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

Authors: E Ahlqvist[1] PhD, P Storm[1] PhD, A Käräjämäki[2†] MD, M Martinell[3†] MD, M Dorkhan[1] PhD, A Carlsson[4] PhD, P Vikman[1] PhD, RB Prasad[1] PhD, D Mansour Aly[1] MSc, P Almgren[1] MSc, Y Wessman[1], N Shaat[1] PhD, P Spegel[1,5] PhD, H Mulder[1] Prof., E Lindholm[1] PhD, O Melander[1] Prof., O Hansson[1] PhD, U Malmqvist[6] PhD, Å Lernmark[1] Prof., K Lahti[2] MD, T Forsén[7] PhD, T Tuomi[7,8,9] PhD, AH Rosengren[1,10] , PhD, L Groop[1,7*] Prof.

Affiliations:

[1*]Lund University Diabetes Centre, Department of Clinical Sciences, Lund University, Skåne University Hospital, SE-20502 Malmö, Sweden.

[2]Department of Primary Health Care, Vaasa Central Hospital, Hietalahdenkatu 2-4, 65130 Vaasa, Finland & Diabetes Center, Vaasa Health Care Center, Sepänkyläntie 14-16, 65100 Vaasa, Finland.

[3]Department of Public Health and Caring Sciences, Uppsala University, Uppsala, Sweden.

[4]Lund University Diabetes Centre, Department of Clinical Sciences, Lund University, Skåne University Hospital, SE-22185 Lund, Sweden.

[5]Department of Chemistry, Centre for Analysis and Synthesis, Lund University, Lund, Sweden

[6]Clinical Research and Trial Center, Lund University Hospital, Sweden.

[7]Folkhälsan Research Center, Helsinki, Finland.

[8]Abdominal Center, Endocrinology, Helsinki University Central Hospital; Research Program for Diabetes and Obesity, University of Helsinki, Helsinki, Finland.

[9]Finnish Institute for Molecular Medicine (FIMM), Helsinki University, Helsinki, Finland.

[10]Department of Neuroscience and Physiology, Wallenberg Center for Molecular and Translational Medicine, University of Gothenburg

[†]Equal contribution

*Correspondence to:

Professor Leif Groop, leif.groop@med.lu.se, phone +46-40-391202

Research in context

Evidence before this study

The current diabetes classification into T1D and T2D relies primarily on presence (T1D) or absence (T2D) of autoantibodies against pancreatic islet beta cell antigens and age at diagnosis (earlier for T1D). With this approach 75-85% of patients are classified as T2D. A third subgroup, Latent Autoimmune Diabetes in Adults (LADA, <10%), is defined by presence of autoantibodies against glutamate decarboxylase (GADA) with onset in adult age. In addition, several rare monogenic forms of diabetes have been described, including Maturity Onset Diabetes of the Young (MODY) and neonatal diabetes. This information is provided by national guidelines (ADA, WHO, IDF, Diabetes UK etc.) but has not been much updated during the past 20 years and very few attempts have been made to explore heterogeneity of T2D. A topological analysis of potential T2D subgroups using electronic health records was published in 2015 but this information has not been implemented in the clinic.

Added value of this study

Here we applied a data-driven cluster analysis of 6 simple variables measured at diagnosis in 4 independent cohorts of newly-diagnosed diabetic patients (N=14,755) and identified 5 replicable clusters of diabetes patients, with significantly different patient characteristics and risk of diabetic complications. Particularly, individuals in the most insulin-resistant cluster 3 had significantly higher risk of diabetic kidney disease.

Implications of the available evidence

This new sub-stratification may help to tailor and target early treatment to patients who would benefit most, thereby representing a first step towards precision medicine in diabetes.

Abstract

Background

Diabetes is presently classified into two main forms, type 1 (T1D) and type 2 diabetes (T2D), but especially T2D is highly heterogeneous. A refined classification could provide a powerful tool individualize treatment regimes and identify individuals with increased risk of complications already at diagnosis.

Methods

We applied data-driven cluster analysis (k-means and hierarchical clustering) in newly diagnosed diabetic patients (N=8,980) from the Swedish ANDIS (All New Diabetics in Scania) cohort, using six variables (GAD-antibodies, age at diagnosis, BMI, HbA1c, HOMA2-B and HOMA2-IR), and related to prospective data on development of complications and prescription of medication from patient records. Replication was performed in three independent cohorts: the Scania Diabetes Registry (SDR, N=1466), ANDIU (All New Diabetics in Uppsala, N=844) and DIREVA (Diabetes Registry Vaasa, N=3485). Cox regression and logistic regression was used to compare time to medication, time to reaching the treatment goal and risk of diabetic complications and genetic associations.

Findings

We identified 5 replicable clusters of diabetes patients, with significantly different patient characteristics and risk of diabetic complications. Particularly, individuals in the most insulin-resistant cluster 3 had significantly higher risk of diabetic kidney disease, but had been prescribed similar diabetes treatment compared to the less susceptible individuals in clusters 4 and 5. The insulin deficient cluster 2 had the highest risk of retinopathy. In support of the clustering, genetic associations to the clusters differed from those seen in traditional T2D.

Interpretation

We could stratify patients into five subgroups with differing disease progression and risk of diabetic complications. This new substratification may 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 one of the greatest threats to human health.[1] Unfortunately, current 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 upon measurement of only one metabolite, glucose, but the disease is very heterogeneous with regard to clinical presentation and progression.

The current diabetes classification into T1D and T2D relies primarily on presence (T1D) or absence (T2D) of autoantibodies against pancreatic islet beta cell antigens and age at diagnosis (earlier for T1D). With this approach 75-85% of patients are classified as T2D. A third subgroup, Latent Autoimmune Diabetes in Adults (LADA, <10%), defined by presence of autoantibodies against glutamate decarboxylase (GADA) is phenotypically indistinguishable from T2D at diagnosis but become more T1D-like with time.[2] With the introduction of gene sequencing for clinical diagnostics several rare monogenic forms of diabetes were described, including Maturity Onset Diabetes of the Young (MODY) and neonatal diabetes.[3, 4]

A limitation of current treatment guidelines is that they respond to poor metabolic control when it has developed but lack means to predict which patients will need intensified treatment. evidence suggests that early treatment is critical for prevention of life-shortening complications since target tissues seem to remember poor metabolic control decades later, also referred to as "metabolic memory".[5, 6]

A refined classification could provide a powerful tool to identify those at greatest risk of complications already at diagnosis, and enable individualized treatment regimes in the same way as a 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, genetically and clinically to the current classification in four separate populations from Sweden and Finland.

