

**Title:** Translating Subphenotypes of Newly Diagnosed Type 2 Diabetes from Cohort Studies to Electronic Health Records in the United States

**Author Names:** Zhongyu Li<sup>1</sup>, Star Liu<sup>2</sup>, Joyce C. Ho<sup>3,4</sup>, K.M. Venkat Narayan<sup>4,5</sup>, Mohammed K. Ali<sup>4,5,6</sup>, Jithin Sam Varghese<sup>4,5</sup>

**Author Affiliations:**

1 Nutrition and Health Sciences Doctoral Program, Laney Graduate School, Emory University, Atlanta, GA, USA

2 Biomedical Informatics and Data Science Doctoral Program, School of Medicine, Johns Hopkins University, Baltimore, MD, USA

3 Department of Computer Science, College of Arts and Sciences, Emory University, Atlanta, GA, USA

4 Emory Global Diabetes Research Center of Woodruff Health Sciences Center and Emory University, Atlanta, GA, USA.

5 Hubert Department of Global Health, Rollins School of Public Health, Emory University, Atlanta, GA, USA

6 Department of Preventive and Family Medicine, School of Medicine, Emory University, Atlanta, USA

**Corresponding Author:** Jithin Sam Varghese, Emory Global Diabetes Research Center, Woodruff Health Sciences Center, Emory University, Atlanta, GA, 30322, USA; Phone: +1 (404) 502-0415, E-mail: [jvargh7@emory.edu](mailto:jvargh7@emory.edu)

## ABSTRACT

Novel subphenotypes of type 2 diabetes mellitus (T2DM) are associated with differences in response to treatment and risk of complications. The most widely replicated approach identified four subphenotypes (severe insulin-deficient diabetes [SIDD], severe insulin-resistant diabetes [SIRD], mild obesity-related diabetes [MOD], and mild age-related diabetes [MARD]). However, the widespread clinical application of this model is hindered by the limited availability of fasting insulin and glucose measurements in routine clinical settings. To address this, we pooled data of adults ( $\geq 18$  years) with newly diagnosed T2DM from six cohort studies ( $n = 3,377$ ) to perform de novo clustering and developed classification algorithms for each of the four subphenotypes using nine variables routinely collected in electronic health records (EHRs). After operationalizing the classification algorithms on the Epic Cosmos Research Platform, we identified that among the 727,076 newly diagnosed diabetes cases, 21.6% were classified as SIDD, 23.8% as MOD, and 40.9% as MARD. Individuals classified as SIDD were more likely to receive insulin and incretin mimetics treatment and had higher risks for microvascular complications (retinopathy, neuropathy, nephropathy). Our findings underscore the heterogeneity in newly diagnosed T2DM and validated T2DM subphenotypes in routine EHR systems. This offers possibilities for the subsequent development of treatment strategies tailored to subphenotypes.

**Keywords:** precision medicine, subphenotypes, machine learning, type 2 diabetes

## 1 INTRODUCTION

2 Type 2 Diabetes Mellitus (T2DM) affects over 500 million people worldwide, including  
3 over 34 million adults in the United States (US).<sup>1,2</sup> Recent studies using unsupervised machine  
4 learning identified subphenotypes of T2DM with differences in pathophysiology, response to  
5 medication, and risk of microvascular complications.<sup>3,4,5</sup> Although several approaches were  
6 employed for subphenotyping of T2DM, the most widely replicated study from the Swedish All  
7 New Diabetics in Scandia (ANDIS) cohort identified four subphenotypes, namely Severe Insulin  
8 Deficient Diabetes (SIDD), Severe Insulin Resistant Diabetes (SIRD), Mild Obesity-related  
9 Diabetes (MOD) and Mild Age-related Diabetes (MARD).<sup>6-8</sup> To date, these subphenotypes have  
10 not yet been translated into health gains from better risk stratification through precision  
11 medicine.<sup>9,10</sup>

12 At least two gaps exist in clinical validation and subsequent translation to real-world  
13 clinical practice. First, not all variables used for subphenotyping are routinely collected during  
14 T2DM diagnosis at clinic visits.<sup>11</sup> A majority of studies that identified the ANDIS subphenotypes  
15 were conducted in observational cohort studies,<sup>12</sup> where homeostatic indices for insulin secretion  
16 and resistance that used fasting levels of c-peptide or insulin are collected.<sup>12</sup> As a result, it  
17 remains unknown whether these subphenotypes can be replicated using variables available in  
18 electronic health records (EHRs). Second, existing data-driven classification from cohort studies  
19 in European countries may not be generalizable across geographies and ethnic groups, including  
20 diverse populations of the US and Asia. For instance, studies in Asian cohorts have found  
21 differences from Europeans in their relative composition of phenotypes (higher proportions of  
22 SIDD), but also identified additional phenotypes such as Combined Insulin Deficient and  
23 Resistant Diabetes.<sup>7,13,14</sup> Additionally, there are variations in subphenotype characteristics even

24 among geographically and ethnically similar populations in Europe.<sup>6,15,16</sup> Therefore, there is a  
25 critical need to identify de-novo clusters in ethnically and socio-demographically diverse source  
26 populations from the US and subsequently enable validation in clinical settings.<sup>17</sup>

27 To address these knowledge gaps in subphenotyping newly diagnosed T2DM and study  
28 their epidemiology in EHRs, we implemented a two-step approach using pooled data from six  
29 diverse cohort studies and a large integrated EHR database in the US (**Figure 1**). First, we  
30 derived novel subphenotypes of newly diagnosed T2DM in cohort studies and developed  
31 classification algorithms for each subphenotype using variables routinely collected in EHRs.  
32 Second, we applied the classification algorithms for subphenotypes in EHRs and studied the time  
33 to the prescription of pharmacological treatments and the incidence of microvascular  
34 complications.

## 35 **RESULTS**

### 36 **Participant Characteristics in the Pooled Cohorts**

37 We pooled data of newly diagnosed T2DM from six US cohort studies: the  
38 Atherosclerosis Risk in Communities (ARIC) Study, Coronary Artery Risk Development in  
39 Young Adults Study (CARDIA), Diabetes Prevention Program (DPP) and Diabetes Prevention  
40 Program Outcomes Study (DPPOS), Jackson Heart Study (JHS), and the Multi-Ethnic Study of  
41 Atherosclerosis (MESA). These cohorts enrolled socio-demographically and geographically  
42 diverse populations with varying clinical profiles (**Supplementary Table 1**), enabling us to  
43 capture a broad spectrum of diabetes presentations.

44 Among these studies, the DPPOS is the observational follow-up to the randomized  
45 controlled trial, DPP. We excluded participants from the intervention arms of the DPP for the

46 main analysis. This exclusion was done to minimize the potential effects of exposure to diabetes  
47 interventions, such as lifestyle modifications or medications, on the biomarkers used in our  
48 classification algorithms. Besides, to ensure the independence of the JHS cohort and avoid  
49 double counting, we excluded participants who were also recruited in the ARIC study from the  
50 JHS cohort. All studies used internally standardized protocols to measure anthropometry and  
51 biomarkers (**Supplementary Table 2**), routinely assessed dysglycemia in their participants, and  
52 had a low risk of left censoring of newly diagnosed T2DM.

53 After harmonizing definitions across six U.S.-based cohort studies (Supplementary Table  
54 3), we identified 7,623 participants with newly diagnosed T2DM. We then excluded individuals  
55 missing key biomarkers, including body mass index (BMI), age at diagnosis, and glycated  
56 hemoglobin (HbA1c), as well as those with implausible homeostatic assessments at diagnosis (n  
57 = 13). This resulted in an analytic sample of 3,377 newly diagnosed T2DM cases for developing  
58 the classification algorithm in the pooled cohort dataset (**Supplementary Figure 1**).

