ORIGINAL ARTICLE

WILEY

Disease heterogeneity of adult diabetes based on routine clinical variables at diagnosis: Results from the German/Austrian Diabetes Follow-up Registry

Julia M. Grimsmann PhD[1,2†] | Sascha R. Tittel MS[1,2†] | Peter Bramlage MD[3] | Benjamin Mayer PhD[4] | Andreas Fritsche MD[5] | Jochen Seufert MD[6] | Markus Laimer MD[7] | Stefan Zimny MD[8] | Sebastian M. Meyhoefer MD[2,9] | Michael Hummel MD[10] | Reinhard W. Holl MD[1,2]

[1]Institute of Epidemiology and Medical Biometry, ZIBMT, Ulm University, Ulm, Germany

[2]German Centre for Diabetes Research (DZD), Munich-Neuherberg, Germany

[3]Institute for Pharmacology and Preventive Medicine, Cloppenburg, Germany

[4]Institute of Epidemiology and Medical Biometry, Ulm University, Ulm, Germany

[5]German Center for Diabetes Research, Eberhard Karl University, Tuebingen, Germany

[6]Division of Endocrinology and Diabetology, Department of Medicine II, Faculty of Medicine, Medical Center, University of Freiburg, Freiburg, Germany

[7]Department of Diabetes, Endocrinology, Nutritional Medicine and Metabolism, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland

[8]Department of General Internal Medicine, Endocrinology and Diabetology, Helios Clinic Schwerin, Schwerin, Germany

[9]Institute for Endocrinology & Diabetes, University of Lübeck, Lübeck, Germany

[10]Specialized Diabetes Practice, Rosenheim, Germany

Correspondence
Julia M. Grimsmann, PhD, Institute of Epidemiology and Medical Biometry, ZIBMT, Ulm University, Albert-Einstein-Allee 41, D-89081, Ulm, Germany.
Email: julia.grimsmann@uni-ulm.de

Abstract

Aim: To cluster adults with diabetes using variables from real-world clinical care at manifestation.

Materials and Methods: We applied hierarchical clustering using Ward's method to 56 869 adults documented in the prospective Diabetes Follow-up Registry (DPV). Clustering variables included age, sex, body mass index (BMI), HbA1c, diabetic ketoacidosis (DKA), components of the metabolic syndrome (hypertension/dyslipidaemia/hyperuricaemia) and beta-cell antibody status. Time until use of oral antidiabetic drugs (OADs), use of insulin, chronic kidney disease (CKD), cardiovascular disease (CVD), retinopathy or neuropathy were assessed using Kaplan–Meier analysis and Cox regression models.

Results: We identified eight clusters: four clusters comprised early diabetes onset (median age 40-50 years) but differed with regard to BMI, HbA1c, DKA and antibody positivity. Two clusters included adults with diabetes onset aged in their early 60s who met target HbA1c, but differed in BMI and sex distribution. Two clusters were

---

characterized by late diabetes onset (median age 69 and 77 years) and comparatively low BMI, but differences in HbA1c. Earlier insulin use was observed in adults with high HbA1c, and earlier OAD use was observed in those with high BMI. Time until CKD or CVD was shorter in those with late onset, whereas retinopathy occurred earlier in adults with late onset and high HbA1c, and in adults with early onset, but high HbA1c and high percentage of antibody positivity.

Conclusions: Adult diabetes is heterogeneous beyond classical type 1/type 2 diabetes, based on easily available variables in clinical practice using an automated clustering algorithm that allows both continuous and binary variables.

KEYWORDS

adults, cluster analysis, diabetes, diabetes complications, diabetes registry, diabetes subgroups, DPV, hierarchical clustering, prospective diabetes follow-up, real-world data

1 | INTRODUCTION

Diabetes is very heterogeneous in terms of age at manifestation, clinical presentation, disease progression, co-morbidities and aetiology,[1,2] which may have implications for its treatment. However, up until today, classification is usually based on blood glucose, beta-cell antibodies, the patient's age and body mass index (BMI), family history and diabetic ketoacidosis (DKA).[3] Commonly, more than 90% of diabetes is classified as type 2 diabetes (T2D).[1]

In an attempt to refine diabetes classification and to personalize diabetes treatment, Ahlqvist et al. tried to identify subgroups of adults with diabetes using clinical features that were measured at diabetes manifestation, including an estimate of insulin secretion and insulin sensitivity based on fasting measurements of glucose and insulin (homeostatic model assessment [HOMA]2-B and HOMA2-IR). They found five replicable clusters of individuals with diabetes, which were significantly different in characteristics and the risk of complications. In support of the clustering, genetic associations in the clusters differed among those traditionally classified as T2D. The authors proposed that this substratification might help to guide optimal early treatment for patients who would benefit most, thereby representing a first step towards precision medicine in diabetes. Ideally, newly diagnosed individuals could be allocated to distinct clusters that may help to predict adverse events and the need for intensified treatment.[2]

Zaharia et al. applied the classification published by Ahlqvist et al. to adults with type 1 diabetes (T1D) or T2D from the German Diabetes Study and observed similar patterns.[4]

A Swedish study from Lugner et al. clustered individuals with clinical T2D from the Swedish National Diabetes Register using nine continuous variables at diabetes onset, but without inclusion of insulin secretion or insulin sensitivity. The authors found no optimal value for the number of clusters and concluded there would be no evidence for a specific number of subgroups within T2D.[5]

Fasting blood samples are not part of standard treatment at diabetes diagnosis in real-world diabetes care,[6] which makes the classification proposed by Ahlqvist et al. difficult to apply in routine clinical care. In the present study, we aimed to find an alternative clustering approach based on routine clinical data obtained at diabetes manifestation, avoiding the requirement of elaborate measures for insulin secretion and sensitivity. In addition, we chose a clustering algorithm that allowed not only for continuous, but also for dichotomous variables such as sex and antibody positivity. Clusters were then analysed by baseline characteristics, disease progression and treatment requirement.