Methods

Study populations

The ANDIS (All New Diabetics in Scania) project (http://andis.ludc.med.lu.se/) aims to recruit all incident cases of diabetes within Scania County in Sweden (~1,200,000 inhabitants). All health care providers in Scania were invited; the current registration covered the period January 1st 2008 until November 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 time was 4.01 years (IQR 2.02-6.00).

The Scania Diabetes Registry (SDR), recruited in the same region 1996- 2009, included >7,400 individuals with diabetes of all types, 1,466 of whom were recruited two years or less after diagnosis and had all data necessary for clustering.[8] Median follow-up time was 11.05 years (IQR 8.33-14.56).

ANDIU (All New Diabetics In Uppsala) is a project similar to ANDIS in the Uppsala region (~300,000 inhabitants) in Sweden (http://www.andiu.se). N=844 patients had complete data for all clustering variables.

DIREVA (Diabetes Registry Vaasa) from Western Finland (~170,000 inhabitants) includes 5,107 individuals with diabetes recruited 2009-2014.

MDC-CVA (Malmö Diet and Cancer CardioVascular Arm) includes subjects (n=3,300), 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]

Measurements

In ANDIS blood samples were drawn at registration. Fasting plasma glucose was analyzed after an overnight fast using the HemoCue Glucose System (HemoCue AB, Ängelholm, Sweden). C-peptide concentrations were determined using ElectroChemi–LuminiscenceImmunoassay on Cobas e411 (Roche Diagnostics, Mannheim, Germany) or radioimmunoassay (Human C-peptide RIA; Linco, St Charles, MO, USA; or Peninsula Laboratories, Belmont, CA, USA). In ANDIS and SDR GADA was measured by Enzyme-Linked Immunosorbent Assay (ELISA) (ref <11

U/ml[10]) or with radiobinding assays (RBA) using 35S-labelled protein[11] (positive cut-off: 5 RU or 32 IU/ml). The RBA showed 62–88% sensitivity and 91–99% specificity, and the ELISA assay showed 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 were measured using ELISA (RSR, Cardiff, UK; positive cut-off 10 IU/ml). ZnT8A antibodies were measured using an RBA as previously described.[12] HbA1c was measured at diagnosis using the Variant II Turbo HbA1c Kit-2.0 (Bio-Rad, Copenhagen, Denmark). Measurements of HbA1c, ALT, ketones and serum creatinine over time were obtained from the Clinical Chemistry database.

Genotyping

Genotyping of ANDIS samples was carried out on frozen DNA samples prepared from blood using Gentra Puregene Blood Kits (Qiagen, Hilden, Germany) using iPlex (Sequenom, San Diego, California, US) or TaqMan assays (Thermo Fisher Scientific) at the Clinical Research Center in Malmö, Sweden. In ANDIS, 5625 of the clustred individuals were genotyped, of which 1714 were excluded due to non-Swedish origin and 164 due to call rate <90%. MDC-CVA samples were genotyped at the Broad genotyping facility using the Infinium OmniExpressExome v1.0 B Beadchip array (Illumina, San Diego, CA, US). Quality control was done as previously described.[13] All SNPs were in Hardy-Weinberg equilibrium in controls.

Definitions of diabetic complications

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

Macroalbuminuria was defined as at least two out of three consecutive visits with albumin excretion rate (AER) ≥200 µg/min, AER ≥300 mg/24 h or albumin-creatinine ratio (ACR) ≥25/35 mg/mmol for men/women.

Diabetic retinopathy was diagnosed by an ophthalmologist based on fundus photographs.[15] Coronary events (CE) were defined by ICD-10 codes I20-21, I24, I251, I253-I259. Stroke was

defined by ICD-10 codes I60-I61 and I63-I64. Individuals with known prior events were excluded.

Cluster analysis

We based the selection of model parameters on the premise that patients develop diabetes when they no longer can increase their insulin secretion (whatever the reason) to meet the increased demands imposed by obesity and insulin resistance. Additionally, parameters should be easily obtainable from different clinical settings without interpretation and include the minimum number of laboratory tests. Therefore we chose BMI, age at onset of diabetes and Homeostasis Model Assessment 2 estimates of beta-cell function (HOMA2-B) and insulin resistance (HOMA2-IR) based upon C-peptide (which performs better than insulin in diabetes patients) calculated using the HOMA calculator (University of Oxford, UK).[16] Presence or absence of GADA was included as a binary variable. Cluster analysis was performed on values centered to mean=0 and SD=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 SD; N=42) were excluded. TwoStep clustering, of which the first step estimates the optimal number of clusters based upon silhouette width and the second performs hierarchical clustering, was performed in SPSS v23 for 2 to 15 clusters using log-likelihood as distance measure and Schwarz's Bayesian criterion for clustering. K-means clustering was performed with k=4 using the kmeansruns function (runs=100) in the fpc package in R. Only GADA negative individuals were included because the k-means method does not accomodate binary variables and all GADA positive individuals clustered together using the TwoStep method. Cluster center coordinates in ANDIS are presented in Table S3.

Clusterwise stability was assesed by resampling the dataset 2,000 times and computing the Jaccard similarities to the original cluster.[17] Generally, stable clusters should yield a Jaccard similarity >0·75.[17] Cluster labels were assigned by examining cluster variable means. The GADA positive cluster was labelled as Severe Autoimmune Diabetes (SAID), the GADA negative cluster with the lowest mean HOMA2-B was labelled Severe Insulin-Deficient Diabetes (SIDD), the cluster with high HOMA2-IR and age at diagnosis was labelled Severe Insulin-Resistant

Diabetes (SIRD), the cluster with high BMI and low age at onset was labelled Mild Obesity-related Diabetes (MOD) and the remaining cluster Mild Age-Related Diabetes (MARD).

