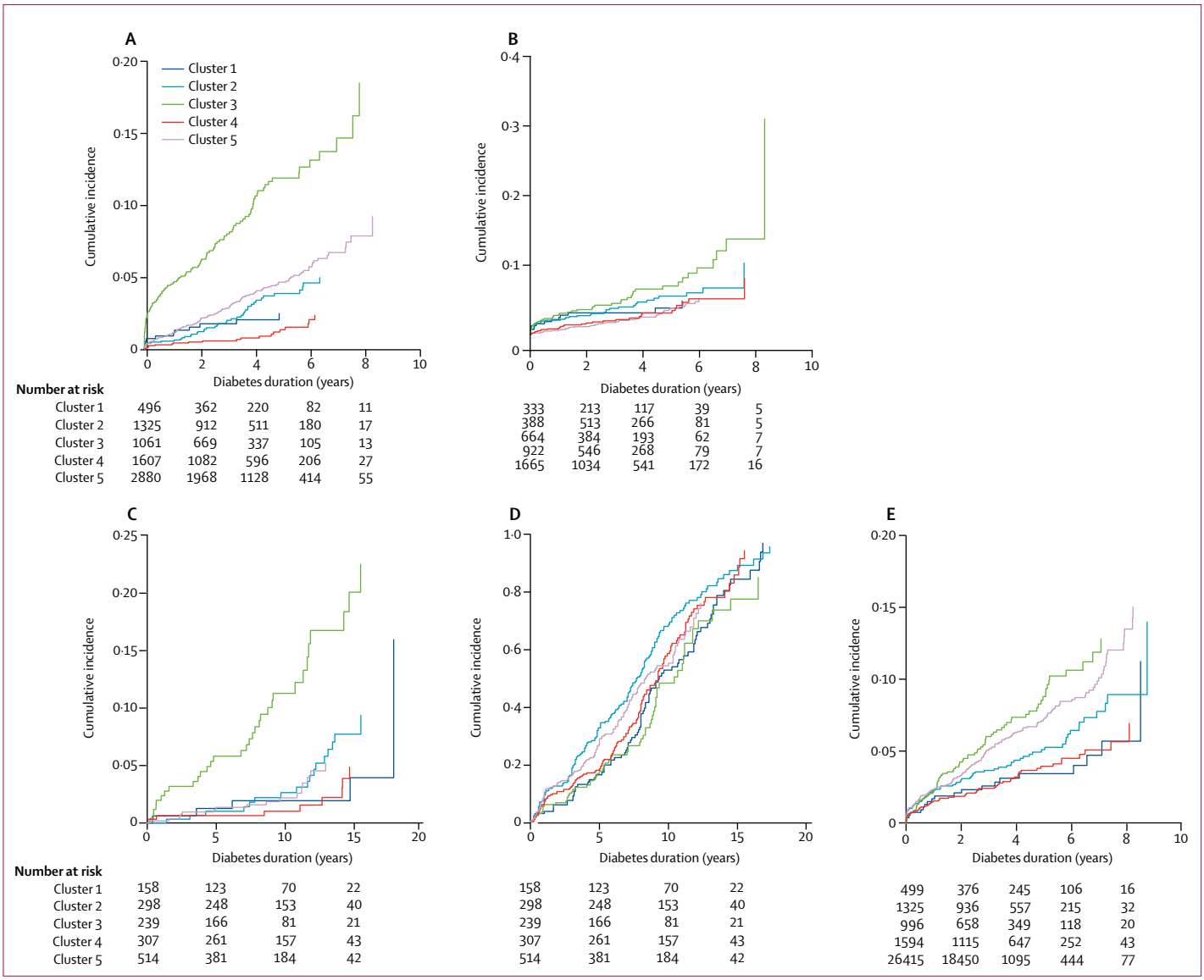

Figure 5: **Progression of disease over time by cluster**
(A) Time to chronic kidney disease (at least stage 3B) in the ANDIS cohort. (B) Time to macroalbuminuria in the ANDIS cohort. (C) Time to end-stage renal disease in the SDR cohort (data presented for SDR rather than ANDIS because of availability of longer-term follow-up). (D) Time to at least mild non-proliferative or proliferative diabetic retinopathy in the SDR cohort (insufficient data for retinopathy available in ANDIS). (E) Time to coronary events in the ANDIS cohort. Kidney function was not tested at diagnosis and, therefore, onset was set to the first screening date; it is not known how many patients were already affected at diagnosis. ANDIS=All New Diabetics in Scania. SDR=Scania Diabetes Registry.