2 | MATERIALS AND METHODS

In September 2020, the multinational Diabetes Follow-up Registry (DPV) consisted of 601 200 individuals with different types of diabetes from 503 participating centres. Detailed information on the documentation system has been published previously.[7] The protocol of DPV was approved by the ethics committee of Ulm University (approval no. 202/09), and data collection was approved by the local review boards at the participating diabetes centres. The registry was conducted in accordance with Good Epidemiological Practice,[8] as well as applicable regulatory and data protection requirements.

2.1 | Participants

We included individuals with a clinical diagnosis of T1D or T2D during 1995-2019, age 18 years or older at manifestation, available baseline data on BMI, HbA1c and at least one of the following variables: blood pressure, lipid values or uric acid (Figure 1).

2.2 | Variable definitions

Baseline data were aggregated ±3 months around diagnosis and included age, sex, BMI (kg/m^2), HbA1c, beta-cell antibody status
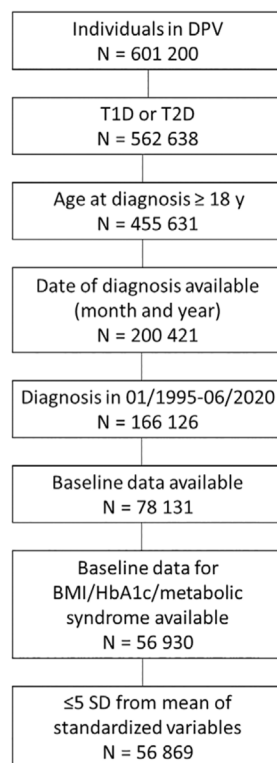
FIGURE 1 Patient selection. BMI, body mass index; DPV, Diabetes Follow-up Registry; T1D, type 1 diabetes; T2D, type 2 diabetes

(positive/negative), DKA, dyslipidaemia, hypertension and hyperuricaemia. HbA1c values were standardized according to the Diabetes Control and Complications Trial.[9,10] DKA was defined as pH less than 7.3 and/or bicarbonate less than 15 mmol/mol.[7] Dyslipidaemia was defined as total cholesterol more than 200 mg/dl and/or HDL cholesterol less than 35 mg/dl and/or LDL cholesterol more than 130 mg/dl and/or triglycerides more than 150 mg/dl (fasting) or more than 350 mg/dl (non-fasting) and/or use of lipid-lowering medication. Hypertension was defined as systolic blood pressure of 140 mmHg or higher and/or diastolic blood pressure of 90 mmHg or higher and/or antihypertensive treatment.[11] Hyperuricaemia was defined as uric acid of more than 6 and of more than 7 mg/dl in females and males, respectively,[12] and 'components of metabolic syndrome' was defined as having at least one of dyslipidaemia, hypertension or hyperuricaemia. In addition, we present '≥3 components of metabolic syndrome' (dyslipidaemia, hypertension, hyperuricaemia, carbohydrate metabolism disorder, obesity [BMI > 30 kg/m^2]).

Follow-up data on HbA1c were obtained after a median [interquartile range] diabetes duration of 1.8 [1.4-2.1] years. Long-term outcomes included time until insulin use, oral antidiabetic drug (OAD) use, chronic kidney disease (CKD), cardiovascular disease (CVD), diabetic retinopathy and diabetic neuropathy. The date of the respective event was extracted from the standardized documentation or by using the string search in free-text fields. CKD was defined as an estimated glomerular filtration rate (eGFR) of less than 60 ml/min/1.73m^2 calculated using the Modification of Diet in Renal Disease (MDRD) formula[13] and/or dialysis and/or transplant and/or albuminuria (urinary albumin excretion > 30 mg/L). CVD was defined as coronary artery disease and/or stroke and/or peripheral arterial occlusive disease and/or any operation on the heart. Diabetic retinopathy was assessed by ophthalmologists in accordance with published guidelines as described before,[14] and diabetic neuropathy was extracted from the string search in free-text fields using diagnosis of (poly)neuropathy or 'PNP' in combination with 'diabetic', or ICD codes (E10.4, E11.4, E12.4, E13.4, E14.4, G59.0, G63.2).

2.3 | Statistical analysis

We applied agglomerative hierarchical clustering to incorporate both continuous and dichotomous variables. In this approach, each observation starts in its own cluster. Similar clusters are merged stepwise until all observations are combined in one cluster, and then the optimal number of clusters is determined.[15] The baseline variables, age at manifestation, sex, BMI, HbA1c, DKA, beta-cell antibodies and 'components of the metabolic syndrome', were used for clustering, with continuous variables being standardized. We used Ward's Minimum Variance method for clustering. As this method is sensitive to outliers, we excluded individuals with standard deviation more than 5 from the mean to remove the most extreme observations.[2] We analysed three criteria to find the optimal number of clusters: Cubic Clustering Criterion (local peaks indicate good clusters), Pseudo-F statistic (comparatively large values indicate good numbers of clusters) and Pseudo-t^2 statistic (a good number of clusters is marked by the start of a peak when going from higher to lower numbers of clusters).[16] The analysis was carried out using SAS procedures PROC CLUSTER and PROC TREE (SAS version 9.4, TS1M7; SAS Institute, Cary, NC).

Baseline data were tabulated overall and by cluster using percentage for dichotomous variables and median [interquartile range] for continuous variables, as Kolmogorov–Smirnov tests indicated non-normality of all continuous variables (all P < .01) (Table 1). Unadjusted comparisons of outcomes between clusters were analysed using the Kruskal–Wallis test for continuous outcomes, and the chi-squared test for dichotomous outcomes. For the time until a certain event (e.g. CKD), we used Kaplan–Meier analysis and compared clusters using log-rank tests. Individuals without an event at the end of documentation were right-censored (Figure 2). Mean (± standard error) event-free survival time is presented by cluster (Figures 3 and 4); the respective survival curves are included as Figure S1 and S2. To account for differences in patient characteristics, we used Cox regression models adjusted for age, sex and use of OADs to analyse the risk of long-term complications. Results are given as hazard ratios with corresponding 95% confidence intervals (Table 1).