Statistical analysis

Risk of complications was calculated using cox regression in SPSS v23. Covariates were included as stated in the text. Post hoc comparisons of effects across clusters were tested in Stata v13.1.

Associations between clusters and genotypes were calculated using the MLE method in SNPtest2 v2.5.2.[18] The equality of odds ratios across strata was tested using seemingly unrelated estimation (suest) in Stata v13.1. Patients from each cluster were used as cases and non-diabetic individuals from the MDC-CVA cohort were used as controls. Patients of non-Swedish origin were excluded. Bonferroni correction was used to determine significance for multiple tests. Genetic risk scores were calculated based on number of risk alleles weighed by their effect sizes reported in previous GWAS studies and logistic regression was performed for each cluster against the controls in SPSS v23.

Funding

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

Ethical approval

The ANDIS and SDR study protocols were approved by the Regional Ethics Review Committee in Lund (ANDIS: Dnr. 584/2006 and 2012/676. SDR: LU 35-99). DIREVA was approved by the Ethical committee in Vasa (Dnr. 6/2007). ANDIU was approved by the Regional Ethics Review Committee in Uppsala (Dnr. 2011/155). All participants have given written informed consent.

Results

We first analyzed a cohort of 14,652 newly diagnosed diabetic patients from Sweden termed ANDIS. Of them, 932 (6·4%) were registered before age 18 and not included in analyses of adult diabetes. Of the adult patients, 204 (1·5%) had T1D (defined as GADA positive and C-peptide < 0·3 nmol/l), 723 (5·3%) LADA (GADA-positive and C-peptide ≥ 0·3 nmol/l), 162 (1·2%) secondary diabetes (coexisting pancreatic disease) and 519 (3·8%) were unclassifiable due to missing data. The remaining 12,112 patients (88·3%) were considered to have T2D (Table S1).

Five quantitative variables (age at diagnosis, BMI, HbA1c, HOMA2-B and HOMA2-IR), plus presence or absence of GADA as a binary variable, were used in cluster analysis to reclassify patients into novel diabetes subgroups. Patients with complete data for the clustering variables (N=8,980) were included in further analyses.

First, we applied the TwoStep clustering method as implemented in SPSS. The minimum silhouette width was found for 5 clusters in both men (N=5,334) and women (N=3,646), exhibiting similar cluster distributions and characteristics (Figure S1). We verified the results using k-means clustering in GADA negative patients, resulting in similar cluster distributions as TwoStep with the same overall cluster characteristics in both sexes (Figure 1B, 2 and S2). Cluster stability was estimated as Jaccard means[17], which were >0·8 for all clusters regardless of sex.

Cluster 1, including 577 (6·4%) of the clustered patients (SAID) was characterized by early onset, relatively low BMI, poor metabolic control, insulin deficiency, and presence of GADA (Table S2). Cluster 2 (SIDD) encompassing 1,575 (17·5%) patients was GADA negative but otherwise similar to SAID: low age at onset, relatively low BMI, low insulin secretion (low HOMA2-B) and poor metabolic control. Cluster 3 (SIRD; n=1,373; 15·3%) was characterized by insulin resistance (high HOMA2-IR) and high BMI. Cluster 4 was also characterized by obesity but not by insulin resistance (MOD; n=1,942; 21·6%). Patients in cluster 5 were older (MARD; n=3,513; 39·1%) but showed, as cluster 4, only modest metabolic derangements.

We used three independent cohorts to replicate the clustering: SDR (N=1,466), ANDIU (N=844) and DIREVA (N=3,485). In SDR, the optimal number of clusters was also estimated to be 5 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 1C, S3A and B). Jaccard bootstrap means were >0·8 for all clusters. K-means clustering in ANDIU also replicated

the results from ANDIS (Figure 1D, S3D). In the DIREVA cohort we tested whether clustering would give similar results in patients with longer diabetes duration (mean 10·15±10·34; N=2,607) as newly-diagnosed diabetes (diabetes duration <2 years, N=878). Encouragingly, the results were comparable (Figure 1E and F, S4 A and C).

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 based on which cluster they were most similar to, calculated as their Euclidian distance from the nearest cluster center derived from ANDIS coordinates, and found similar distributions (Figure S3 C and E, Figure S4 B and D). Sensitivity and specificity was highest in ANDIU and DIREVA patients recruited near diagnosis (Table S4), likely reflecting how and when clustering variables were obtained.

We then compared disease progression, treatment and development of diabetic complications between clusters in ANDIS. SAID and SIDD had markedly higher HbA1c at diagnosis compared to other clusters, a difference persisting throughout the follow-up period (Figure 3A). Ketoacidosis at diagnosis was most frequent in SAID (30·5%) and SIDD (25·1%), compared to others (<5%, Figure S5). HbA1c was the strongest predictor of ketoacidosis at diagnosis (OR 2·73[2·46-3·03], $p=2{\cdot}0x10^{-82}$, per 1SD change, Table S5). SIRD had the highest prevalence of non-alcoholic fatty liver disease (NAFLD, Figure S6). Zinc transporter 8A antibodies were primarily seen in SAID (27·3% positive compared to <1·5% in other clusters; Figure S7).

At registration, insulin had been prescribed to 41·9% of patients in SAID and 29·1% in SIDD but < 4% of patients in clusters 3-5 (Table S2, Figure S8). Time to insulin was shortest in SAID (HR 17·05[14·34-20·28] compared to MARD, Figure 4A, Table S6), followed by SIDD (HR 9·23[7·88-10·81]). The proportion of patients on metformin was highest in SIDD and lowest in SAID (Figure S8, 4B), but also surprisingly low in SIRD which should benefit most from metformin, demonstrating 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 (Figure S9). SIDD had the shortest time to a second oral diabetes treatment (Figure 4C, Table S6) and the longest time to reaching the treatment goal (HbA1c <52 mmol/mol; Figure 4D).