At registration, insulin had been prescribed to 212 (42%) of 506 patients in cluster 1 and 389 (29%) of 1339 patients in cluster 2, but to less than 4% of patients in clusters 3–5 (appendix). Time to sustained insulin use was shortest in cluster 1 (hazard ratio [HR] 26·87, 95% CI 21·17–34·11, *vs* cluster 5; figure 4; appendix), followed by cluster 2 (10·97, 8·73–13·77, *vs* cluster 5). The proportion of patients on metformin was highest in cluster 2 and lowest in cluster 1 (figure 4; appendix), but was also low in cluster 3, which would be expected to benefit the most from metformin, showing that traditional classification is unable to tailor treatment to the underlying pathogenic defects. Kidney function and adverse reactions had no major effect on the proportions of patients taking metformin at this early stage of disease (appendix). Patients in cluster 2 had the shortest time to second oral diabetes treatment (figure 4; appendix) and the longest time to reach the treatment goal (HbA_{1c}<6·9% [52 mmol/mol]; figure 4).

In ANDIS, patients in cluster 3 had the highest risk of developing chronic kidney disease during mean follow-up of 3·9 years (SD 2·3; appendix). For stage 3A chronic kidney disease (eGFR <60 mL/min), the age-adjusted and sex-adjusted risk was more than two times

	EA/NEA	MAF	Cluster 1 (SAID; n=313)	Cluster 2 (SIDD; n=676)	Cluster 3 (SIRD; n=603)	Cluster 4 (MOD; n=727)	Cluster 5 (MARD; n=1646)	p value of difference among clusters 2–5
TCF7L2 (rs7903146)	T/C	0·26	1·17 (0·97–1·40); p=0·077	1·51 (1·33–1·71); p<0·0001	1·00 (0·87–1·15); p=0·86	1·38 (1·21–1·56); p<0·0001	1·41 (1·28–1·55); p<0·0001	<0·0001*
KCNQ1 (rs2237895)	C/T	0·41	1·08 (0·91–1·28); p=0·31	1·13 (1·00–1·28); p=0·052	0·85 (0·74–0·97); p=0·0272	0·98 (0·86–1·10); p=0·88	1·13 (1·03–1·23); p=0·0196	0·0008
HHEX/IDE (rs1111875)	G/A	0·41	1·16 (0·98–1·38); p=0·10	1·21 (1·07–1·37); p=0·0045	1·05 (0·92–1·19); p=0·51	0·94 (0·84–1·06); p=0·31	1·11 (1·02–1·22); p=0·0228	0·0106
IGF2BP2 (rs4402960)	T/G	0·29	1·04 (0·87–1·24); p=0·50	1·23 (1·08–1·40); p=0·0002	1·01 (0·88–1·16); p=0·53	1·04 (0·92–1·18); p=0·31	1·22 (1·11–1·33); p<0·0001	0·0117
CDKN2B (rs10811661)	T/C	0·16	0·87 (0·70–1·08); p=0·24	1·33 (1·11–1·59); p=0·0014	0·98 (0·83–1·17); p=0·85	0·99 (0·84–1·16); p=0·92	1·18 (1·04–1·33); p=0·0054	0·0149
MTNR1B (rs10830963)	G/C	0·29	0·84 (0·70–1·01); p=0·05	0·93 (0·82–1·07); p=0·26	0·89 (0·77–1·02); p=0·056	1·13 (1·00–1·28); p=0·067	1·05 (0·96–1·15); p=0·29	0·0151
SLC30A8 (rs13266634)	T/C	0·31	0·98 (0·82–1·17); p=0·78	0·93 (0·82–1·06); p=0·23	1·11 (0·97–1·27); p=0·11	1·07 (0·94–1·21); p=0·30	0·92 (0·83–1·01); p=0·0457	0·0160
MC4R (rs12970134)	G/A	0·27	0·95 (0·79–1·14); p=0·52	0·97 (0·85–1·11); p=0·55	0·99 (0·86–1·13); p=0·59	0·87 (0·77–0·99); p=0·0229	1·07 (0·97–1·18); p=0·18	0·0230
TM6SF2 (rs10401969)	T/C	0·10	0·75 (0·58–0·97); p=0·038	0·69 (0·58–0·83); p=0·0002	0·62 (0·52–0·75); p<0·0001	0·89 (0·73–1·07); p=0·26	0·77 (0·67–0·89); p=0·0005	0·0233
ADAMTS9-AS2 (rs4607103)	T/C	0·24	1·05 (0·87–1·27); p=0·54	0·89 (0·77–1·03); p=0·15	0·93 (0·80–1·08); p=0·42	1·12 (0·98–1·27); p=0·064	0·92 (0·83–1·01); p=0·13	0·0278
VPS13C (rs17271305)	G/A	0·40	1·00 (0·84–1·19); p=0·93	0·97 (0·86–1·10); p=0·84	1·11 (0·98–1·26); p=0·092	0·88 (0·78–0·99); p=0·0491	0·93 (0·85–1·02); p=0·17	0·0281
SLC2A2 (rs11920090)	T/A	0·13	0·94 (0·74–1·20); p=0·54	0·83 (0·70–0·99); p=0·0162	0·91 (0·76–1·09); p=0·23	0·97 (0·82–1·16); p=0·63	1·08 (0·95–1·24); p=0·44	0·0368
KCNJ11 (rs5219)	T/C	0·38	1·05 (0·88–1·25); p=0·61	1·18 (1·04–1·34); p=0·0121	1·03 (0·90–1·18); p=0·67	1·28 (1·13–1·44); p=0·0001	1·10 (1·01–1·21); p=0·0324	0·0453
TSPAN8 (rs7961581)	T/C	0·26	0·97 (0·80–1·17); p=0·69	1·05 (0·92–1·21); p=0·55	1·13 (0·98–1·31); p=0·11	0·99 (0·87–1·13); p=0·80	0·92 (0·84–1·02); p=0·11	0·0464