P values of unadjusted comparisons were adjusted for multiple testing using the Bonferroni–Holm method. Two-sided P values less than .05 were considered statistically significant.

TABLE 1 Demographics and outcomes overall and by cluster

	Overall (n = 56 869)	Cluster 1 (n = 6974)	Cluster 2 (n = 6824)	Cluster 3 (n = 6523)	Cluster 4 (n = 8464)	Cluster 5 (n = 7287)	Cluster 6 (n = 8329)	Cluster 7 (n = 3304)	Cluster 8 (n = 9164)	P value
Male sex (%)	58.5	58.1	70.3	67.2	58.0	67.9	66.1	35.3	38.3	<.001
Age at diagnosis (y)	58.9 [47.4-70.2]	39.9 [30.8-47.6]	48.8 [37.8-55.0]	45.7 [38.4-52.8]	62.7 [56.6-68.6]	61.0 [55.2-66.3]	69.0 [62.4-75.9]	48.6 [40.0-55.6]	76.6 [72.4-81.1]	<.001
BMI at onset (kg/m^2)	29.4 [25.7-34.0]	28.7 [25.0-31.7]	24.6 [21.9-27.2]	34.9 [31.2-38.6]	34.4 [31.8-37.3]	26.1 [24.2-27.9]	29.3 [26.0-32.1]	44.6 [40.6-49.3]	27.9 [25.0-30.8]	<.001
HbA1c at onset (%)	7.9 [6.4-10.6]	7.0 [6.1-8.5]	11.9 [10.6-13.6]	11.0 [10.0-12.1]	6.7 [6.1-7.6]	6.5 [5.9-7.4]	10.8 [9.5-12.2]	7.0 [6.2-8.4]	6.5 [6.0-7.2]	<.001
HbA1c at onset (mmol/mol)	62 [46-92]	53 [43-69]	107 [92-125]	97 [86-109]	50 [43-60]	48 [41-57]	95 [80-110]	53 [44-68]	48 [42-55]	<.001
DKA at onset (%)	3.6	2.4	13.9	6.1	0.1	0.1	3.3	2.2	1.9	<.001
Positive β cell AB (%)	5.0	5.6	20.6	6.1	0.4	6.4	0.6	3.0	0.0	<.001
≥3 components of the metabolic syndrome† (%)	56.5	39.4	29.2	74.5	70.7	37.8	68.9	81.6	58.5	<.001
Clinical type 1/2 diabetes (%)	13.2/86.8	31.5/68.5	44.8/55.2	11.9/88.1	2.0/98.0	9.7/90.3	3.9/96.1	2.3/97.7	1.9/98.1	
HbA1c at 2-y follow-up (%) (n)	6.5 [5.9 - 7.3] (n = 12 650)	6.4 [5.8 - 7.4] (n = 2008)	6.7 [6.0 - 7.9] (n = 1170)	6.9 [6.1 - 8.2] (n = 1251)	6.4 [5.9 - 7.1] (n = 2219)	6.3 [5.8 - 7.0] (n = 1967)	6.9 [6.1 - 7.8] (n = 1348)	6.4 [5.8 - 7.2] (n = 862)	6.4 [5.9 - 7.1] (n = 1825)	
Chronic kidney disease: HR (95% CI) ‡		Reference	**1.3 (1.2, 1.3)**	**1.6 (1.5, 1.8)**	**0.9 (0.9, 1.0)**	**0.8 (0.7, 0.8)**	**1.3 (1.2, 1.4)**	**1.3 (1.2, 1.4)**	1.0 (1.0, 1.1)	
Cardiovascular disease: HR (95% CI) ‡		Reference	1.1 (1.0, 1.3)	**1.3 (1.1, 1.5)**	1.0 (0.9, 1.1)	**1.2 (1.1, 1.4)**	**1.2 (1.0, 1.4)**	**1.3 (1.2, 1.5)**	**1.2 (1.0, 1.4)**	
Retinopathy: HR (95% CI) ‡		Reference	**2.2 (1.7, 2.8)**	**1.6 (1.2, 2.1)**	0.9 (0.7, 1.2)	1.0 (0.8, 1.4)	**1.9 (1.4, 2.5)**	0.7 (0.5, 1.0)	0.8 (0.6, 1.2)	
Neuropathy: HR (95% CI) ‡		Reference	**1.4 (1.3, 1.5)**	**1.6 (1.4, 1.7)**	**1.1 (1.0, 1.2)**	1.1 (1.0, 1.2)	**1.4 (1.3, 1.5)**	**1.3 (1.2, 1.4)**	1.0 (0.9, 1.1)	

Abbreviations: BMI, body mass index; DKA, diabetic ketoacidosis; HR, hazard ratio.

Data are presented as median [interquartile range], percentage or hazard ratio (HR) with 95% confidence interval. P values were obtained by Kruskal–Wallis or chi-squared tests. P values were adjusted for multiple testing using the Bonferroni–Holm method.

†Hypertension, dyslipidaemia, hyperuricaemia, carbohydrate metabolism disorder, obesity (BMI > 30 kg/m^2).

‡Obtained from Cox regression models adjusted for age, sex and use of oral antidiabetic drugs.

Significant HRs (P < .05) for comparison with cluster 1 are highlighted in bold.