In ANDIS, SIRD had the highest risk of developing chronic kidney disease (CKD) during follow-up of 3·9±2·3 years (Table S7). For CKD stage 3A (eGFR<60 ml/min) the age and sex adjusted risk was >2-times higher (HR 2·41[2·08-2·79], $p=1{\cdot}4x10^{-31}$, Figure S10A) and for stage 3B (eGFR<45 ml/min) >3-times higher compared to MARD (HR 3·34[2·59-4·30], $p=8{\cdot}3x10^{-21}$, Figure 3B). SIRD also showed higher risk of diabetic kidney disease defined as persistent macroalbuminuria (Figure S10B, HR 2·28[1·6-3·23], $p=3{\cdot}0x10^{-6}$). Also in the SDR cohort (follow-up 11·0±4·4 years), SIRD had the highest risk of CKD (Table S9), and macroalbuminuria (HR 2·18[1·31-3·63], p=0.0026, Figure 3D). Strikingly, SIRD patients had almost five times higher risk of ESRD than MARD (HR 4·89[2·68-8·93], $p=2{\cdot}4x10^{-7}$, Figure 3E). The increased prevalence of kidney disease in SIRD was also confirmed in the DIREVA cohort (Figure S12).

Early signs of diabetic retinopathy (mean duration 135 days) were more common in SIDD than in other clusters (OR 1·6[1·3-1·9], $p=9{\cdot}7x10^{-7}$ compared to MARD; Figure S11A). The higher prevalence of retinopathy in SIDD was replicated in ANDIU (Figure S11B) and SDR (HR 1·33[1·15-1·54], p=0·0001; Figure 3F, Table S10).

Although unadjusted risk of coronary events and stroke was lowest in SAID, SIDD and MOD there was no significant difference in age-adjusted risk (Figure 3C, S10, Table S8 and S11).

Finally, we analyzed genetic loci previously shown to be associated with diabetes and related traits[19] (Table 1). Each cluster was compared to a non-diabetic cohort (MDC-CVA) from the same geographical region.[9] Notably, no genetic variant was associated (p<0·01) with all clusters (Table S12). Strikingly, the strongest T2D-associated variant in the *TCF7L2* (rs7903146) gene[20] was associated with SIDD, MOD and MARD, but not with SIRD (only significant difference after correction for multiple testing; Table 1). The variant rs10401969 in the *TM6SF2* gene previously associated with NAFLD[21] was associated with SIRD but not MOD suggesting that SIRD is characterized by more unhealthy (metabolic syndrome) obesity. Importantly, rs2854275 in the *HLA* locus (previously associated with T1D) was strongly associated with SAID (OR 2.05[1.69-2.56]; $p=5.7x10^{-10}$), but not with SIDD (OR 0.82[0.66-1.00]; p=0.0777) supporting the non-autoimmune nature of the SIDD cluster. A genetic risk score for T2D (Tables S13, S14) was significantly associated with all clusters (p<0.0008) except SIRD (p=0.1602). An insulin secretion risk score was significantly associated with MOD (p=0.0002) and MARD ($p=1.0x10^{-6}$)

and nominally with SIDD (p=0.0143) but showed no evidence of association with SAID or SIRD (p>0.5).

Discussion

Taken together, this study demonstrates that this new clustering of adult-onset diabetes patients is superior to the classical diabetes classification since it identifies patients with high risk of diabetic complications and provides information about underlying disease mechanisms, thereby guiding choice of therapy. Importantly, this information is available already at diagnosis. In contrast to previous attempts to dissect the heterogeneity of diabetes[22] we used variables reflecting key aspects of diabetic disease that are monitored in patients. Thus, this clustering can easily be applied to both existing diabetes cohorts (e.g. from drug trials) and patients in the diabetes clinic. A web-tool to assign patients to specific clusters, provided above variables have been measured, is under development.

While SAID overlapped with T1D and LADA, SIDD and SIRD represent two novel severe forms of diabetes previously masked within T2D. It would be reasonable to target intensified treatment resources to these clusters to prevent diabetic complications. SIRD had a markedly increased risk of kidney complications, reinforcing the association between insulin resistance and kidney disease.[23] Insulin resistance has been associated with higher salt sensitivity, glomerular hypertension, hyperfiltration, and declining renal function, all hallmarks of diabetic kidney disease (DKD).[24] The increased incidence of DKD in this study was seen in spite of relatively low HbA1c, suggesting that glucose-lowering therapy is not the ultimate way of preventing DKD. In support of this, mice with podocyte-specific knockout of the insulin receptor, mimicking the reduced insulin signaling seen in insulin resistant individuals, developed DKD even during normoglycemic conditions.[25] Although differences were not as pronounced as for DKD, insulin deficiency and/or hyperglycemia seem to be important triggers of retinopathy with the highest prevalence observed in SIDD.

The fact that clustering gave similar results in newly diagnosed patients and patients with longer diabetes duration, and that the key variable C-peptide remained relatively stable over time (Figure S13), 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. Especially the lack of association of the genetic risk scores for T2D and insulin secretion with SIRD indicate that this group might have a different etiology than the other clusters. Notably, hepatic insulin resistance seems to be a feature of NAFLD, as the NAFLD-associated SNP in the *TM6SF2* gene was associated with SIRD but not with MOD.

Limitations

We cannot at this stage claim that the new clusters represent different etiologies of diabetes, nor that this represents the optimal classification of diabetes subtypes, Also, it still needs to be shown in prospective studies whether patients (especially from the periphery of clusters) can move between clusters 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 by including additional cluster variables e.g. 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 determine the inherited proportion of each cluster using heritability partitioning models.[26] This classification was derived primarily on Northern Europeans with limited non-Scandinavian representation, and the applicability of this strategy to patients of other ethnicity needs to be assessed. Only two types of auto-antibodies were measured and the influence of other antibodies on clustering performance is unknown. We also 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.

Conclusions

Taken together, the current data demonstrate that the combined information from a few variables central to the development of diabetes is superior to measurement of only one metabolite, glucose. By 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 opens up for randomized trials targeting insulin secretion in SIDD and insulin resistance in SIRD.