Maximum likelihood estimation using geographically matched individuals without diabetes as controls (n=2754). EA=effect allele. NEA=non-effect allele. MAF=minor allele frequency. SAID=severe autoimmune diabetes. SIDD=severe insulin-deficient diabetes. SIRD=severe insulin-resistant diabetes. MOD=mild obesity-related diabetes. MARD=mild age-related diabetes. ANDIS=All New Diabetics in Scania. *Significant after correction for multiple testing (77 tests).

Table: **Genetic associations with specific ANDIS clusters reaching at least nominal significance for difference among clusters 2–5**

higher than for patients in cluster 5 (HR 2·41, 95% CI 2·08–2·79; p<0·0001; appendix); for stage 3B chronic kidney disease (eGFR <45 mL/min), the adjusted risk was more than three times higher than for cluster 5 (3·34, 2·59–4·30; p<0·0001; figure 5A). Patients in cluster 3 also had higher risk of diabetic kidney disease, defined as persistent macroalbuminuria (2·89, 1·92–4·35; p<0·0001; figure 5B). Similarly, in the SDR cohort (follow-up 11·0 years [SD 4·4]), patients in cluster 3 had the highest risk of chronic kidney disease (appendix) and macroalbuminuria (2·18, 1·31–3·63; p=0·0026; appendix). Patients in cluster 3 in SDR had almost five times higher risk of end-stage renal disease than did patients in cluster 5 (4·89, 2·68–8·93; p<0·0001; figure 5C). The increased prevalence of kidney disease in cluster 3 was also confirmed in the DIREVA cohort (appendix).

Early signs of diabetic retinopathy (mean duration 135 days [SD 299]) were more common in cluster 2 than in the other clusters in ANDIS (OR 1·6, 95% CI 1·3–1·9; p<0·0001 *vs* cluster 5; appendix). The higher prevalence of retinopathy in cluster 2 than in other clusters was replicated in ANDIU (appendix) and SDR (HR 1·33, 95% CI 1·15–1·54; p=0·0001; figure 5D; appendix).

Although unadjusted risk of coronary events and stroke was lowest in clusters 1, 2, and 4, no significant difference was seen between the clusters in age-adjusted and sex-adjusted risk in ANDIS and SDR (figure 5E; appendix).