59 Descriptive characteristics of the pooled cohort sample and each cohort study are  
60 presented in **Table 1** and **Supplementary Table 4**. The average age of participants at diagnosis  
61 of T2DM was 63.4 years (SD: 12.4), with younger ages at diagnosis in CARDIA (48.8 years  
62 [SD: 4.9]) and older age at diagnosis in ARIC (75.3 years [SD: 5.1]). The overall pooled cohort  
63 were 59.8% female, 49.9% White, 35.5% Black and 14.6% from other racial groups with  
64 average HbA1c of 6.3% (IQR: 5.8-6.7%) and BMI of 33.2 kg/m<sup>2</sup> (SD: 7.0). ARIC was  
65 predominantly White (72.1%) while CARDIA and JHS had higher proportions of Black  
66 participants (70.6% and 100% respectively) (**Supplementary Table 4**). The pooled cohort had  
67 high median values of insulin resistance and beta cell function, defined using homeostatic model  
68 assessment indices of HOMA2-IR (2.8 [interquartile range: 1.7-4.7]) and HOMA2-B% (108%

69 [IQR: 73.3-156.4]). There was variability across the six cohort studies for all metabolic  
70 biomarkers (range of median values; systolic blood pressure [SBP]: 106.23 to 129.6 mmHg,  
71 diastolic blood pressure [DBP]: 66.8 to 99.5 mmHg, LDL cholesterol: 97.3 to 121.5 mg/dL,  
72 HDL cholesterol: 42.1 to 49.5 mg/dL, triglycerides: 128.8 to 175.3 mg/dL, triglyceride-to-HDL  
73 ratio: 2.1 to 3.6).

#### 74 **De-novo clustering of Subphenotypes of Incident Diabetes in the Pooled Cohorts**

75 Consistent with the other studies of subphenotypes of newly diagnosed T2DM, we  
76 conducted a de-novo hierarchical clustering and k-means clustering utilizing five key variables:  
77 age at diagnosis, BMI, HbA1c (%), HOMA2 %B, and HOMA2 IR. Both approaches identified  
78 an optimal solution of four clusters (**Supplementary Figure 2**). The clusters identified through  
79 k-means were subsequently labeled according to their similarity in the distribution of the five  
80 clustering variables (**Figure 1**) to those subphenotypes described in the original ANDIS study:  
81 Mild Obesity-Related Diabetes (MOD), Mild Age-Related Diabetes (MARD), Severe Insulin-  
82 Deficient Diabetes (SIDD), and Severe Insulin-Resistant Diabetes (SIRD). A sensitivity analysis  
83 showed that individuals were predominantly classified into the same four subphenotypes when  
84 using cluster centroids from the ANDIS study (**Supplementary Table 5**).

85 In the pooled cohort sample, MARD was the most common subphenotype, representing  
86 44.5% of those with newly diagnosed T2DM, followed by MARD (37.5%), SIRD (14.5%) and  
87 SIDD (3.2%). Cases identified as MARD were older (72.5 years [SD: 7.6]) with lower BMI  
88 (29.4 kg/m<sup>2</sup> [SD: 4.6]) and HOMA2-IR (1.9 [IQR: 1.3-2.9]). MOD was characterized by  
89 younger age (52 years [SD: 7.9]), higher BMI (37.9 kg/m<sup>2</sup> [SD: 7.0]) and a high HOMA2-IR (3.7  
90 [IQR: 2.5-5.2]) (**Table 1**). The majority of newly diagnosed T2DM in CARDIA (83.8%), DPP  
91 (79.8%), DPPOS (62.1%), and JHS (63.7%) were identified as MOD (**Supplementary Table 4**).

92 Nearly all newly diagnosed cases in ARIC (91.3%) were identified as MARD. The SIDD and  
93 SIRD subphenotypes had worse cardiometabolic profiles, compared to MOD and MARD. SIRD  
94 was characterized by high BMI (32.4 kg/m<sup>2</sup> [SD: 5.8]), HOMA2-B% (303 [IQR: 250.6-374.8])  
95 and HOMA2-IR (9.2 [IQR: 7.4-11.8]). SIRD also had a later age at diagnosis (65.7 years [SD:  
96 9.4]) than MOD and SIDD but earlier than MARD. SIDD was marked by early age at onset (61.1  
97 years [SD: 13.7]), high HbA1c (9.4% [IQR: 8.7-10.4]) and poor beta-cell function (HOMA2-  
98 B%: 41.9 [IQR: 21.9-72.8]).

99 To evaluate the robustness of the clustering algorithms to pooling cohort studies, we first  
100 examined whether the inclusion of any single large cohort meaningfully influenced the results.  
101 We calculated the Adjusted Rand Index (ARI) and Cohen's  $\kappa$  between the original classification  
102 from the pooled cohort studies and those obtained using a leave-one-cohort-out approach. Both  
103 the ARI and Cohen's  $\kappa$  values suggest a high level of stability and concordance (ARI  $\geq$  0.96,  $\kappa \geq$   
104 0.98) in diabetes subphenotype clustering when excluding CARDIA, DPP, DPPOS and JHS, and  
105 moderate stability when excluding ARIC (ARI = 0.49,  $\kappa$  = 0.69) and MESA (ARI = 0.63,  $\kappa$  =  
106 0.68) (**Supplementary Table 5**). Second, to examine if the imputation of missing clustering  
107 variables (i.e., blood pressure, LDL cholesterol, HDL cholesterol, triglycerides) using k-nearest  
108 neighbors imputation influenced the results, we conducted a de-novo clustering using complete  
109 cases (**Supplementary Table 7**). Subphenotypes identified using the pooled analytic sample (N=  
110 3,377) and the complete case sample (N = 2,775) displayed high concordance (**Supplementary**  
111 **Table 8**).

## 112 **Classification of Subphenotypes Using Routine Clinical Variables from the EHR**

113 Replication studies of the four subphenotypes are usually conducted using HOMA2  
114 indices, which are based on fasting insulin or C-peptide levels. However, these measurements are

115 not commonly collected in routine clinical care (**Supplementary Table 9**), thereby limiting the  
116 validation and translation of these novel subphenotypes to clinical practice. To address this  
117 challenge, we explored alternative methods to classify T2DM subgroups using only routinely  
118 available clinical variables. First, we conducted unsupervised k-means clustering of the analytic  
119 sample using biomarkers available in EHRs, namely age of diagnosis, HbA1c, BMI, blood  
120 pressure, HDL, LDL, triglycerides, and triglycerides to HDL ratio (a proxy for insulin  
121 resistance).<sup>18</sup> However, this approach yielded clusters that exhibited low concordance with the  
122 four ANDIS subphenotypes identified in our de-novo clustering analysis (**Supplementary Table**  
123 **10**).

124         Next, we sought to refine the classification process by developing and validating four  
125 One-vs-All logistic regression models using the same routinely available clinical variables. This  
126 supervised learning approach allowed us to model an individual's probability of membership in  
127 each subphenotype. Each model was fitted on a 70% training subset ( $n = 2,363$ ) of the pooled  
128 cohort dataset and subsequently validated on a 30% held-out test dataset ( $n = 1,014$ ). To  
129 determine the optimal cutoff to predict subphenotype membership, we employed a fivefold  
130 cross-validation approach and selected the predicted probability threshold that maximizes the F1  
131 score. These thresholds were validated on the held-out test dataset to evaluate discrimination for  
132 each subphenotype. Key indicators of discrimination for each model in the training dataset and  
133 test dataset, i.e. sensitivity, specificity, area under the receiver operating characteristic curve  
134 (AUC), positive predictive value (PPV), negative predictive value (NPV), and F1 score, are  
135 presented in **Table 2**. Coefficients and standard errors for the logistic regression models are  
136 presented in **Supplementary Table 11**.



137           The models demonstrated strong overall performance, with minimal loss of  
138 discrimination between the training and test datasets (**Supplementary Table 12**). The optimal  
139 predicted probability thresholds for best classification based on the five-fold cross-validation  
140 were 0.16, 0.38, and 0.32, respectively, for SIDD, MOD, and MARD. In particular, the model  
141 for classifying SIDD exhibited high discrimination: AUC (0.99, 95% CI = 0.99, 1.00), sensitivity  
142 (0.94), specificity (0.99), and F1 score (0.90). Similarly, the models for MOD and MARD  
143 showed high sensitivity of 0.93 and 0.95, with F1 scores of 0.88 and 0.87, respectively, though  
144 they had lower specificity (MOD: 0.89, MARD: 0.81) compared to the SIDD model. The  
145 classification model for SIRD showed lower performance across all metrics, with an AUC of  
146 0.62 (95% CI = 0.58, 0.67) and F1 score of 0.28. If an individual's membership probability in a  
147 subphenotype was higher than the optimal threshold, we classified them sequentially as SIDD,  
148 MARD, and MOD based on the known risk of complications. Given lower discrimination  
149 indices for the SIRD model in reliably distinguishing this subphenotype from others, individuals  
150 not classified into any of the other three groups were categorized as "Unclassified". Alternative  
151 classification approaches for subphenotypes (multinomial regression, random forests) did not  
152 improve discrimination (**Supplementary Table 13**).