FIGURE 2 Schematic characterization of the eight clusters. BMI, body mass index; CKD, chronic kidney disease; CVD, cardiovascular disease; T1D, type 1 diabetes
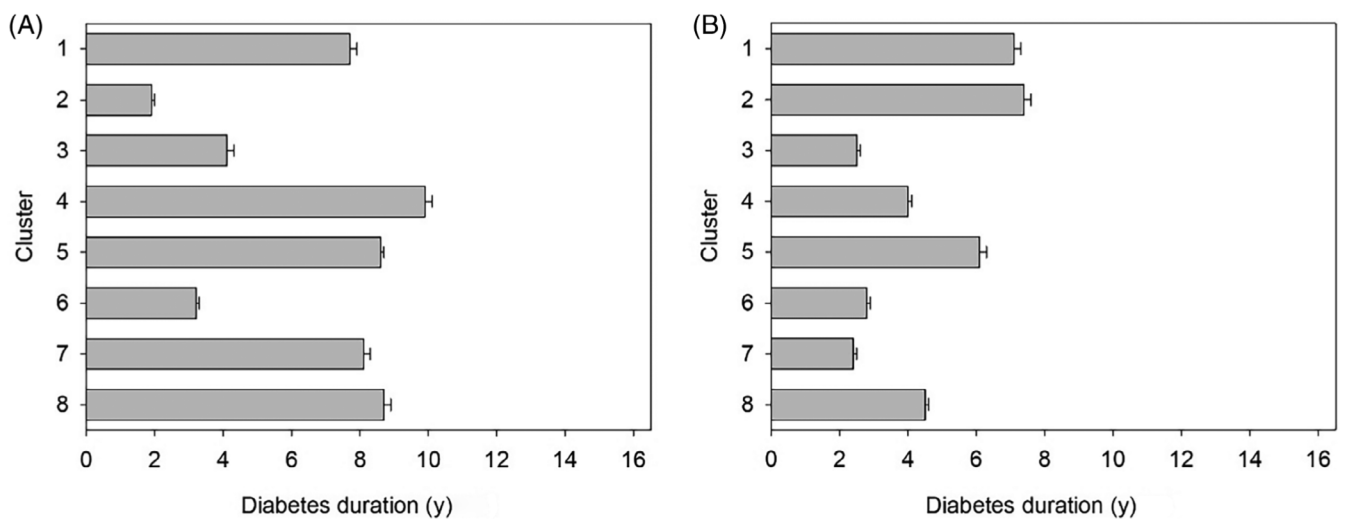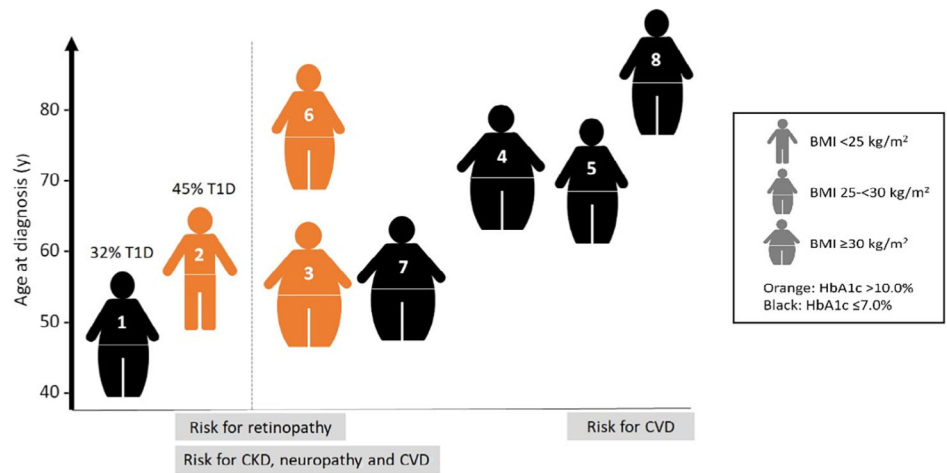

FIGURE 3 Mean estimated time until A, Use of insulin, and B, Use of oral antidiabetic drugs by cluster. Estimated time based on Kaplan–Meier analysis (unadjusted). Estimates are given as mean with SE

3 | RESULTS

Inclusion criteria were met by 56 869 individuals from 257 centres in Germany and 12 centres in Austria (Figure 1). Overall, 13% of all individuals included were clinically classified as T1D and 87% as T2D. Median age at diagnosis was 40 [31-48] years; 59% were men (Table 1).

3.1 | Cluster analysis

The Cubic Clustering Criterion had a local peak at eight clusters, whereas the PseudoF statistic showed a smooth curve without clearly indicating the optimal number of clusters. The Pseudo t^2 statistic indicated good clustering at eight or three clusters. Considering the three statistics together, we chose eight clusters that account for 54% of the variance.