Finally, we analysed genetic loci previously associated with diabetes and related traits[19] (table). Each cluster in ANDIS was compared with a non-diabetic cohort (MDC-CVA) from the same geographical region.[9] No genetic variant was associated with all clusters (appendix). A variant in the *TCF7L2* gene (rs7903146), previously associated with type 2 diabetes,[20] was also associated with SIDD, MOD, and MARD, but not with SIRD (only difference significant after correction for multiple testing; table). The rs10401969 variant in the *TM6SF2* gene, previously associated with non-alcoholic fatty liver disease,[21] was associated with SIRD but not with MOD, suggesting that SIRD is characterised by more unhealthy (metabolic syndrome) obesity than MOD. The rs2854275

variant in the *HLA* locus (previously associated with type 1 diabetes[22]) was strongly associated with SAID (OR 2·05, 95% CI 1·69–2·56; p<0·0001), but not with SIDD (0·82, 0·66–1·00; p=0·078), reflecting the non-autoimmune nature of the SIDD cluster. A genetic risk score for type 2 diabetes (appendix) was significantly associated with all clusters (p<0·0008), except for cluster 3 (p=0·16). An insulin secretion risk score was significantly associated with MOD (p=0·0002) and MARD (p<0·0001), and nominally with SIDD (p=0·0143), but showed no evidence of association with SAID (p=0·59) or SIRD (p=0·65). We did not analyse genetic associations in cohorts other than ANDIS because of insufficient data.

Discussion

Taken together, the results of our study suggest that this new clustering of patients with adult-onset diabetes is superior to the classic diabetes classification because it identifies patients at high risk of diabetic complications at diagnosis and provides information about underlying disease mechanisms, thereby guiding choice of therapy. By contrast with previous attempts to dissect the heterogeneity of diabetes,[23] we used variables reflective of key aspects of diabetes that are monitored in patients. Thus, this clustering can easily be applied to both existing diabetes cohorts (eg, from drug trials) and patients in diabetes clinics. A web-based tool to assign patients to specific clusters, provided the appropriate variables have been measured, is under development.

Whereas SAID overlapped with type 1 diabetes and LADA, SIDD and SIRD represent two new, severe forms of diabetes previously masked within type 2 diabetes. It would be reasonable to target individuals in these clusters with intensified treatment to prevent diabetic complications. The risk of kidney complications was substantially increased in patients with SIRD, reinforcing the association between insulin resistance and kidney disease.[24] Insulin resistance has been associated with increased salt sensitivity, glomerular hypertension, hyperfiltration, and reduced renal function, all hallmarks of diabetic kidney disease.[25] The increased incidence of diabetic kidney disease in this study was in spite of reasonably low HbA_{1c}, suggesting that glucose-lowering therapy is not the optimum way of preventing this complication. In support of this hypothesis, mice with podocyte-specific knockout of the insulin receptor, mimicking the reduced insulin signaling seen in patients who are insulin resistant, developed diabetic kidney disease, even during normoglycaemic conditions.[26] Although differences in retinopathy were not as pronounced as for diabetic kidney disease, insulin deficiency or hyperglycaemia appeared to be important triggers of retinopathy, with the highest prevalence observed in cluster 2 (SIDD).

The fact that clustering led to similar results in newly diagnosed patients and patients with longer-term diabetes, and that C-peptide remained relatively stable over time (appendix), suggests that the clusters are stable and at least partially mechanistically distinct rather than representing different stages of the same disease. The differences in genetic associations also support this view. In particular, the absence of associations between the genetic risk scores for type 2 diabetes and insulin secretion and SIRD indicate that this group might have a different aetiology to the other clusters. Hepatic insulin resistance seems to be a feature of non-alcoholic fatty liver disease, because the SNP in the *TM6SF2* gene usually associated with non-alcoholic fatty liver disease was associated with SIRD in this study, but not with MOD.