### 153 **Epidemiology of Diabetes Subphenotype**

154           Based on the high overall discrimination of SIDD, MARD and MOD in our pooled  
155 cohort analysis and previously reported associations of SIDD with a higher risk of diabetic  
156 complications, we subsequently applied the classification models to newly diagnosed T2DM  
157 cases (n = 727,076) identified in the Epic Cosmos Research Platform. This approach enabled us  
158 to characterize the epidemiology of subphenotypes in a diverse, real-world patient population,  
159 providing critical insights into its demographic, clinical, and treatment profiles. Newly diagnosed

160 T2DM between January 2012 and December 2023 were identified using the SUPREME-DM  
161 computable phenotype. The newly diagnosed cases were on average 64.4 years (SD = 13.3), 52%  
162 female, 68% Non-Hispanic [NH] White (68%) 17% NH Black, 7% Hispanic and 8.6% NH  
163 Other, with an average BMI of 33.2 kg/m<sup>2</sup> (SD: 6.8) and median HbA1c of 7.0% (IQR: 6.6-7.9).  
164 Median values of BMI, HbA1c and other biomarkers were similar to those observed in the  
165 pooled cohort data.

166 Applying logistic regression to predict subphenotype membership in Epic Cosmos  
167 identified 156,951 individuals (21.6%) as SIDD, 172,922 (23.8%) as MOD, and 297,568  
168 (40.9%) as MARD. The average age at diagnosis was least for MOD (51.9 years [SD: 9.4]),  
169 followed by SIDD (59.8 years [SD: 13.5]), the unclassified group (62.9 years [SD: 5.7]), and  
170 MARD (74.8 years [SD: 7.9]) (**Table 2**). MOD had the highest average BMI (38.7 kg/m<sup>2</sup> [SD:  
171 6.1]), while MARD had the lowest average BMI (29.4 kg/m<sup>2</sup> [SD: 4.0]). SIDD had the highest  
172 average HbA1c at diagnosis (9.5% [IQR: 8.7-11.0]), meaningfully higher than other  
173 subphenotypes, which ranged from 6.7% (MARD) to 7.1% (Unclassified). The unclassified  
174 group had a cardiometabolic profile resembling that of the MOD group, particularly in terms of  
175 BMI, blood pressure, and triglyceride to HDL cholesterol ratio. The social vulnerability index  
176 was higher for SIDD (64.1 [IQR: 37.0, 84.6]) and MOD (62.8 [IQR: 35.6, 83.8]) compared to  
177 MARD (57.3 [IQR: 31.1, 79.6])

178 The distribution of SIDD, MOD, and MARD among newly diagnosed type 2 diabetes  
179 cases across the US shows regional variations (**Figure 3**). For instance, the District of Columbia  
180 (27.8%) and South Carolina (26.1%) had the highest proportions of cases classified as SIDD,  
181 while Idaho (18.7%) and Kansas (18.9%) had the lowest proportions (**Figure 3, Panel A**).  
182 MARD is widespread across much of the U.S., with high prevalence in the Midwest (South

183 Dakota: 49.4%), and the Northeast (Rhode Island: 46.6%) (**Figure 3, Panel B**). Utah showed the  
184 highest prevalence of MOD (29.2%), followed by Colorado (28.7%) and Alaska (27.4%)  
185 (**Figure 3, Panel C**).

186 We also observed regional variations in the proportions of SIDD and MOD by race &  
187 ethnicity, although we did not conduct statistical tests for these differences. For example, a high  
188 proportion of non-Hispanic Black and Hispanic patients were classified as SIDD ( $\geq 20\%$ ) and  
189 MOD ( $\geq 30\%$ ) across several states (**Supplementary Figure 5, Panels A and C**). In contrast, the  
190 proportion of MARD was highest among non-Hispanic White patients (**Supplementary Figure**  
191 **5, Panel B**).

## 192 **Time to pharmacological treatments**

193 We utilized longitudinal EHRs and cumulative incidence functions to estimate the time to  
194 prescription of glucose-lowering medications (**Figure 4**). Hazard ratios (HR), adjusting for sex  
195 and age at diagnosis, were estimated using Cox proportional hazard models to study the time to  
196 prescription of insulin, metformin, and incretin mimetics (GLP-1 RA or GLP-1 RA/GIP) with  
197 subphenotype membership, relative to those classified as MOD (**Supplementary Table 14**).  
198 Relative to MOD in the first 60 months after diagnosis, prescription of insulin (adjusted HR:  
199 1.65; 95% CI: 1.62, 1.67) and incretin mimetics (adjusted HR: 1.22; 95% CI: 1.20, 1.24) were  
200 earlier among SIDD. SIDD were also less likely to receive metformin (adjusted HR: 0.92,  
201 95% CI: 0.91-0.94). Relative to MOD, initiation of prescription of insulin (adjusted HR: 0.87;  
202 95% CI: 0.86, 0.89), metformin (adjusted HR: 0.81; 95% CI: 0.80, 0.83), and incretin mimetics  
203 (adjusted HR: 0.42; 95% CI: 0.41, 0.43) were later among MARD. Those unclassified were less  
204 likely to receive insulin (adjusted HR: 0.96, 95% CI: 0.94, 0.98) and incretin mimetics (adjusted

205 HR: 0.84, 95% CI: 0.82, 0.85) but were more likely to receive metformin (adjusted HR: 1.08,  
206 95% CI: 1.06, 1.10).

## 207 **Time to microvascular complications**

208 We estimated the time to microvascular complications based on ICD-10-CM codes among  
209 patients free of these complication at diagnosis. SIDD exhibited a higher cumulative incidence of  
210 microvascular complications ten years after diagnosis (**Figure 4 and Supplementary Table 14**).  
211 Relative to MOD in the first 10 years after diagnosis, SIDD were more likely to develop  
212 retinopathy (adjusted HR: 2.83, 95% CI: 2.73, 2.93), neuropathy (adjusted HR: 1.57, 95% CI:  
213 1.54, 1.60) and nephropathy (adjusted HR: 1.34, 95% CI: 1.32, 1.37). Relative to MOD in the  
214 first 10 years after diagnosis, MARD were more likely to develop retinopathy (adjusted HR:  
215 1.21, 95% CI: 1.16, 1.26) but less likely to develop neuropathy (adjusted HR: 0.90, 95% CI:  
216 0.88, 0.92) and nephropathy (adjusted HR: 0.95, 95% CI: 0.93, 0.97). Those unclassified were  
217 more likely to develop retinopathy (adjusted HR: 1.43, 95% CI: 1.37, 1.49), neuropathy  
218 (adjusted HR: 1.10, 95% CI: 1.08, 1.13) and nephropathy (adjusted HR: 1.03, 95% CI: 1.01,  
219 1.05). When comparing SIDD to all non-SIDD, the former were 2.41 times (95% CI: 2.36, 2.47),  
220 1.61 times (95% CI: 1.59, 1.63) and 1.37 times (95% CI: 1.35, 1.38) more likely to develop  
221 retinopathy, neuropathy and nephropathy (**Supplementary Figure 6**).

## 222 **DISCUSSION**

223 In this study, we successfully translated subphenotypes first identified in European  
224 cohort studies to a diverse EHR population from the US. By identifying four subphenotypes  
225 from newly diagnosed T2DM within pooled cohort studies from the United States, we developed  
226 and validated reliable classification models for detecting the novel subphenotypes of SIDD,

227 MOD, and MARD. This replication and validation of diabetes subphenotypes in routine clinical  
228 care are critical steps towards successful translation of precision medicine into practice, since  
229 subphenotypes reflect variations in pathophysiology, treatment responses, and complication  
230 risks. The findings from this study have significant implications for diabetes surveillance and  
231 health policy.