Acknowledgements

We thank all the patients and the health care providers for their support and willingness to participate. We would also like to 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. Finally we would like to thank Rita Jedlert and Region Skåne (Scania County) as well as the ANDIS steering committee for their support.

Financial Support

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.

Declaration of interest

The authors have no conflicts of interest.

Author contributons

EA, PS, PV, TT, AHR and LG contributed with the conception of the work. EA, PS, AK, MM, MD, AC, PV, YW, NS, PS, HM, EL, OM, OH, UM, ÅL, KL, TF, TT, AHR and LG contributed to the data collection. EA, PS, MM, RP, DMA and PA contributed to the data analysis. EA, PS , 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.

References

1. Collaboration NCDRF. Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4.4 million participants. *Lancet* 2016; **387**(10027): 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**(2): 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**(10): 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**(6608): 455-8.
5. Reddy MA, Zhang E, Natarajan R. Epigenetic mechanisms in diabetic complications and metabolic memory. *Diabetologia* 2015; **58**(3): 443-55.
6. Brownlee M. The pathobiology of diabetic complications: a unifying mechanism. *Diabetes* 2005; **54**(6): 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**(5): 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**(11): 983-9.
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**(6): 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**(7-8): 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**(1): 150-7.
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. *Journal of immunological methods* 2011; **371**(1-2): 25-37.

13. Almgren P, Lindqvist A, Krus U, et al. Genetic determinants of circulating GIP and GLP-1 concentrations. *JCI insight* 2017; **2**(21).

14. Levey AS, Stevens LA, Schmid CH, et al. A new equation to estimate glomerular filtration rate. *Ann Intern Med* 2009; **150**(9): 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**(8): 1456-61.

16. Levy JC, Matthews DR, Hermans MP. Correct homeostasis model assessment (HOMA) evaluation uses the computer program. *Diabetes Care* 1998; **21**(12): 2191-2.

17. Hennig C. Cluster-wise assessment of cluster stability. *Comput Stat Data An* 2007; **52**(1): 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**(7): 906-13.

19. Prasad RB, Groop L. Genetics of type 2 diabetes-pitfalls and possibilities. *Genes (Basel)* 2015; **6**(1): 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**(8): 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. *Nature communications* 2014; **5**: 4309.

22. Li L, Cheng WY, Glicksberg BS, et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science translational medicine* 2015; **7**(311): 311ra174.

23. 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**(7): 642-7.

24. Gnudi L, Coward RJ, Long DA. Diabetic Nephropathy: Perspective on Novel Molecular Mechanisms. *Trends Endocrinol Metab* 2016; **27**(11): 820-30.

25. Welsh GI, Hale LJ, Eremina V, et al. Insulin signaling to the glomerular podocyte is critical for normal kidney function. *Cell metabolism* 2010; **12**(4): 329-40.
26. Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet* 2011; **88**(3): 294-305.

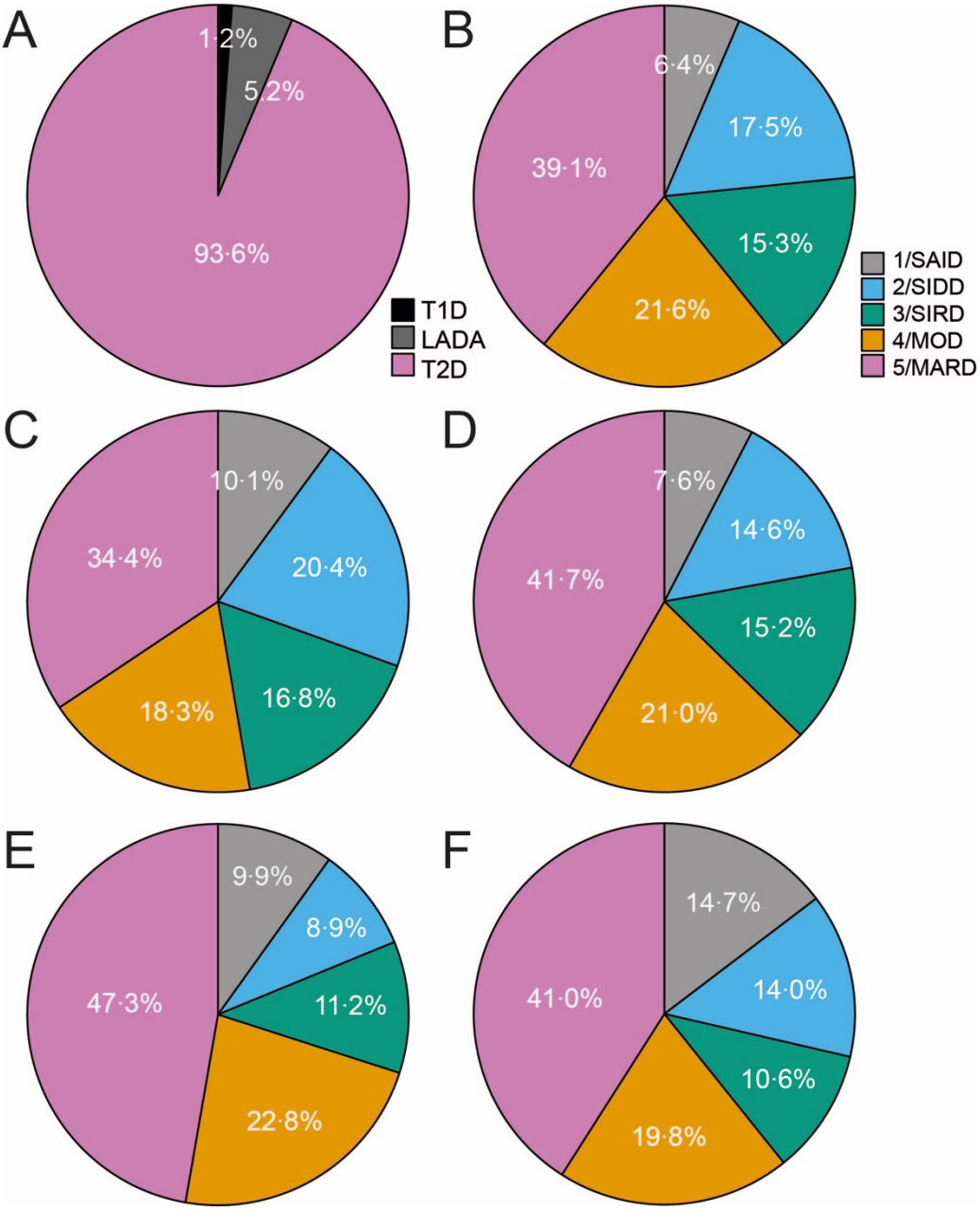

Figure 1. Patient distribution using different methods for classification.