We cannot at this stage claim that the new clusters represent different aetiologies of diabetes, nor that this clustering is the optimal classification of diabetes subtypes. Additionally, whether patients (particularly from the periphery of clusters) can move between clusters needs to be shown in future prospective studies, and the exact overlap of weaker association signals will need to be investigated in larger cohorts. It might be possible to refine the stratification further through inclusion of additional cluster variables, such as biomarkers, genotypes, or genetic risk scores. Future genome-wide association studies might also be able to better describe the genetic architecture of the different clusters and establish the inherited proportion of each cluster with heritability partitioning models.[27] This classification was derived primarily with patients from northern Europe, with limited non-Scandinavian representation, and the applicability of this strategy to patients of other ethnicities needs to be assessed. Only two types of autoantibodies were measured, and the effects of other antibodies on clustering performance are unknown. Moreover, we did not have data on some known risk factors for diabetic complications, such as blood pressure and blood lipids, and could therefore not include these in the analysis.

In conclusion, our data suggest that the combined information from a few variables central to the development of diabetes is superior to measurement of only one metabolite, glucose. Through combining this information from diagnosis with information in the health-care system, this study provides a first step towards a more precise, clinically useful stratification, representing an important step towards precision medicine in diabetes. This clustering also paves the way for randomised trials targeting insulin secretion in SIDD and insulin resistance in SIRD.

Contributors

EA, PSt, PV, TT, AHR, and LG contributed to the conception of the work. EA, PSt, AK, MM, MD, AC, PV, YW, NS, PSp, HM, EL, OM, OH, UM, ÅL, KL, TF, TT, AHR, and LG contributed to the data collection. EA, PSt, MM, RBP, DMA, and PA contributed to the data analysis. EA, PSt, AK, and LG drafted the article. All authors contributed to the interpretation of data and critical revision of the Article. All authors gave final approval of the version to be published.

Declaration of interests

We declare no competing interests.

Acknowledgments

This study was supported by grants from the Swedish Research Council (project grant 521-2010-3490 and infrastructure grants 2010-5983, 2012-5538, and 2014-6395 to LG; project grant 2017-02688 to EA; Linnaeus grant 349-2006-237; and a strategic research grant 2009-1039 to LG), a European Research Council Advanced Research grant (GA 269045), a Vinnova Swelife grant, and grants from the Academy of Finland (263401 and 267882 to LG), Sigrid Juselius Foundation, Novo Nordisk Foundation, and Scania University Hospital (ALF grant). This project has also received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreements 115974 (BEAt-DKD) and 115881 (RHAPSODY). This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and the European Federation of Pharmaceutical Industries and Associations. Furthermore, this project was financially supported by the Swedish Foundation for Strategic Research (IRC15-0067). DIREVA was supported by the Vasa Hospital district, Jakobstadsnejden Heart Foundation, Folkhalsan Research Foundation, and Ollqvist Foundation (to TT and AK). We thank all patients and health-care providers for their support and willingness to participate. We also thank Johan Hultman, Jasmina Kravic, Maria Fälemark, Christina Rosborn, Gabriella Gremsperger, Maria Sterner, Malin Neptin, Lisa Sundman, Paula Kokko, Carin Gustavsson, and Ulrika Blom-Nilsson for excellent technical and administrative support; Rita Jedlert and Region Skåne (Scania County); and the ANDIS steering committee for their support.

References

1 NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4·4 million participants. *Lancet* 2016; **387:** 1513–30.

2 Tuomi T, Groop LC, Zimmet PZ, Rowley MJ, Knowles W, Mackay IR. Antibodies to glutamic acid decarboxylase reveal latent autoimmune diabetes mellitus in adults with a non-insulin-dependent onset of disease. *Diabetes* 1993; **42:** 359–62.

3 Froguel P, Zouali H, Vionnet N, et al. Familial hyperglycemia due to mutations in glucokinase. Definition of a subtype of diabetes mellitus. *N Engl J Med* 1993; **328:** 697–702.

4 Yamagata K, Oda N, Kaisaki PJ, et al. Mutations in the hepatocyte nuclear factor-1alpha gene in maturity-onset diabetes of the young (MODY3). *Nature* 1996; **384:** 455–58.

5 Reddy MA, Zhang E, Natarajan R. Epigenetic mechanisms in diabetic complications and metabolic memory. *Diabetologia* 2015; **58:** 443–55.