232         Leveraging the largest integrated EHR database in the world, we revealed novel insights  
233 into the epidemiology of novel subphenotypes in the US. We estimated that one in five cases of  
234 newly diagnosed T2DM belong to the SIDD subphenotype, with higher proportions among racial  
235 and ethnic minorities. Estimates remain consistent with prior studies using smaller samples of  
236 new and previously diagnosed T2DM using national surveys and with those from a university  
237 health system.<sup>17,19</sup> Furthermore, the prevalence of MARD and MOD subphenotypes in our study  
238 aligns with the patterns observed in other studies globally, where MARD is the most common  
239 subphenotype, particularly in non-Hispanic white populations, while MOD is more prevalent  
240 among younger individuals and those with higher body mass across racial and ethnic groups.<sup>7</sup>

241         Treatment patterns across subtypes revealed opportunities for optimizing diabetes care,  
242 prioritizing resource-limited treatments to high-risk groups. Notably, patients classified as SIDD  
243 were more likely to receive insulin and incretin mimetics (GLP-1 RA or GLP-1 RA/GIP) earlier  
244 than other subphenotypes during the follow-up, consistent with the ANDIS study and the  
245 American Diabetes Association's recommendations for management of severe hyperglycemia.<sup>20</sup>  
246 A recent randomized controlled trial demonstrated that stratifying patients into subtypes  
247 translated to tailored therapies that improved responses to semaglutide and dapagliflozin, with  
248 SIDD showing greater reductions in HbA1c and improved postprandial glucose control,  
249 highlighting the value of precision medicine in diabetes care<sup>21</sup>. Although current approaches in

250 clinical practice may prioritize those with insulin deficiency (SIDD), nearly half of SIDD cases  
251 were not treated with therapy appropriate to insulin deficiency within the first five years of  
252 diagnosis, indicating a gap in proper and timely treatment for this high-risk diabetic subgroup.

253         Conversely, overtreatment should be carefully managed, particularly among older adults  
254 in the mild age-related diabetes (MARD) subphenotype. Recent evidence suggests that  
255 overtreatment in older multimorbid patients can lead to adverse outcomes, such as higher  
256 mortality rates, without significant benefits in hospitalization or functional decline<sup>22</sup>. In our  
257 study, we observed that more than 20% of MARD received insulin treatment within the first two  
258 years of diagnosis. Alternatives such as GLP1-RAs may minimize the risk of falls from insulin-  
259 induced hypoglycemia in this population, provided concerns of loss of muscle mass and bone  
260 density are adequately studied.<sup>23</sup> These results, therefore, underscore the need for research into a  
261 tailored treatment paradigm to address the unique T2DM management goals for each  
262 subphenotype.

263         We also observed differential risks of microvascular complications across subphenotypes,  
264 consistent with previous studies, highlighting the potential of precision prognosis for  
265 personalized microvascular complication prevention. SIDD exhibited a significantly higher  
266 incidence of diabetic retinopathy and neuropathy.<sup>6,15</sup> However, the differences in cumulative  
267 incidence of nephropathy among SIRD, MOD and MARD were less distinguishable, relative to  
268 other studies.<sup>6,24</sup> For instance, the risk of chronic kidney disease and albuminuria for SIDD was  
269 lower than SIRD and higher than MOD in the ANDIS study but was similar to MOD in the  
270 ADOPT trial.<sup>6,24</sup>

271         Findings from this study also have implications for diabetes surveillance and health  
272 policy. Current efforts towards achieving precision in public health skew towards genomics.

273 Although useful for characterizing the burden and tailoring interventions for some forms of  
274 diabetes like monogenic diabetes of the young and potentially type 1 diabetes, genomics may be  
275 limited in its ability to stratify risk of complications after type 2 diabetes, a polygenic disease  
276 with only 19% of heritability explained by genetics.<sup>27–29</sup> Most of the variability in treatment  
277 outcomes is, therefore, likely driven by structural socio-economic factors. For instance, we  
278 observed higher proportions of the SIDD subphenotype among non-Hispanic Black, Hispanic,  
279 and non-Hispanic Other adults, who also experience a higher prevalence of T2DM and worse  
280 social determinants of health. Geographic variability in SIDD proportions within racial and  
281 ethnic groups suggests the need for further investigation into the role of specific environmental  
282 factors and gene-environmental interactions contributing to the higher risk of SIDD in these  
283 populations. Additionally, individuals classified as SIDD and MOD reported higher social  
284 vulnerability, further complicating their diabetes management. Such insights could guide  
285 targeted interventions that address not only the biological underpinnings of diabetes but also the  
286 socio-economic factors influencing health outcomes.

287 This study has several strengths. First, we used data from rigorously conducted cohort  
288 studies with a low risk of undiagnosed diabetes or missing data due to left censoring of newly  
289 diagnosed T2DM. Second, we applied harmonized definitions across cohort studies, combining  
290 self-reported data and glycemic biomarkers to minimize information bias. Third, we utilized a  
291 validated computable phenotype of newly diagnosed T2DM and characterized the epidemiology  
292 and progression of a novel subphenotype, leveraging the largest integrated EHR database in the  
293 United States to characterize the epidemiology and progression of a novel subphenotype. Finally,  
294 our simple classification algorithm utilizing commonly available clinical measurements enables  
295 the clinical translation of our findings, facilitating practical application in routine clinical settings.

296           This study also has several limitations and challenges regarding subtyping diabetes for  
297 clinical validation. First, we excluded half of the newly diagnosed cases from the pooled cohort  
298 studies due to missing data, largely because of the absence of key biomarkers such as HbA1c in  
299 the earlier study visits in those cohort studies. While the analytic sample may not capture the full  
300 variability in newly diagnosed T2DM, the excluded and analytic samples were similar in  
301 distributions. Additionally, our methodological approach was designed to be robust against non-  
302 representativeness relative to the target population of all new T2DM cases. The data from six  
303 geographically diverse cohorts likely captures a broad phenotypical representation of diabetes  
304 cases. Moreover, applying our classification algorithm to a more representative large EHR  
305 dataset validated its robustness and enabled a nationally representative characterization of  
306 diabetes subphenotypes. Second, the SUPREME-DM computable phenotype does not specify the  
307 sequence of criterion (labs, medications, diagnostic codes) to meet for a case to be classified as  
308 incident T2DM in electronic health records. Therefore, nearly 1 in 3 cases and 1 in 6 cases had a  
309 history of insulin use and diabetic nephropathy, respectively, on or before the inclusion date in  
310 the analytic sample and were therefore excluded from the analysis of cumulative incidence.  
311 Third, the pooled cohorts might not fully represent the heterogeneity in diabetes within the US,  
312 particularly in its representation of other minorities such as Alaska Natives, Native Americans,  
313 South Asians and Pacific Islanders. Besides, although Epic Cosmos includes data from 250  
314 million individuals in the US, only half the healthcare systems use Epic software.<sup>30</sup> Nevertheless,  
315 a comparison of socio-demographic characteristics between Epic Cosmos and the US Census  
316 suggests similarities in socio-demographic characteristics.<sup>31</sup> Finally, the unavailability of fasting  
317 insulin and fasting glucose in the EHRs prevented us from identifying SIRD. We hypothesize  
318 that distinguishing among the SIRD, MOD, and MARD subphenotypes may require additional



319 biomarkers, such as liver and kidney function markers, to better capture the insulin-resistant  
320 pathophysiology.

321         Given the novel focus on integrating precision medicine into public health,<sup>25</sup> we  
322 underscore two takeaways. First, prognostic models built using cohort studies and variables  
323 routinely collected in large electronic health record databases could enhance geographic  
324 surveillance of high-risk subphenotypes of newly diagnosed T2DM and monitor their  
325 prognosis.<sup>26</sup> Second, the overlap of high-risk subphenotypes and social vulnerability emphasizes  
326 the importance of ensuring equitable access to high-quality diabetes care, including early  
327 diagnosis and access to highly efficacious medications like incretin mimetics, particularly in  
328 underserved communities, given their disproportionate burden of SIDD. This study, therefore,  
329 addresses critical gaps in translating T2DM subphenotypes into clinical practice, advancing  
330 opportunities for tailored treatment strategies. Our findings also support the development of  
331 predictive models for early subphenotype detection driving precision prevention. By enabling  
332 clinical validation, this research takes a key step toward integrating precision medicine into  
333 diabetes care.

## 334 **ONLINE METHODS**

### 335 **Data Sources**

#### 336 *Cohort studies*

337 Data for subphenotyping of newly diagnosed T2DM consisted of six US-based  
338 longitudinal studies funded by the National Institutes of Health: Atherosclerosis Risk in  
339 Communities (ARIC), Coronary Artery Risk Development in Young Adults (CARDIA),  
340 Diabetes Prevention Program (DPP), DPP Outcomes Study (DPPOS is the long-term follow-up  
341 of DPP), Jackson Heart Study (JHS), and Multi-Ethnic Study of Atherosclerosis (MESA).<sup>32–36</sup>  
342 Details of each cohort are provided in **Supplementary Note 1**. Our data, henceforth referred to  
343 as ‘pooled cohort studies’ consisted of a mix of observational cohorts (n =4) and follow-up of  
344 randomized trials (n=2).