Distribution of ANDIS patients included in the clustering using (A) traditional classification and (B) k-means clustering N=8,980. Distribution of patients using k-means clustering in SDR, N=1,466 (C), ANDIU, N=844 (D) and in DIREVA stratified for newly diagnosed, N=878 (E) and long duration, 2,607 (F).

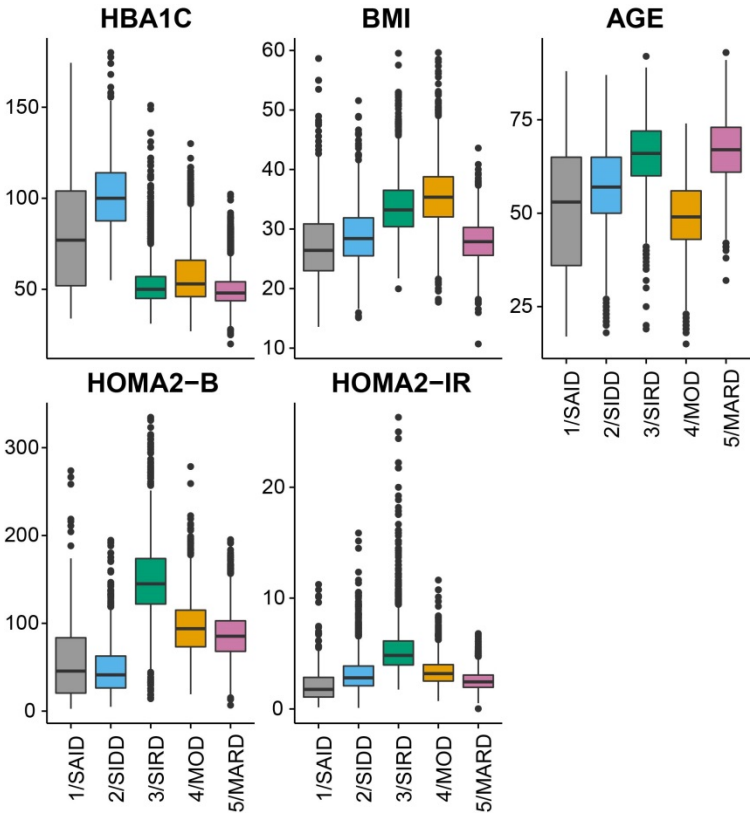

Figure 2. Cluster characteristics in ANDIS.

Distributions of HbA1c (mmol/mol) at diagnosis, and BMI (kg/m2), age (years), HOMA2-B (%) and HOMA2-IR at registration in ANDIS for each cluster. K-means clustering was performed separately for men and women, pooled data are shown here (cluster 2-5).

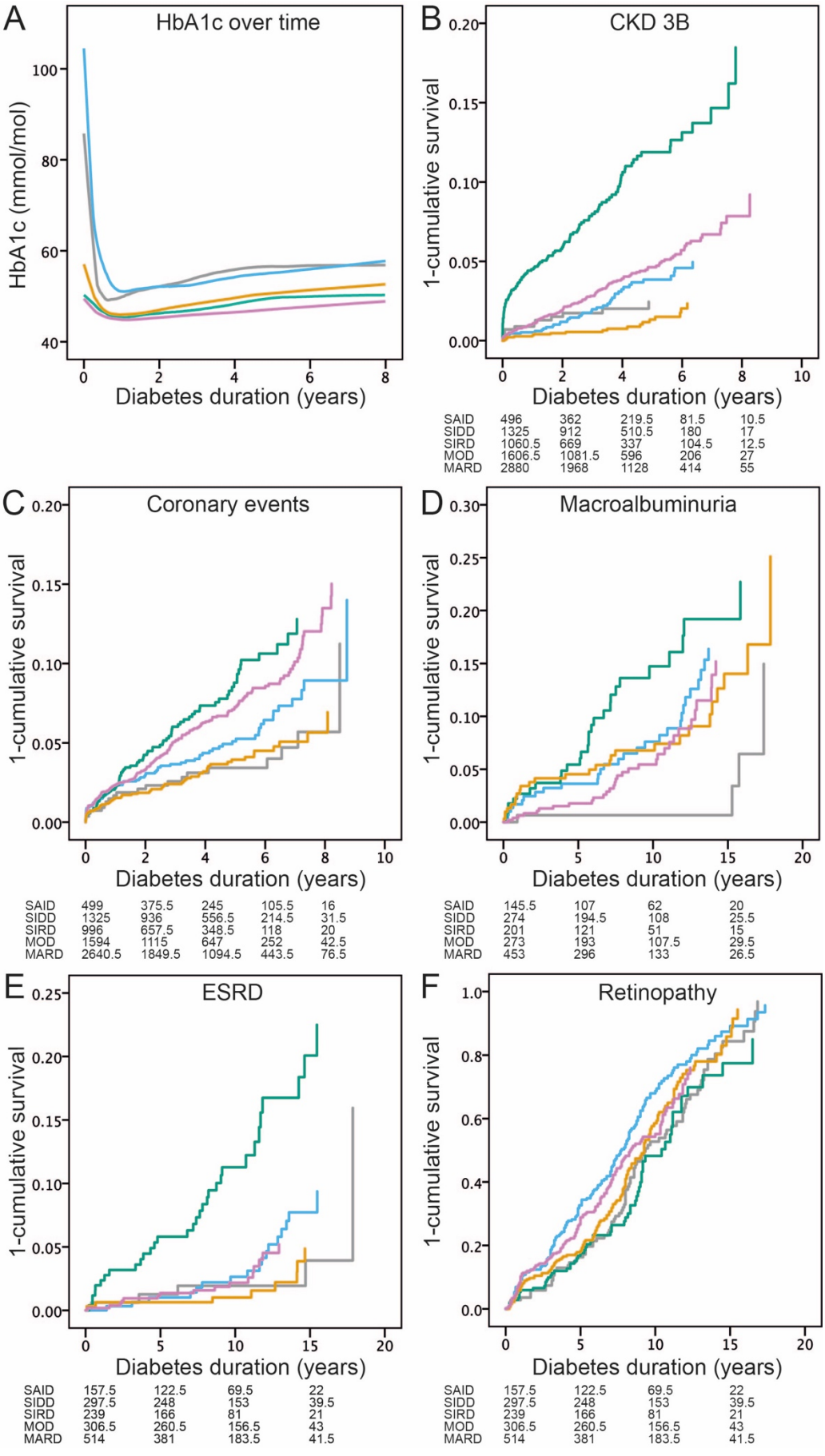
A
HbA1c over time
HbA1c (mmol/mol)
Diabetes duration (years)
B
CKD 3B
1-cumulative survival
Diabetes duration (years)
SAID 496 362 219.5 81.5 10.5
SIDD 1325 912 510.5 180 17
SIRD 1060.5 669 337 104.5 12.5
MOD 1606.5 1081.5 596 206 27
MARD 2880 1968 1128 414 55
C
Coronary events
1-cumulative survival
Diabetes duration (years)
SAID 499 375.5 245 105.5 16
SIDD 1325 936 556.5 214.5 31.5
SIRD 996 657.5 348.5 118 20
MOD 1594 1115 647 252 42.5
MARD 2640.5 1849.5 1094.5 443.5 76.5
D
Macroalbuminuria
1-cumulative survival
Diabetes duration (years)
SAID 145.5 107 62 20
SIDD 274 194.5 108 25.5
SIRD 201 121 51 15
MOD 273 193 107.5 29.5
MARD 453 296 133 26.5
E
ESRD
1-cumulative survival
Diabetes duration (years)
SAID 157.5 122.5 69.5 22
SIDD 297.5 248 153 39.5
SIRD 239 166 81 21
MOD 306.5 260.5 156.5 43
MARD 514 381 183.5 41.5
F
Retinopathy
1-cumulative survival
Diabetes duration (years)
SAID 157.5 122.5 69.5 22
SIDD 297.5 248 153 39.5
SIRD 239 166 81 21
MOD 306.5 260.5 156.5 43
MARD 514 381 183.5 41.5