A description of the seven clustering variables can be found in Table 1, and a schematic characterization of the clusters is presented in Figure 2. Cluster 1 (including 12% of the individuals) was characterized by early diabetes onset, comparatively low BMI and low HbA1c. Cluster 2 (12%) also had early diabetes onset and low BMI, but high HbA1c. Furthermore, cluster 2 included more men (70%), very high percentage of DKA (14%) and positive beta-cell antibodies (21%), and the lowest percentage of at least three components of the metabolic syndrome (29%). Cluster 3 (11%) was characterized by early diabetes onset, high BMI, high HbA1c and higher percentage of men. Individuals in cluster 4 (15%) were older at diabetes onset, had high BMI and low HbA1c. Cluster 5 (13%) was similar with regard to age and HbA1c, but had lower BMI than cluster 4 (26.1 [24.2-27.9] vs. 34.4 [31.8-37.3] kg/m^2), higher percentage of positive beta-cell antibodies (6% vs. <1%), lower percentage of '≥3 components of the metabolic syndrome' (38% vs. 71%) and included more men (68% vs. 58%). Both clusters 4 and 5 included almost no individuals with DKA at diabetes onset. Cluster
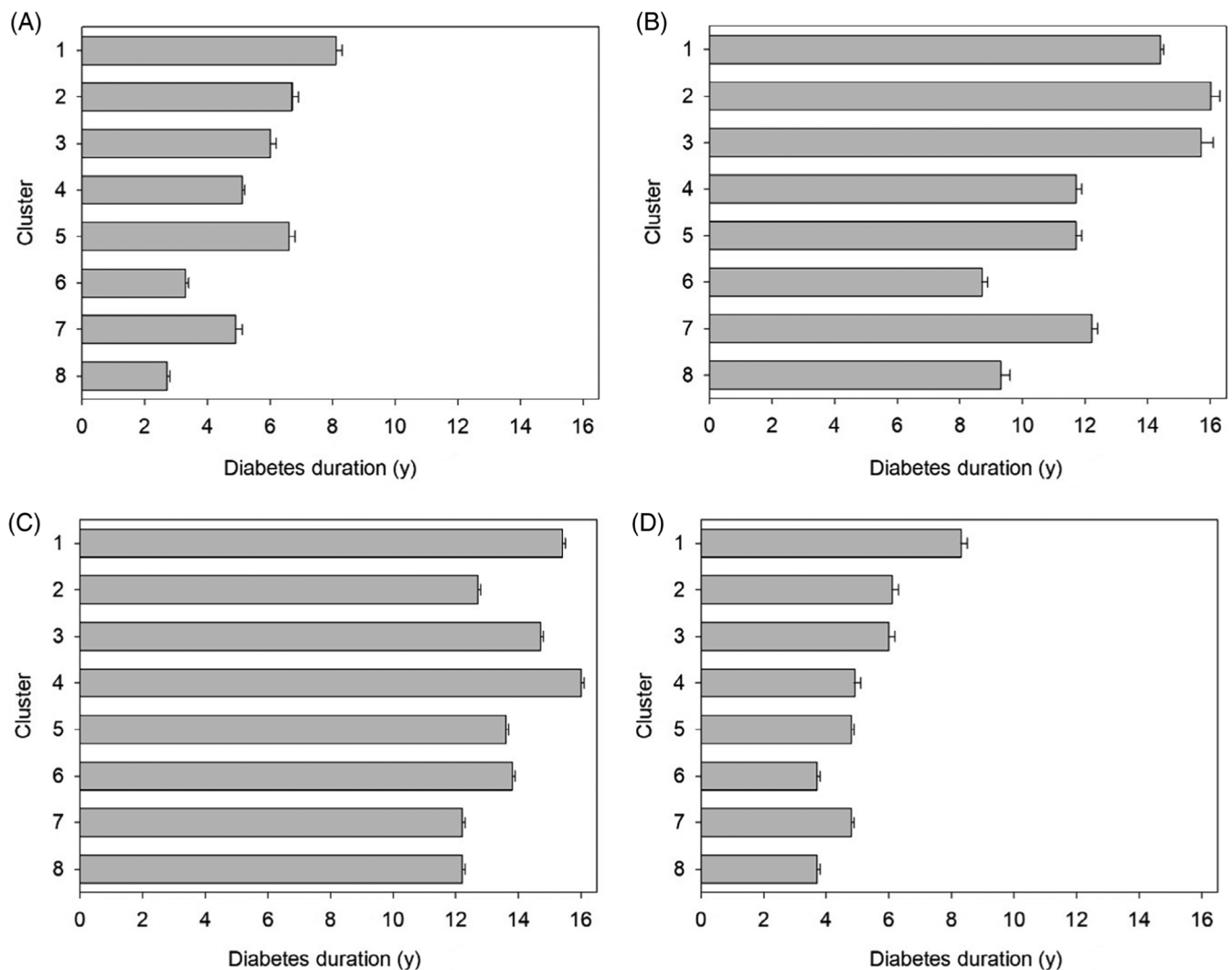
(A)

(B)

(C)

(D)

FIGURE 4 Mean estimated event-free survival time by cluster for A, Chronic kidney disease, B, Cardiovascular disease, C, Retinopathy, and D, Neuropathy. Estimated time based on Kaplan–Meier analysis (unadjusted). Estimates are given as mean with SE

6 (15%) was characterized by late diabetes onset, BMI in the upper overweight range, high HbA1c and a very low percentage of positive beta-cell antibodies (<1%). Cluster 7 was the smallest cluster (6%) and included individuals with early diabetes onset, low HbA1c, very high BMI (44.6 [40.6-49.3] kg/m^2) and a lower percentage of men (35%). This cluster had the highest percentage of individuals having '≥3 components of the metabolic syndrome' (82%). Cluster 8 (16%) had the latest diabetes onset (76.6 [72.4-81.1] years), comparatively low BMI and low HbA1c. In this cluster, 38% were men, and almost no one had positive beta-cell antibodies. Clusters 1 and 2 comprised the highest percentage of clinical T1D (32% and 45%, respectively), whereas clusters 4, 6, 7 and 8 comprised mostly clinical T2D (96% to 98%).

3.2 | HbA1c at 2-year follow-up by cluster

Two-year follow-up data on metabolic control were available for 12 650 individuals. HbA1c was low in all clusters with median values between 6.3% and 6.9% (45 and 52 mmol/mol); however, clusters with high baseline HbA1c (clusters 2, 3 and 6) still had higher HbA1c at follow-up than clusters with low baseline HbA1c (Table 1).

3.3 | Use of insulin and OADs by cluster

Kaplan–Meier analysis showed significant differences in mean estimated time until first documented use of insulin or OADs among clusters (Figures 3 and S1, both P < .001). Cluster 2 (highest baseline HbA1c, 45% clinical T1D) had the earliest usage of insulin with a mean (SE) time of 1.9 (0.1) years after diabetes onset, followed by clusters 3 and 6 (both with high baseline HbA1c). Time until OAD use was shortest in clusters with high baseline BMI and HbA1c (clusters 3 and 6), or HbA1c near target value but very high BMI (cluster 7). Latest OAD use was observed in clusters with a high proportion of clinical T1D diagnoses (clusters 1 and 2).

3.4 | Long-term complications by cluster

For CKD, CVD, retinopathy and neuropathy, the mean estimated event-free survival time was significantly different among the clusters (Figure 4, all P < .001). Time until CKD was shortest in cluster 8 (highest age at diabetes manifestation) and cluster 6 (higher age and high HbA1c at baseline). Clusters 4 and 5 with higher age but low HbA1c at baseline had a later diagnosis of CKD. The longest time until CKD was observed in cluster 1 (young age and low HbA1c at diabetes manifestation). A similar pattern was observed for time until CVD or neuropathy with the shortest time until an event in clusters 8 and 6, and a long time in cluster 1. Overall, retinopathy was less frequent than CKD and CVD. Around 10% of all individuals in clusters 1 and 7 had retinopathy 15 years after diabetes manifestation, whereas it was around 20% in clusters 2 and 6 (Figure S2).