#### 345 *Epic Cosmos*

346 Data for validating the subphenotypes of newly diagnosed T2DM were from the Epic  
347 Cosmos Research Platform.<sup>37</sup> Epic Cosmos includes HIPAA-compliant de-identified  
348 longitudinal electronic health records on nearly 250 million (at the time of analysis) unique  
349 patients from all 50 states of the United States, the District of Columbia, and Lebanon. Patients  
350 were linked longitudinally across health systems through an internal privacy-preserving process  
351 by Epic Cosmos. Socio-demographic variables were also harmonized across health systems by  
352 Epic Cosmos upon inclusion in the centralized platform and are broadly representative of the US  
353 population.<sup>31</sup> Epic Cosmos additionally limited the geographic resolution to the state level and  
354 date-shifted the encounters at a patient level to prevent re-identification.

355

## 356 **Newly Diagnosed Type 2 Diabetes Mellitus**

### 357 *Pooled Cohort Studies*

358 A newly diagnosed diabetes case definition was harmonized across the different cohorts  
359 based on the combination of questionnaire and biomarker data. Participants were classified as  
360 having newly diagnosed T2DM if they did not have a T2DM diagnosis at enrollment and either  
361 self-reported a new T2DM diagnosis, reported a diagnosis by a physician or health provider, or  
362 used diabetes medications in the last year (**Supplementary Table 3**). In ARIC, MESA, JHS, and  
363 CARDIA, participants who indicated diabetes at baseline but had missing age of diagnosis or  
364 diabetes duration or the earliest age of diagnosis was more than one year earlier than the baseline  
365 visit age were classified as having previously diagnosed diabetes. In the DPP trial and DPPOS,  
366 the development of diabetes was a primary outcome, so all diabetes cases were considered newly  
367 diagnosed. For those who did not self-report a diagnosis of diabetes, we relied on the 2024 ADA  
368 diagnostic criterion for identifying the cases, namely HbA1c  $\geq 6.5\%$  or fasting plasma glucose  
369  $\geq 126$  mg/dL; or 2-hour oral glucose tolerance test  $\geq 200$  mg/dL (**Supplementary Table 1**). High  
370 random glucose was not considered as criteria for detecting new onset T2DM since data on  
371 diabetes symptoms were not available. We excluded all individuals for whom age, body mass  
372 index (BMI) and HbA1c were missing at diagnosis of T2DM (n = 3,390) or for whom  
373 homeostatic indices of beta cell function and insulin resistance were implausible (n = 13,  
374 acceptable input range for HOMA2 %B and HOMA2-IR calculation: insulin [20 to 400 pmol/L],  
375 glucose [3.0 to 25.0 mmol/L]). The analytic sample of newly diagnosed T2DM for the pooled  
376 cohort dataset consisted of 3,377 individuals (**Supplementary Figure 1**).

### 377 *Epic Cosmos*

378 Newly diagnosed type 2 diabetes was identified using the SUPREME-DM computable  
379 phenotype based on inpatient diagnosis codes or any combination of labs, outpatient diagnosis  
380 codes (International Classification of Diseases-10 or ICD-10-CM) and diabetes medications  
381 occurring within two years of each other (**Supplementary Table 13**).<sup>38,39</sup> We considered the date  
382 of detection of the second criterion of SUPREME-DM as the date of diagnosis. To identify new-  
383 onset T2DM, we restricted our analysis to adult patients (18-99 years) who had in-person  
384 encounters (inpatient or outpatient) in each of the two years before detection. The final analytic  
385 sample of newly detected type 2 diabetes in Epic Cosmos consisted of 727,076 patients  
386 (**Supplementary Table 15**).

## 387 **Data Collection**

### 388 *Pooled Cohort Studies*

389 We identified and obtained additional variables relevant for clinical phenotyping at or  
390 within one year after T2DM diagnosis. Nine clinical variables that are routinely measured during  
391 clinical visits were extracted, namely age at diagnosis, body mass index (BMI, kg/m<sup>2</sup>), systolic  
392 blood pressure (SBP), diastolic blood pressure (DBP), Hemoglobin A1c (HbA1c), Low-Density  
393 Lipoprotein (LDL) cholesterol, High-Density Lipoprotein (HDL) cholesterol, triglycerides, and  
394 the triglyceride-to-HDL cholesterol ratio (an indirect marker of insulin resistance).<sup>18</sup>  
395 Measurement protocols and availability of key clinical variables for each cohort are presented in  
396 **Supplementary Table 2** and **Supplementary Table 4**. Additionally, demographic information  
397 such as gender, race and ethnicity, education level, and lifestyle factors, including drinking and  
398 smoking status, were harmonized across cohorts.

### 399 *Epic Cosmos*

400 Anthropometry (body mass index, blood pressure) and laboratory parameters (HbA1c,  
401 LDL cholesterol, HDL cholesterol, triglycerides) at the most recent visit in the year after  
402 detection of the SUPREME-DM computable phenotype were extracted. We additionally  
403 extracted other clinical and socio-demographic characteristics prior to the index visit: year of  
404 birth, biological sex (male, female), race-ethnicity (Hispanic, NH White, NH Black, NH Other),  
405 insurance status (Medicare, Medicaid, Private/unspecified, Self-pay) for the month of diagnosis,  
406 history of comorbidities based on ICD-10-CM codes, and prescriptions for the year prior to  
407 diagnosis. The percentile ranking for socio-economic position for the zip code of residence was  
408 available through linked Social Vulnerability Index 2020 data from the Center for Disease  
409 Control & Prevention. The most recent residential location was categorized as urban or rural  
410 based on US Department of Agriculture's Rural-Urban Commuting Area 2010 primary codes.

#### 411 **Data Cleaning**

412 We harmonized variable names and units for the cardiometabolic biomarkers across the  
413 pooled cohorts: HbA1c was measured in percentages (%); blood pressure in millimeters of  
414 mercury (mmHg); LDL cholesterol, HDL cholesterol, triglycerides, and fasting blood glucose in  
415 milligrams per deciliter (mg/dL); and fasting insulin in micro-international units per milliliter  
416 ( $\mu$ IU/mL). Data from different studies were then integrated using common identifiers and  
417 harmonized definitions of variables. For the pooled cohort studies, we estimated the Homeostasis  
418 Model Assessment indicators, HOMA2 %B and HOMA2-IR, based on fasting insulin and  
419 fasting blood glucose, using the calculator published by the University of Oxford.<sup>40</sup> For EHRs,  
420 vitals and laboratory parameters are harmonized by Epic Cosmos across all participating health  
421 systems. K-nearest neighbor imputation ( $k = 5$ ) was used to impute missing values of  
422 cardiometabolic biomarkers in pooled cohorts and electronic health records.

## 423 **Classification and Statistical Analysis**

### 424 *Subphenotypes of Newly Diagnosed Type 2 Diabetes in Pooled Cohorts*

425 We conducted an initial hierarchical clustering analysis using five variables (age of  
426 diagnosis, BMI, HbA1c, HOMA2 %B, and HOMA2 IR). Next, we conducted a k-means  
427 clustering analysis with varying numbers of clusters (k ranging from 2 to 10) and identified the  
428 optimal number using the Kneedle algorithm.<sup>41</sup> The optimal number of clusters was identified as  
429 four from both unsupervised approaches. We used the clusters from the k-means clustering and  
430 labelled them based on their clinical similarity to the original ANDIS study subphenotypes as  
431 Mild Age-Related Diabetes (MARD), Mild Obesity-Related Diabetes (MOD), Severe Insulin-  
432 Deficient Diabetes (SIRD), and Severe Insulin-Resistant Diabetes (SIDD). The distribution of  
433 the five key variables is shown in **Supplementary Figure 3**.