Figure 3. Progression of disease over time by cluster

Figure 3 shows mean HbA1c over time by loess regression (A), time to CKD at least stage 3B (B) and coronary events (C) in ANDIS; Macroalbuminuria (D), ESRD (E) and mild non-proliferative to proliferative diabetic retinopathy (F), in the SDR cohort. Kidney function was not tested at diagnosis and therefore set to the first screening date. Thus it is not known how many were already affected at diagnosis.

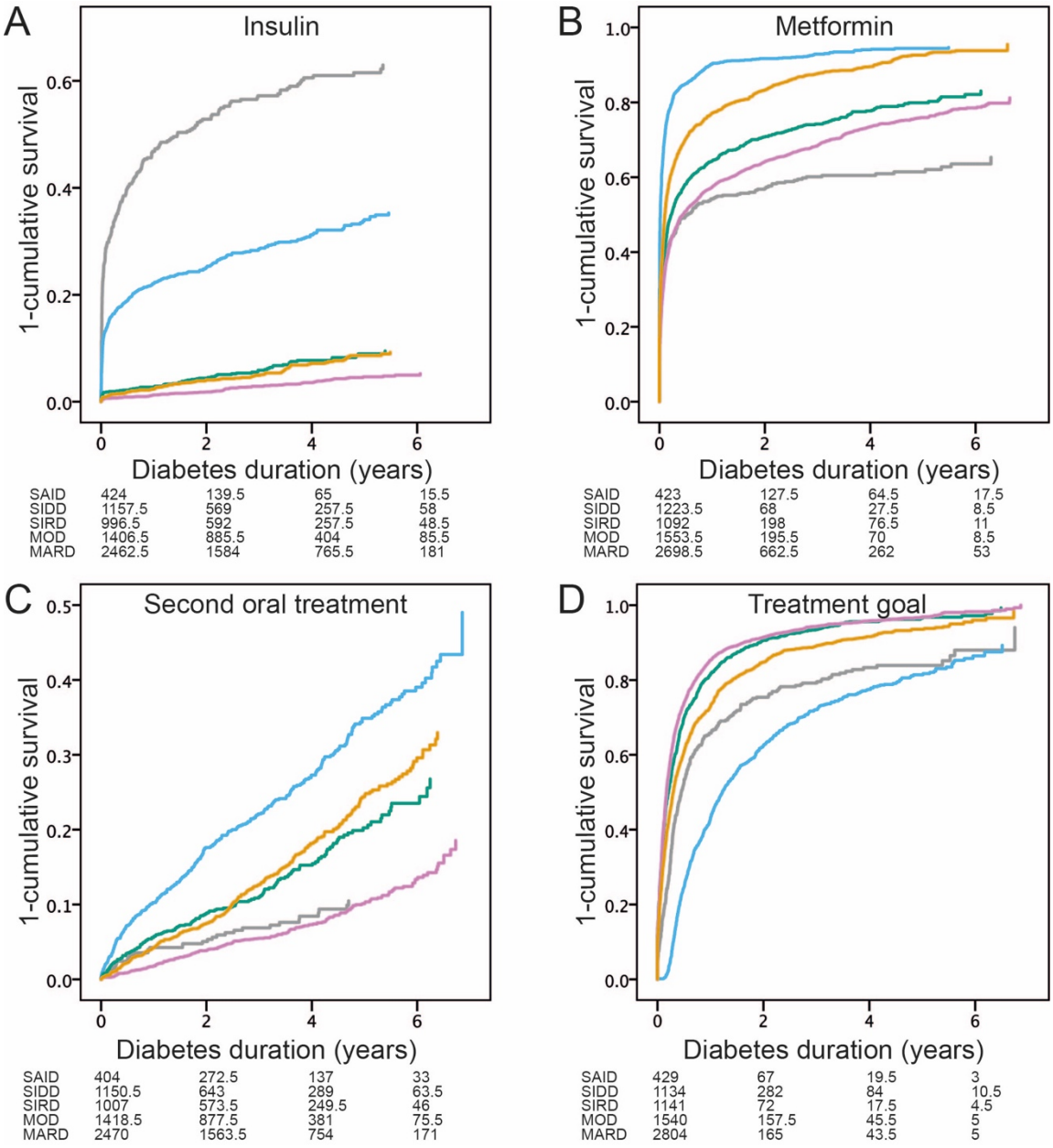

Figure 4. Antidiabetic therapy in ANDIS during follow-up.