The risk for long-term complications after adjustment for age, sex and use of OADs is presented in Table 1 and in the schematic characterization in Figure 2. A high risk for CKD, retinopathy and neuropathy was observed in clusters 3, 6 and 2 (high HbA1c at manifestation). Individuals in cluster 7 (very high BMI) had a high risk of developing CKD, CVD or neuropathy.

4 | DISCUSSION

In this large observational study, we found eight data-driven clusters of adults with diabetes based on routine clinical data collected at diabetes manifestation. These clusters were significantly related to therapeutic decisions and the development of diabetes complications. We consider the use of broadly available routine clinical data as essential when newly diagnosed individuals should be assigned to subgroups that may eventually be used to guide early treatment decisions. As patient heterogeneity is reflected by both continuous and categorical variables, we chose a hierarchical clustering approach accommodating both. We included age at onset, BMI and metabolic control, similar to previous studies,[1,2,17] but also sex, DKA at onset, beta-cell antibody positivity and components of the metabolic syndrome. We did not use pathophysiological variables like insulin sensitivity and insulin secretion as they are not easily available in daily routine care.

4.1 | Cluster characterization and comparison with clusters proposed by Ahlqvist et al.

Our clusters are not directly comparable with those proposed by Ahlqvist et al., which were mainly dependent on insulin secretion and insulin sensitivity. Ahlqvist et al. found five clusters, among them one cluster labelled 'severe autoimmune diabetes' (SAID, 6.4%) including all individuals with T1D or latent autoimmune diabetes in adults (LADA), defined by the presence of glutamic acid decarboxylase antibodies (GADA).[2,18] The authors had to prespecify this group in their k-means analysis based on results from the hierarchical

clustering approach. In our study, cluster 2 (12.0% of all individuals) resembled the SAID and 'severe insulin-deficient diabetes' (SIDD, 12.0%) clusters from Ahlqvist et al. with regard to early diabetes onset, low BMI and high metabolic control. In addition, cluster 2 included more men than women and a very high percentage of positive beta-cell antibodies and DKA, which was in accordance with the results of Ahlqvist et al., who evaluated DKA at manifestation after clustering.

Cluster 7 from our study (5.8%) was similar to Ahlqvist et al.'s 'mild obesity-related diabetes' (21.6%) cluster with early onset, high BMI and metabolic control near target level. In contrast to Ahlqvist et al., we found two additional clusters comprising individuals with early diabetes manifestation: cluster 1 (12.3%, low HbA1c and BMI) and cluster 3 (11.5%, high HbA1c and BMI). In addition, we observed sex differences between the three clusters, and adults in cluster 3 more often presented with DKA.

Cluster 4 resembled Ahlqvist et al.'s cluster labelled 'severe insulin-resistant diabetes' (14.9%) with late diabetes onset, high BMI and median HbA1c of ∼6.7% (50 mmol/mol). Two clusters were most similar to the 'mild age-related diabetes' (MARD) cluster with regard to low BMI and low metabolic control: individuals in cluster 5 were younger (61 [55-66] years) and more often male (68%), whereas cluster 8 comprised older age (77 [72-81] years) and fewer men (38%). We found one additional cluster with higher age (69 [62-76] years, cluster 6) that also had low BMI; however, metabolic control was high.

Of note, we did not detect a single cluster that comprised clinical T1D only, but three clusters (1-3) that included 32%, 45% and 12% of clinically diagnosed T1D, respectively.

4.2 | Association of clusters with long-term diabetes complications

We observed the highest risk for CKD and neuropathy in those clusters with median HbA1c more than 10% at manifestation (clusters 2, 3 and 6) and in cluster 7, characterized by the highest median BMI. Clusters 6 and 3 also had short time until retinopathy. Furthermore, short time until retinopathy was observed in cluster 2, in line with the SIDD cluster.[2,4,19]

In our study, time until CVD decreased with increasing age at manifestation for unadjusted data. After adjustment for age, sex and use of OADs, the risk for CVD was highest in clusters 3 and 7. Ahlqvist et al. observed a higher risk of coronary events and stroke in clusters with late diabetes onset for unadjusted data as well, but no significant difference after adjustment for age and sex.[2]

4.3 | Clustering variables, algorithms and reproducibility

Ahlqvist et al. used data from the All New Diabetics in Scania (ANDIS) database that documents more than 90% of all incident cases of

diabetes in Scania County in Sweden. They replicated the analysis in three other Swedish databases and one Finnish database.[2]

In our German/Austrian multicentre study including adults with T1D and T2D, we applied a different clustering algorithm based on a different set of clustering variables (hierarchical clustering using Ward's method, routine care variables) compared with Ahlqvist et al. (hierarchical clustering with log-likelihood as distance measure, k-means, including variables for insulin secretion and sensitivity). Some clusters resembled those from Ahlqvist et al., but others did not have a clear equivalent.

Slieker et al. performed clustering in GADA-negative individuals with T2D from ANDIS or two large cohorts from Scotland (GoDARTS) and the Netherlands (DCS). They used C-peptide as a proxy for beta-cell function and HDL cholesterol as a proxy for insulin sensitivity. In principle, their clusters mapped to those identified by Ahlqvist et al.; however, adding HDL cholesterol resulted in division of the MARD cluster into two clusters.[17]

Lugner et al. applied k-means clustering based on nine routine clinical variables (age, HbA1c, BMI, systolic and diastolic blood pressure, HDL and LDL cholesterol, triacylglycerol and eGFR) to adults with T2D from the Swedish National Diabetes Register. They were not able to find a specific number of clusters and hence concluded that the cluster division might be arbitrary.[5]