### 434 *Classification Model for Subphenotypes in Electronic Health Records*

435 We constructed one-vs-all logistic regression models to predict the probability of being  
436 classified into each of the four subphenotype as a function of the nine routine clinical variables  
437 (age of diagnosis, BMI, HbA1c, SBP, DBP, LDL, HDL, triglycerides, triglyceride-to-HDL  
438 ratio). For example, the SIDD model estimated the probability of an individual being classified  
439 as SIDD and non-SIDD as a function of the nine variables. The pooled cohort data were split into  
440 70% training and 30% test datasets. We identified the probability threshold for each model that  
441 maximized the F1 score (a combination of sensitivity and positive predictive value [PPV]). The  
442 models were used to predict subphenotype membership in the held-out test data. We evaluated  
443 the performance of our prediction models, relative to the cluster analysis labels, using the Area  
444 Under the Receiver Operating Characteristic Curves (AUC), sensitivity, specificity, PPV,

445 negative predictive value (NPV) and F1 score. We applied the logistic regression models to the  
446 data from Epic Cosmos to sequentially estimate the probability of membership and classify cases  
447 into SIDD, MOD and MARD subphenotypes, given the low specificity of the SIRD model. The  
448 distribution of the nine clinical variables used in the logistical regression models is shown in  
449 **Supplementary Figure 4.**

#### 450 *Time to pharmacological Treatment and Risk of Microvascular Complications*

451 In Epic Cosmos, we estimated cumulative incidence curves for prescription of glucose-  
452 lowering medication classes (insulin, metformin and incretin mimetics - glucagon-like peptide-1  
453 receptor agonists or glucagon-like peptide-1 receptor agonists/glucose-dependent insulintropic  
454 polypeptide). We also estimated the cumulative incidence of new-onset microvascular  
455 complications based on diagnostic codes (diabetic nephropathy [E11.2], retinopathy [E11.3], and  
456 neuropathy [E11.4]) among those free of each complication at diagnosis. Unadjusted Kaplan-  
457 Meier curves are shown in **Supplementary Figure 7 and 8.** We estimated hazard ratios and  
458 plotted survival curves of each of these outcomes for membership in SIDD, MARD, and  
459 unclassified cases after adjusting for sex, age, or both, relative to MOD (**Figure 4**). Because of  
460 the elevated risks associated with the SIDD subphenotype and its classification model's high  
461 performance, we compared SIDD to all other subphenotypes grouped into non-SIDD, adjusting  
462 for age and sex (**Supplementary Figure 6**).

#### 463 *Sensitivity Analysis*

464 First, to evaluate if one large and non-representative cohort influenced the final  
465 clustering, we assessed the concordance using the Adjusted Rand Index and Cohen's  $\kappa$  between

466 the original classification using the pooled cohort studies and those generated using a leave-one-  
467 cohort-out clustering approach. Second, we clustered the participants based on the nine routinely  
468 collected variables and compared these new clusters as an alternative to the original clusters.  
469 Third, we trained multiclass prediction models to complement the clustering results as well as to  
470 assess whether non-linearity and statistical interactions between covariates can affect  
471 discrimination. We trained multinomial regression and random forest models using the training  
472 dataset. There were limited improvements in model performance across all subphenotypes  
473 (**Supplementary Table 13**). We observed that misdiagnosed SIRD cases were predominantly  
474 classified as MOD or MARD. A framework of the analysis plan is provided in **Supplementary**  
475 **Figure 1**. All analysis was conducted using Python 3.12.1 and R 4.2.3.



**Ethics approval and consent to participate:** We were exempt from ethical approval for analysis of secondary datasets. All participants of cohort studies and trials gave written informed consent before participation. Epic Cosmos data was HIPAA-limited and expertly de-identified.

**Consent for publication:** Not applicable

**Competing interests:** None declared

**Data availability:** The code for the analysis is available on [https://github.com/jvargh7/diabetes\\_endotypes\\_cohorts](https://github.com/jvargh7/diabetes_endotypes_cohorts). Data for the National Institutes of Health funded cohort studies are available for request from the NIDDK Central Repository (NIDDK-CR; <https://repository.niddk.nih.gov>) and NHLBI Biolincc (<https://biolincc.nhlbi.nih.gov/home/>) to registered users. Epic Cosmos access is available through institutional representatives of participating institutions and the Epic Cosmos team after completing certification requirements.

**Funding:** None

**Author contributions:** JSV and ZL conceptualized the study. JSV, ZL and SL developed the analytic plan with inputs from JCH. ZL and JSV led the data extraction, analysis and wrote the first draft. All authors reviewed and edited the subsequent drafts.

**Acknowledgments:** We thank Kayla Yates, Meghan Howat, and Danessa Sandman of the Epic Cosmos team for their support.

## Reference

1. Centers for Disease Control & Prevention. National Diabetes Statistics Report. Published November 29, 2023. Accessed February 13, 2024. <https://www.cdc.gov/diabetes/data/statistics-report/index.html>
2. International Diabetes Federation. *IDF Diabetes Atlas*. International Diabetes Federation Accessed October 6, 2022. <https://diabetesatlas.org/>
3. Herder C, Herder C, Christian Herder, Roden M, Roden M, Michael Roden. A novel diabetes typology: towards precision diabetology from pathogenesis to treatment. *Diabetologia*. Published online January 4, 2022. doi:10.1007/s00125-021-05625-x
4. Stefan N, Schulze MB. Metabolic health and cardiometabolic risk clusters: implications for prediction, prevention, and treatment. *The Lancet Diabetes & Endocrinology*. Published online May 2023:S2213858723000864. doi:10.1016/S2213-8587(23)00086-4
5. Xing L, Peng F, Liang Q, Dai X, Ren J, Wu H, et al. Clinical Characteristics and Risk of Diabetic Complications in Data-Driven Clusters Among Type 2 Diabetes. *Front Endocrinol*. 2021;12:617628. doi:10.3389/fendo.2021.617628
6. Ahlqvist E, Storm P, Käräjämäki A, Martinell M, Dorkhan M, Carlsson A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *The Lancet Diabetes & Endocrinology*. 2018;6(5):361-369. doi:10.1016/S2213-8587(18)30051-2
7. Varghese JS, Narayan KMV. Ethnic differences between Asians and non-Asians in clustering-based phenotype classification of adult-onset diabetes mellitus: A systematic narrative review. *Primary Care Diabetes*. Published online 2022:4. doi:10.1016/j.pcd.2022.09.007
8. Herder C, Roden M. A novel diabetes typology: towards precision diabetology from pathogenesis to treatment. *Diabetologia*. Published online January 4, 2022. doi:10.1007/s00125-021-05625-x
9. Chung WK, Erion K, Florez JC, Hattersley AT, Hivert MF, Lee CG, et al. Precision Medicine in Diabetes: A Consensus Report From the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). *Diabetes Care*. 2020;43(7):1617-1635. doi:10.2337/dci20-0022
10. Li X, van Giessen A, Altunkaya J, Slieker RC, Beulens JWJ, 't Hart LM, et al. Potential Value of Identifying Type 2 Diabetes Subgroups for Guiding Intensive Treatment: A Comparison of Novel Data-Driven Clustering With Risk-Driven Subgroups. *Diabetes Care*. Published online May 5, 2023:dc222170. doi:10.2337/dc22-2170
11. Ho JC, Staimez LR, Narayan KMV, Ohno-Machado L, Simpson RL, Hertzberg VS. Evaluation of available risk scores to predict multiple cardiovascular complications for patients with type 2 diabetes mellitus using electronic health records. *Computer Methods and Programs in Biomedicine Update*. 2023;3:100087. doi:10.1016/j.cmpbup.2022.100087