6 Brownlee M. The pathobiology of diabetic complications: a unifying mechanism. *Diabetes* 2005; **54:** 1615–25.

7 Pearson ER, Flechtner I, Njolstad PR, et al. Switching from insulin to oral sulfonylureas in patients with diabetes due to Kir6.2 mutations. *N Engl J Med* 2006; **355:** 467–77.

8 Lindholm E, Agardh E, Tuomi T, Groop L, Agardh CD. Classifying diabetes according to the new WHO clinical stages. *Eur J Epidemiol* 2001; **17:** 983–89.

9 Manjer J, Carlsson S, Elmstahl S, et al. The Malmo Diet and Cancer study: representativity, cancer incidence and mortality in participants and non-participants. *Eur J Cancer Prev* 2001; **10:** 489–99.

10 Rahmati K, Lernmark A, Becker C, et al. A comparison of serum and EDTA plasma in the measurement of glutamic acid decarboxylase autoantibodies (GADA) and autoantibodies to islet antigen-2 (IA-2A) using the RSR radioimmunoassay (RIA) and enzyme linked immunosorbent assay (ELISA) kits. *Clin Lab* 2008; **54:** 227–35.

11 Tuomi T, Carlsson A, Li H, et al. Clinical and genetic characteristics of type 2 diabetes with and without GAD antibodies. *Diabetes* 1999; **48:** 150–57.

12 Vaziri-Sani F, Delli AJ, Elding-Larsson H, et al. A novel triple mix radiobinding assay for the three ZnT8 (ZnT8-RWQ) autoantibody variants in children with newly diagnosed diabetes. *J Immunol Methods* 2011; **371:** 25–37.

13 Almgren P, Lindqvist A, Krus U, et al. Genetic determinants of circulating GIP and GLP-1 concentrations. *JCI insight* 2017; **2:** 93306.

14 Levey AS, Stevens LA, Schmid CH, et al. A new equation to estimate glomerular filtration rate. *Ann Intern Med* 2009; **150:** 604–12.

15 Martinell M, Dorkhan M, Stalhammar J, Storm P, Groop L, Gustavsson C. Prevalence and risk factors for diabetic retinopathy at diagnosis (DRAD) in patients recently diagnosed with type 2 diabetes (T2D) or latent autoimmune diabetes in the adult (LADA). *J Diabetes Complications* 2016; **30:** 1456–61.

16 Levy JC, Matthews DR, Hermans MP. Correct homeostasis model assessment (HOMA) evaluation uses the computer program. *Diabetes Care* 1998; **21:** 2191–92.

17 Hennig C. Cluster-wise assessment of cluster stability. *Comput Stat Data An* 2007; **52:** 258–71.

18 Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 2007; **39:** 906–13.

19 Prasad RB, Groop L. Genetics of type 2 diabetes—pitfalls and possibilities. *Genes* 2015; **6:** 87–123.

20 Lyssenko V, Lupi R, Marchetti P, et al. Mechanisms by which common variants in the TCF7L2 gene increase risk of type 2 diabetes. *J Clin Invest* 2007; **117:** 2155–63.

21 Liu YL, Reeves HL, Burt AD, et al. TM6SF2 rs58542926 influences hepatic fibrosis progression in patients with non-alcoholic fatty liver disease. *Nat Commun* 2014; **5:** 4309.

22 Nguyen C, Varney MD, Harrison LC, Morahan G. Definition of high-risk type 1 diabetes HLA-DR and HLA-DQ types using only three single nucleotide polymorphisms. *Diabetes* 2013; **62:** 2135–40.

23 Li L, Cheng WY, Glicksberg BS, et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci Transl Med* 2015; **7:** 311ra174.

24 Groop L, Ekstrand A, Forsblom C, et al. Insulin resistance, hypertension and microalbuminuria in patients with type 2 (non-insulin-dependent) diabetes mellitus. *Diabetologia* 1993; **36:** 642–47.

25 Gnudi L, Coward RJ, Long DA. Diabetic nephropathy: perspective on novel molecular mechanisms. *Trends Endocrinol Metab* 2016; **27:** 820–30.

26 Welsh GI, Hale LJ, Eremina V, et al. Insulin signaling to the glomerular podocyte is critical for normal kidney function. *Cell Metab* 2010; **12:** 329–40.

27 Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet* 2011; **88:** 294–305.