These discrepancies may raise the question of how many and which variables should be included for a clinically meaningful clustering, and whether the clustering variables must be adjusted to different populations. Moreover, clustering newly diagnosed individuals may lead to different results than clustering individuals with long-term diabetes. Zaharia et al. assigned 1105 individuals with newly diagnosed diabetes to the diabetes clusters proposed by Ahlqvist et al. While the pattern at baseline resembled that of Ahlqvist et al., cluster reproducibility reduced to 77% after 5 years of disease progression. The authors suspected that cluster membership could be affected by differences in treatment over time as well as by alterations in glucose homeostasis, triglycerides and liver steatosis.[4]

Furthermore, the choice of the classification algorithm is probable to affect results, as the algorithms differ, among other aspects, regarding distance metrics, cluster shape and type of variables. The k-means algorithm is not applicable to mixed-type data.[20] However, binary variables like sex or DKA at onset are known to have a strong association with HbA1c trajectory, use of medications[21] and long-term outcomes.[14,21] Our results based on hierarchical clustering indicate that they may also play a role in clustering individuals at diabetes manifestation, as we observed differences between clusters with respect to sex, antibody positivity and DKA. Ahlqvist et al. partly addressed this issue by stratifying their analyses by sex. However, GADA positivity could only be analysed in their hierarchical clustering approach, whereas for k-means the authors classified all GADA-positive individuals as one group before applying clustering to GADA-negative individuals only.[2] Another characteristic of the k-means algorithm is that it partitions observations into convex clusters, with data being split halfway between cluster means. Overlapping groups may therefore not be partitioned reliably.[20]

Other clustering approaches like k-prototype, an extension to the k-means algorithm for clustering mixed-type data, are able to overcome some of the drawbacks mentioned.[22] Alternatively, dissimilarity measures for mixed-type data can be created using Random Forest analysis and subsequently be used for various clustering algorithms.[23] Another way to detect subgroups is latent class growth modelling that can model distinct trajectories over time.[24] A previous DPV analysis in adults with T2D has identified four distinct HbA1c trajectories over 5 years, which differed with regard to age at diabetes manifestation, BMI, sex and insulin therapy.[21] In youth with T1D, five distinct HbA1c trajectories over 10 years were identified.[25] Also, a multi-trajectory approach identified five distinct curves of HbA1c, BMI and insulin dose in youth with T1D.[26] Cluster switchers, as observed by Zaharia et al., may indicate that clustering could benefit from including not only variables obtained at diagnosis, but also development over time (e.g. response to treatment after 6 months). Additional work is required to clarify which clustering approach, based on selection of variables as well as computer algorithm, is superior. Also, as proposed by Dennis et al.,[1] the predictive power of a clustering approach should be compared with a score based on various risk factors (e.g. the UKPDS risk engine[27] or the Framingham risk function[28]). In our view, data published to date do not show a superiority of the clustering approach over traditional categorization of diabetes heterogeneity.

5 | STRENGTHS AND LIMITATIONS

The strengths of our study include the large sample size of newly diagnosed adults with T1D or T2D and the long follow-up period to address complications. We used continuous as well as binary routine clinical variables for clustering; hence, the classification would be easily adaptable to other datasets without the need for fasting measurements. However, several limitations need to be addressed. The determination of the optimal number of clusters was not consistent for all three methods. In addition, hierarchical clustering using Ward's method tends to produce clusters of a similar size,[15] which may not fully reflect the underlying subgroups. Furthermore, diabetes therapy including medications and technology has changed during the long documentation period (1995-2019), which affects disease progression. We did not investigate genetic differences among clusters, hence clusters were built based on routine clinical variables only, but did not include disease pathogenesis. Data from our European study population are not necessarily generalizable to other cohorts, as diabetes phenotypes and drug responses can differ between ethnic groups.[29] Bancks et al.[30] used k-means clustering in individuals with South Asian, non-Hispanic White, Chinese, African American or Hispanic ethnicity and observed ethnic differences across the subgroups. They reported differences in BMI and age at diabetes onset, which affected subgroup assignment. This finding highlights the need for further studies that examine the applicability of the subgroup concept in different populations and ethnic groups. Future research may also lead to combination of clusters to make them more practical and

easier to follow. However, at this stage, we consider knowledge gain as the primary aim of clustering approaches.

Additionally, diabetes onset is difficult to determine for diabetes forms with slow progression and may depend on the frequency of routine laboratory tests. Therefore, different healthcare systems may include more or less early diagnosed individuals, which could lead to different results. Potential misdiagnosis of type of diabetes affected our analysis in only that way as only individuals with T1D (including LADA) or T2D were included, but no other types of diabetes. Misdiagnosis of, for example, LADA as T2D, would not affect clustering results, as type of diabetes was not included as a clustering variable. Misdiagnosis of maturity-onset diabetes of the young (MODY) would affect results; however, as absolute numbers of MODY are small, results are probable to change only marginally.

6 | CONCLUSIONS

Based on routine real-world clinical data at diabetes onset, hierarchical clustering results in distinct patient subgroups with different treatment and complication rates. However, clustering results should be interpreted cautiously, as they depend heavily on the choice of clustering variables. Independent confirmation of the subgroups, for example using genetic polymorphisms, is needed to account for the underlying pathophysiology. Large registries from different parts of the world are necessary to clarify these questions before a reclassification of diabetes is 'ready to use'.

AUTHOR CONTRIBUTIONS

PB, AF, JS, ML, SZ, SMM and MH contributed to data collection. SRT and RWH designed the analysis. JMG and SRT drafted the manuscript. JMG created the figures. SRT and RWH were responsible for the statistical analyses. All the authors contributed to the discussion and reviewed/edited the manuscript. SRT and RWH had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. All the authors approved the final manuscript submitted.

CONFLICT OF INTEREST

JS reports grants and personal fees from Abbott, AstraZeneca and Sanofi, outside the submitted work. PB reports to have received consultancy honoraria from Sanofi and Abbott. All the other authors have no conflicts of interest to declare.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/dom.14812.

DATA AVAILABILITY STATEMENT

Data available on request from the authors.

AVAILABILITY OF DATA AND MATERIAL

The datasets generated and analysed during the current study are not publicly available because of data privacy but are available from the corresponding author on reasonable request.