12. Varghese JS, Carrillo-Larco RM, Narayan KV. Achieving replicable subphenotypes of adult-onset diabetes. *The Lancet Diabetes & Endocrinology*. Published online July 2023:S221385872300195X. doi:10.1016/S2213-8587(23)00195-X
13. Anjana RM, Baskar V, Nair ATN, Jebarani S, Siddiqui MK, Pradeepa R, et al. Novel subgroups of type 2 diabetes and their association with microvascular outcomes in an Asian Indian population: a data-driven cluster analysis: the INSPIRED study. *BMJ Open Diab Res Care*. 2020;8(1):e001506. doi:10.1136/bmjdr-2020-001506
14. Ke C, Narayan KMV, Chan JCN, Jha P, Shah BR. Pathophysiology, phenotypes and management of type 2 diabetes mellitus in Indian and Chinese populations. *Nat Rev Endocrinol*. Published online May 4, 2022. doi:10.1038/s41574-022-00669-4
15. Zaharia OP, Strassburger K, Strom A, Bönhof GJ, Karusheva Y, Antoniou S, et al. Risk of diabetes-associated diseases in subgroups of patients with recent-onset diabetes: a 5-year follow-up study. *The Lancet Diabetes & Endocrinology*. 2019;7(9):684-694. doi:10.1016/S2213-8587(19)30187-1
16. Christensen DH, Nicolaisen SK, Ahlqvist E, Stidsen JV, Nielsen JS, Hojlund K, et al. Type 2 diabetes classification: a data-driven cluster study of the Danish Centre for Strategic Research in Type 2 Diabetes (DD2) cohort. *BMJ Open Diab Res Care*. 2022;10(2):e002731. doi:10.1136/bmjdr-2021-002731
17. Antonio-Villa NE, Fernández-Chirino L, Vargas-Vázquez A, Fermín-Martínez CA, Aguilar-Salinas CA, Bello-Chavolla OY. Prevalence Trends of Diabetes Subgroups in the United States: A Data-driven Analysis Spanning Three Decades From NHANES (1988-2018). *The Journal of Clinical Endocrinology & Metabolism*. 2022;107(3):735-742. doi:10.1210/clinem/dgab762
18. Giannini C, Santoro N, Caprio S, Kim G, Lartaud D, Shaw M, et al. The Triglyceride-to-HDL Cholesterol Ratio. *Diabetes Care*. 2011;34(8):1869-1874. doi:10.2337/dc10-2234
19. Lu B, Li P, Crouse AB, Grimes T, Might M, Ovalle F, et al. Data-driven Cluster Analysis Reveals Increased Risk for Severe Insulin-Deficient Diabetes in Black/African Americans. *The Journal of Clinical Endocrinology & Metabolism*. Published online July 30, 2024:dgae516. doi:10.1210/clinem/dgae516
20. American Diabetes Association Professional Practice Committee, ElSayed NA, Aleppo G, Bannuru RR, Bruemmer D, Collins BS, et al. 9. Pharmacologic Approaches to Glycemic Treatment: *Standards of Care in Diabetes—2024*. *Diabetes Care*. 2024;47(Supplement\_1):S158-S178. doi:10.2337/dc24-S009
21. Dwivedi C, Ekström O, Brandt J, Adiels M, Franzén S, Abrahamsson B, et al. Randomized open-label trial of semaglutide and dapagliflozin in patients with type 2 diabetes of different pathophysiology. *Nat Metab*. 2024;6(1):50-60. doi:10.1038/s42255-023-00943-3
22. Christiaens A, Baretella O, Del Giovane C, Rodondi N, Knol W, Peters M, et al. Association between diabetes overtreatment in older multimorbid patients and clinical

- outcomes: an ancillary European multicentre study. *Age and Ageing*. 2023;52(1):afac320. doi:10.1093/ageing/afac320
23. Schwartz AV, Vittinghoff E, Sellmeyer DE, Feingold KR, Rekeneire N de, Strotmeyer ES, et al. Diabetes-Related Complications, Glycemic Control, and Falls in Older Adults. *Diabetes Care*. 2008;31(3):391-396. doi:10.2337/dc07-1152
  24. 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. *The Lancet Diabetes & Endocrinology*. 2019;7(6):442-451. doi:10.1016/S2213-8587(19)30087-7
  25. Khoury MJ, Iademarco MF, Riley WT. Precision Public Health for the Era of Precision Medicine. *American Journal of Preventive Medicine*. 2016;50(3):398-401. doi:10.1016/j.amepre.2015.08.031
  26. Roberts MC, Holt KE, Del Fiol G, Baccarelli AA, Allen CG. Precision public health in the era of genomics and big data. *Nat Med*. 2024;30(7):1865-1873. doi:10.1038/s41591-024-03098-0
  27. Florez JC, Pearson ER. A roadmap to achieve pharmacological precision medicine in diabetes. *Diabetologia*. Published online June 24, 2022. doi:10.1007/s00125-022-05732-3
  28. Vujkovic M, Keaton JM, Lynch JA, Miller DR, Zhou J, Tcheandjieu C, et al. Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nat Genet*. 2020;52(7):680-691. doi:10.1038/s41588-020-0637-y
  29. Pociot F, Lernmark Å. Genetic risk factors for type 1 diabetes. *The Lancet*. 2016;387(10035):2331-2339. doi:10.1016/S0140-6736(16)30582-7
  30. Joseph S. Epic's Market Share: Who Should Control The Levers Of Healthcare Innovation? *Forbes*. Published February 26, 2024. Accessed June 11, 2024. <https://www.forbes.com/sites/sethjoseph/2024/02/26/epics-antitrust-paradox-who-should-control-the-levers-of-healthcare-innovation/>
  31. About | Epic Cosmos. Accessed July 28, 2024. <https://cosmos.epic.com/about/>
  32. Rosamond WD, Folsom AR, Chambless LE, Wang CH. Coronary heart disease trends in four United States communities. The Atherosclerosis Risk in Communities (ARIC) Study 1987–1996. *International Journal of Epidemiology*. 2001;30:S17-S22. doi:10.1093/ije/30.suppl\_1.s17
  33. Liu K, Daviglus ML, Loria CM, Colangelo LA, Spring B, Moller AC, et al. Healthy Lifestyle Through Young Adulthood and the Presence of Low Cardiovascular Disease Risk Profile in Middle Age: The Coronary Artery Risk Development in (Young) Adults (CARDIA) Study. *Circulation*. 2012;125(8):996-1004. doi:10.1161/CIRCULATIONAHA.111.060681

34. White NH, Pan Q, Knowler WC, Schroeder EB, Dabelea D, Chew EY, et al. The Effect of Interventions to Prevent Type 2 Diabetes on the Development of Diabetic Retinopathy: The DPP/DPPOS Experience. *Diabetes Care*. 2022;45(7):1640-1646. doi:10.2337/dc21-2417
35. Echouffo-Tcheugui JB, Musani SK, Bertoni AG, Correa A, Fox ER, Mentz RJ. Patients phenotypes and cardiovascular risk in type 2 diabetes: the Jackson Heart Study. *Cardiovasc Diabetol*. 2022;21(1):89. doi:10.1186/s12933-022-01501-z
36. Bild DE. Multi-Ethnic Study of Atherosclerosis: Objectives and Design. *American Journal of Epidemiology*. 2002;156(9):871-881. doi:10.1093/aje/kwf113
37. Tarabichi Y, Frees A, Honeywell S, Huang C, Naidech AM, Moore JH, et al. The Cosmos Collaborative: A Vendor-Facilitated Electronic Health Record Data Aggregation Platform. *ACI open*. 2021;05(01):e36-e46. doi:10.1055/s-0041-1731004
38. Richesson RL, Rusincovitch SA, Wixted D, Batch BC, Feinglos MN, Miranda ML, et al. A comparison of phenotype definitions for diabetes mellitus. *J Am Med Inform Assoc*. 2013;20(e2):e319-e326. doi:10.1136/amiajnl-2013-001952
39. Nichols GA, Schroeder EB, Karter AJ, Gregg EW, Desai J, Lawrence JM, et al. Trends in Diabetes Incidence Among 7 Million Insured Adults, 2006–2011. *American Journal of Epidemiology*. 2015;181(1):32-39. doi:10.1093/aje/kwu255
40. Levy JC, Matthews DR, Hermans MP. Correct Homeostasis Model Assessment (HOMA) Evaluation Uses the Computer Program. *Diabetes Care*. 1998;21(12):2191-2192. doi:10.2337/diacare.21.12.2191
41. Satopaa V, Albrecht J, Irwin D, Raghavan B. Finding a “Kneedle” in a Haystack: Detecting Knee Points in System Behavior. In: *2011 31st International Conference on Distributed Computing Systems Workshops*. IEEE; 2011:166-171. doi:10.1109/ICDCSW.2011.20