Cox regressions of time to treatment with insulin (A), metformin (B), oral medication other than metformin (C) or (D) reaching treatment goal (HbA1c <52mmol/mol). Cluster 1/SAID had the shortest time to insulin. Cluster 2/SIDD had a shorter time to insulin, metformin and any other oral medication than clusters 3 to 5. Despite this, cluster 2/SIDD reached the treatment goal significantly later than other clusters. For statistics see table S6.

Table 1. Genetic associations with specific ANDIS clusters reaching at least nominal significance for difference between clusters 2 to 5.

				1/SAID		2/SIDD		3/SIRD		4/MOD		5/MARD		Difference cluster 2-5
				N=313		N=676		N=603		N=727		N=1646		
SNP	Gene	EA/ NEA	MAF	OR	P	OR	P	OR	P	OR	P	OR	P	P
rs7903146	*TCF7L2*	T/C	0·26	1·17(0·97-1·40)	0·0766	1·51(1·33-1·71)	$2·8x10^{-10}$	1·00(0·87-1·15)	0·8626	1·38(1·21-1·56)	$5·7x10^{-7}$	1·41(1·28-1·55)	$1·1x10^{-12}$	$9·6x10^{-6}$*
rs2237895	*KCNQ1*	C/T	0.41	1.08(0.91-1.28)	0.3106	1.13(1.00-1.28)	0.0518	0.85(0.74-0.97)	0.0272	0.98(0.86-1.10)	0.8770	1.13(1.03-1.23)	0.0196	0.0008
rs1111875	*HHEX/IDE*	G/A	0·41	1·16(0·98-1·38)	0·1044	1·21(1·07-1·37)	0·0045	1·05(0·92-1·19)	0·5104	0·94(0·84-1·06)	0·3139	1·11(1·02-1·22)	0·0228	0·0106
rs4402960	*IGF2BP2*	T/G	0·29	1·04(0·87-1·24)	0·5013	1·23(1·08-1·40)	0·0002	1·01(0·88-1·16)	0·5279	1·04(0·92-1·18)	0·3089	1·22(1·11-1·33)	$2·1x10^{-6}$	0·0117
rs10811661	*CDKN2B*	T/C	0·16	0·87(0·70-1·08)	0·2421	1·33(1·11-1·59)	0·0014	0·98(0·83-1·17)	0·8494	0·99(0·84-1·16)	0·9221	1·18(1·04-1·33)	0·0054	0·0149
rs10830963	*MTNR1B*	G/C	0·29	0·84(0·70-1·01)	0·0540	0·93(0·82-1·07)	0·2643	0·89(0·77-1·02)	0·0555	1·13(1·00-1·28)	0·0673	1·05(0·96-1·15)	0·2859	0·0151
rs13266634	*SLC30A8*	T/C	0·31	0·98(0·82-1·17)	0·7814	0·93(0·82-1·06)	0·2302	1·11(0·97-1·27)	0·1071	1·07(0·94-1·21)	0·2986	0·92(0·83-1·01)	0·04573	0·0160
rs12970134	*MC4R*	G/A	0·27	0·95(0·79-1·14)	0·5238	0·97(0·85-1·11)	0·5494	0·99(0·86-1·13)	0·5942	0·87(0·77-0·99)	0·0229	1·07(0·97-1·18)	0·1847	0·0230
rs10401969	*TM6SF2*	T/C	0·10	0·75(0·58-0·97)	0·0376	0·69(0·58-0·83)	0·0002	0·62(0·52-0·75)	$3·1x10^{-6}$	0·89(0·73-1·07)	0·2603	0·77(0·67-0·89)	0·0005	0·0233
rs4607103	*ADAMTS9-AS2*	T/C	0·24	1·05(0·87-1·27)	0·5399	0·89(0·77-1·03)	0·1547	0·93(0·80-1·08)	0·4245	1·12(0·98-1·27)	0·0642	0·92(0·83-1·01)	0·1314	0·0278
rs17271305	*VPS13C*	G/A	0·40	1·00(0·84-1·19)	0·9325	0·97(0·86-1·10)	0·8396	1·11(0·98-1·26)	0·0921	0·88(0·78-0·99)	0·0491	0·93(0·85-1·02)	0·1678	0·0281
rs11920090	*SLC2A2*	T/A	0·13	0·94(0·74-1·20)	0·5404	0·83(0·70-0·99)	0·01624	0·91(0·76-1·09)	0·2263	0·97(0·82-1·16)	0·6305	1·08(0·95-1·24)	0·4351	0·0368
rs5219	*KCNJ11*	T/C	0·38	1·05(0·88-1·25)	0·6114	1·18(1·04-1·34)	0·0121	1·03(0·90-1·18)	0·6737	1·28(1·13-1·44)	0·0001	1·10(1·01-1·21)	0·0324	0·0453
rs7961581	*TSPAN8*	T/C	0·26	0·97(0·80-1·17)	0·6936	1·05(0·92-1·21)	0·5490	1·13(0·98-1·31)	0·1145	0·99(0·87-1·13)	0·7963	0·92(0·84-1·02)	0·1135	0·0464

Maximum likelihood estimation using geographically matched non-diabetic individuals as controls (N=2,754). EA=Effect allele; NEA=Non effect allele

*Significant after correction for multiple testing (77 tests).