ORCID

Julia M. Grimsmann https://orcid.org/0000-0003-3627-1163
Sascha R. Tittel https://orcid.org/0000-0001-7913-6135
Jochen Seufert https://orcid.org/0000-0001-5654-7310

REFERENCES

1. 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;7(6):442-451.
2. Ahlqvist E, Storm P, Karajamaki 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;6(5): 361-369.
3. American Diabetes Association. 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes-2021. *Diabetes Care.* 2021;44(Suppl 1):S15-S33.
4. Zaharia OP, Strassburger K, Strom A, 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; 7(9):684-694.
5. Lugner M, Gudbjornsdottir S, Sattar N, et al. Comparison between data-driven clusters and models based on clinical features to predict outcomes in type 2 diabetes: Nationwide observational study. *Diabetologia.* 2021;64(9):1973-1981.
6. Petersmann A, Müller-Wieland D, Müller UA, et al. Definition, classification and diagnosis of diabetes mellitus. *Exp Clin Endocrin Diabetes.* 2019;127(S01):S1-S7.
7. Galler A, Tittel SR, Baumeister H, et al. Worse glycemic control, higher rates of diabetic ketoacidosis, and more hospitalizations in children,

adolescents, and young adults with type 1 diabetes and anxiety disorders. *Pediatr Diabetes*. 2021;22(3):519-528.

8. Hoffmann W, Latza U, Baumeister SE, et al. Guidelines and recommendations for ensuring good epidemiological practice (GEP): a guideline developed by the German Society for Epidemiology. *Eur J Epidemiol*. 2019;34(3):301-317.

9. The DCCT research group. Feasibility of centralized measurements of glycated hemoglobin in the diabetes control and complications trial: a multicenter study. *Clin Chem*. 1987;33(12):2267-2271.

10. DCCT Research Group. Diabetes control and complications trial (DCCT). *Update. Diabetes Care*. 1990;13(4):427-433.

11. van Mark G, Tittel SR, Sziegoleit S, et al. Type 2 diabetes in older patients: an analysis of the DPV and DIVE databases. *Ther Adv Endocrinol Metab*. 2020;11:2042018820958296.

12. Gois PHF, Souza ERM. Pharmacotherapy for hyperuricaemia in hypertensive patients. *Cochrane Database Syst Rev*. 2020;9:CD008652.

13. Schwandt A, Bergis D, Denkinger M, et al. Risk factors for decline in renal function among young adults with type 1 diabetes. *J Diabetes Complications*. 2018;32(10):940-946.

14. Hammes HP, Kerner W, Hofer S, et al. Diabetic retinopathy in type 1 diabetes-a contemporary analysis of 8,784 patients. *Diabetologia*. 2011;54(8):1977-1984.

15. Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning. Data Mining, Inference, and Prediction*. 2nd ed. New York, NY: Springer; 2009.

16. SAS Institute Inc. *SAS/STAT(R) 13.1 User's guide. Chapter 33: the CLUSTER procedure*. Cary, NC: SAS Institute Inc.; 2013.

17. Slieker RC, Donnelly LA, Fitipaldi H, et al. Replication and cross-validation of type 2 diabetes subtypes based on clinical variables: an IMI-RHAPSODY study. *Diabetologia*. 2021;64(9):1982-1989.

18. Ahlqvist E, Prasad RB, Groop L. Subtypes of type 2 diabetes determined from clinical parameters. *Diabetes*. 2020;69(10):2086-2093.

19. Tanabe H, Saito H, Kudo A, et al. Factors associated with risk of diabetic complications in novel cluster-based diabetes subgroups: a Japanese retrospective cohort study. *J Clin Med*. 2020;9(7):2083.

20. Bacher J. A probabilistic clustering model for variables of mixed type. *Qual Quant*. 2000;34(3):223-235.

21. Rathmann W, Schwandt A, Hermann JM, et al. Distinct trajectories of HbA1c in newly diagnosed type 2 diabetes from the DPV registry using a longitudinal group-based modelling approach. *Diabet Med*. 2019;36(11):1468-1477.

22. Huang Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mini Knowl Discovery*. 1998; 2(3):283-304.

23. Shi T, Horvath S. Unsupervised learning with Random Forest predictors. *J Comput Graph Stat*. 2006;15(1):118-138.

24. Nagin DS, Jones BL, Passos VL, Tremblay RE. Group-based multi-trajectory modeling. *Stat Methods Med Res*. 2018;27(7):2015-2023.

25. Sherr JL, Schwandt A, Phelan H, et al. Hemoglobin A1c patterns of youth with type 1 diabetes 10 years post diagnosis from 3 continents. *Pediatrics*. 2021;148(2):e2020048942.

26. Schwandt A, Kuss O, Dunstheimer D, et al. Three-variate longitudinal patterns of metabolic control, body mass index, and insulin dose during puberty in a type 1 diabetes cohort: a group-based multitrajectory analysis. *J Pediatr*. 2020;218:64-71.e3.

27. Kothari V, Stevens RJ, Adler AI, et al. UKPDS 60: risk of stroke in type 2 diabetes estimated by the UK prospective diabetes study risk engine. *Stroke*. 2002;33(7):1776-1781.

28. Anderson KM, Odell PM, Wilson PW, Kannel WB. Cardiovascular disease risk profiles. *Am Heart J*. 1991;121(1 Pt 2):293-298.

29. Herder C, Roden M. A novel diabetes typology: towards precision diabetology from pathogenesis to treatment. *Diabetologia*. 2022. doi: 10.1007/s00125-021-05625-x

30. Bancks MP, Bertoni AG, Camethon M, et al. Association of Diabetes Subgroups with Race/ethnicity, risk factor burden and complications: the MASASLA and MESA studies. *J Clin Endocrinol Metab*. 2021; 106(5):e2106-e2115.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Grimsmann JM, Tittel SR, Bramlage P, et al. Disease heterogeneity of adult diabetes based on routine clinical variables at diagnosis: Results from the German/Austrian Diabetes Follow-up Registry. *Diabetes Obes Metab*. 2022;1-10. doi:10.1111/dom.14812