**Table 1. Descriptive characteristics of newly diagnosed T2DM subphenotypes in pooled cohort studies**

	<b>Overall</b>	<b>SIDD</b>	<b>SIRD</b>	<b>MOD</b>	<b>MARD</b>
<i>N</i>	3,377	108	488	1,269	1,512
Age at diagnosis (SD)	63.4 (12.4)	61.1 (13.7)	65.7 (9.4)	52 (7.9)	72.5 (7.6)
Female %	2,018 (59.8%)	49(45.4%)	268(54.9%)	896(70.6%)	805(53.2%)
<b>Race<sup>a</sup></b>					
<i>White</i>	1685 (49.9%)	30 (27.8%)	159 (32.6%)	532 (41.9%)	964 (63.8%)
<i>Black</i>	1199 (35.5%)	66 (61.1%)	145 (29.7%)	547 (43.1%)	441 (29.2%)
<i>Other</i>	493 (14.6%)	12 (11.1%)	184 (37.7%)	190 (15%)	107 (7.1%)
<b>Cohort</b>					
ARIC	1062 (31.4%)	44 (40.7%)	10 (2%)	38 (3%)	970 (64.2%)
CARDIA	228 (6.8%)	16 (14.8%)	1 (0.2%)	191 (15.1%)	20 (1.3%)
DPP	285 (8.4%)	5 (4.6%)	5 (1%)	225 (17.7%)	50 (3.3%)
DPPOS	1013 (30%)	4 (3.7%)	62 (12.7%)	629 (49.6%)	318 (21%)
JHS	245 (7.3%)	19 (17.6%)	2 (0.4%)	156 (12.3%)	68 (4.5%)
MESA	544 (16.1%)	20 (18.5%)	408 (83.6%)	30 (2.4%)	86 (5.7%)
<b>Key Biomarkers</b>					
HbA1c (%)	6.3 (5.8, 6.7)	9.4 (8.7, 10.4)	6.5 (6, 6.7)	6.3 (5.9, 6.7)	6.1 (5.7, 6.5)
Body mass index (kg/m <sup>2</sup> )	33.2 (7)	33.7 (7.3)	32.4 (6)	37.9 (7)	29.4 (4.6)
HOMA2-B (%)	108 (73.3, 156.4)	41.9 (21.9, 72.8)	300 (249.8, 367.4)	129.4 (99.3, 166.9)	81.8 (59.2, 111.8)
HOMA2-IR	2.8 (1.7, 4.7)	2.9 (1.6, 5.6)	9 (7.4, 11.8)	3.7 (2.5, 5.2)	1.9 (1.3, 2.9)
Systolic BP (mmHg)	125.6 (18.6)	126.4 (20.6)	126.6 (18.9)	122.5 (18.5)	127.8 (18.2)
Diastolic BP (mmHg)	74.2 (14.6)	77.1 (16.1)	71.1 (10.9)	80.9 (16)	69.3 (11.9)
LDL cholesterol (mg/dL)	108 (34.5)	110.3 (36.6)	105.5 (34.4)	115.9 (33.6)	102.1 (33.8)
HDL cholesterol (mg/dL)	47.6 (12.8)	43.9 (11.4)	46.5 (13.1)	45.4 (11.2)	50 (13.5)
Triglycerides (mg/dL)	148.3 (104.3)	198.8 (158.7)	163 (120)	154.2 (120.5)	135.1 (72.6)
TGL:HDL ratio	2.8 (1.8, 4.2)	3.5 (2.4, 5.6)	3 (2, 4.6)	3 (2, 4.6)	2.5 (1.7, 3.7)

Values are mean (standard deviation) or median (25th percentile, 75th percentile) for continuous variables, and counts (percentages) for categorical variables. <sup>a</sup> Ethnicity information was not available for some cohorts and hence not reported.

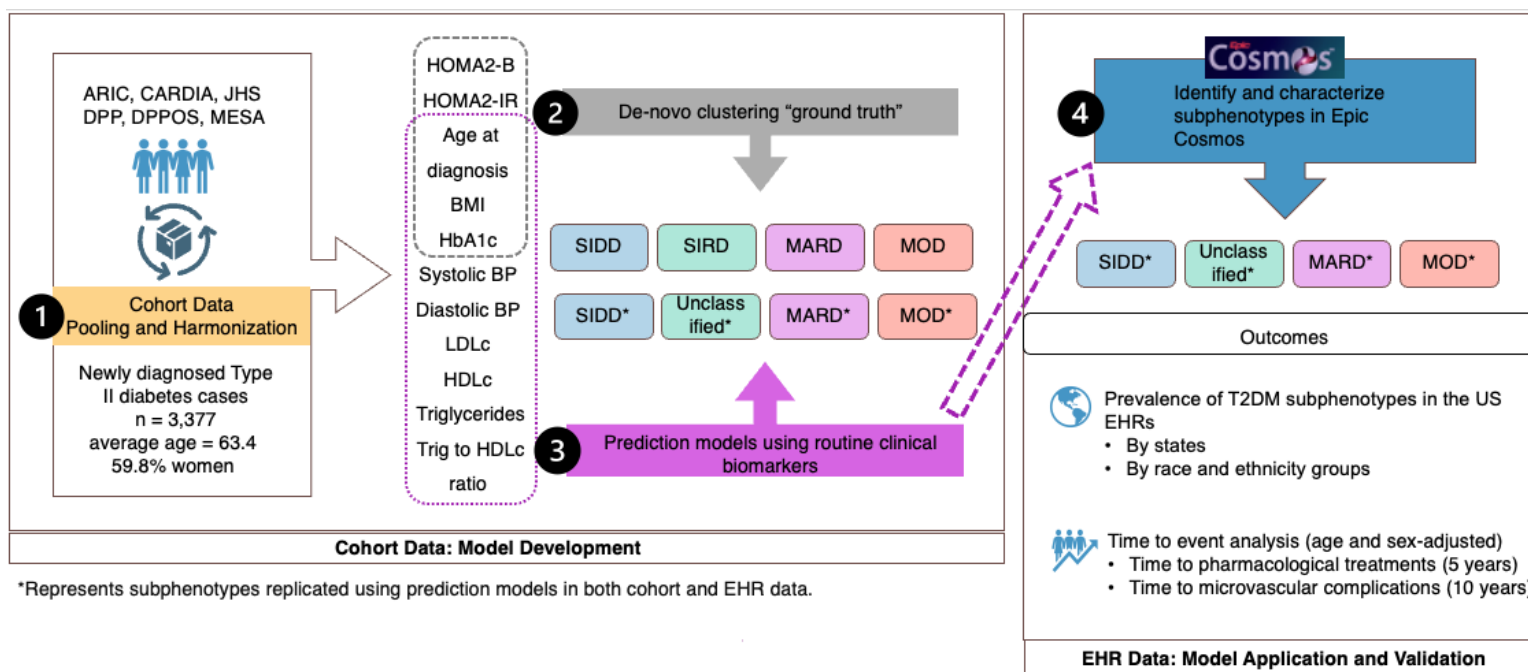
**Table 2. Descriptive characteristics of newly detected type 2 diabetes by subphenotype in Epic Cosmos**

	<b>Overall</b>	<b>SIDD</b>	<b>MARD</b>	<b>MOD</b>	<b>Unclassified</b>
	<i>727,076</i>	<i>156,951</i>	<i>297,568</i>	<i>172,922</i>	<i>99,635</i>
Age at detection of SUPREME-DM	64.4 (13.3)	59.9 (13.5)	74.8(7.9)	51.9 (9.4)	62.9 (5.7)
Female	52%	50%	50%	58%	49%
<b>Race &amp; Ethnicity</b>					
<i>NH White</i>	68%	63%	73%	61%	69%
<i>NH Black</i>	17%	20%	13%	21%	15%
<i>Hispanic</i>	7%	9.3%	4.8%	8.9%	6.8%
<i>NH Other</i>	8.6%	8.0%	8.8%	8.5%	8.8%
Social Vulnerability Index (0: Low, 100: High vulnerability)	60.4 (33.5, 82.1)	64.1 (37.0, 84.6)	57.4(31.2, 79.7)	62.7(35.5, 83.8)	59.6 (32.8, 81.4)
<b>Insurance</b>					
<i>Medicare</i>	40%	31%	62%	15%	32%
<i>Medicaid</i>	14%	16%	18%	18%	13%
<b>Key Biomarkers</b>					
HbA1c (%)	7.0 (6.6, 7.9)	9.5 (8.7, 11.0)	6.7 (6.5, 7.0)	6.8 (6.5, 7.2)	7.1 (6.8, 7.6)
Body mass index	33.2 (6.8)	34.0 (6.9)	29.5 (5.1)	38.8 (6.1)	33.8 (5.3)
Systolic BP	131.2 (15.5)	132.5 (16.3)	131.1(15.7)	130.9 (14.8)	129.9 (15.0)
Diastolic BP	74.7 (10.2)	76.2 (10.4)	71.4 (9.5)	78.7 (9.7)	74.9 (9.3)
LDL cholesterol	89.0 (37.8)	96.2 (41.7)	81.2 (34.8)	97.5 (37.2)	84.5 (35.5)
HDL cholesterol	44.9 (14.6)	42.9 (15.1)	47.8 (15.1)	43.7 (13.0)	42.3 (13.9)
Triglycerides	165.0 (92.4)	185.9(106.2)	141.6 (73.1)	173.3 (93.5)	183.9 (101.8)
TGL:HDL ratio	3.3 (2.1, 5.3)	3.9 (2.4, 6.4)	2.8 (1.8, 4.2)	3.6 (2.3, 5.5)	3.9 (2.5, 6.4)

Values are mean (standard deviation) or median (25<sup>th</sup> percentile, 75<sup>th</sup> percentile) for continuous variables, and percentages for categorical variables. Number of observations where biomarkers are missing is presented in **Supplementary Table 8**.



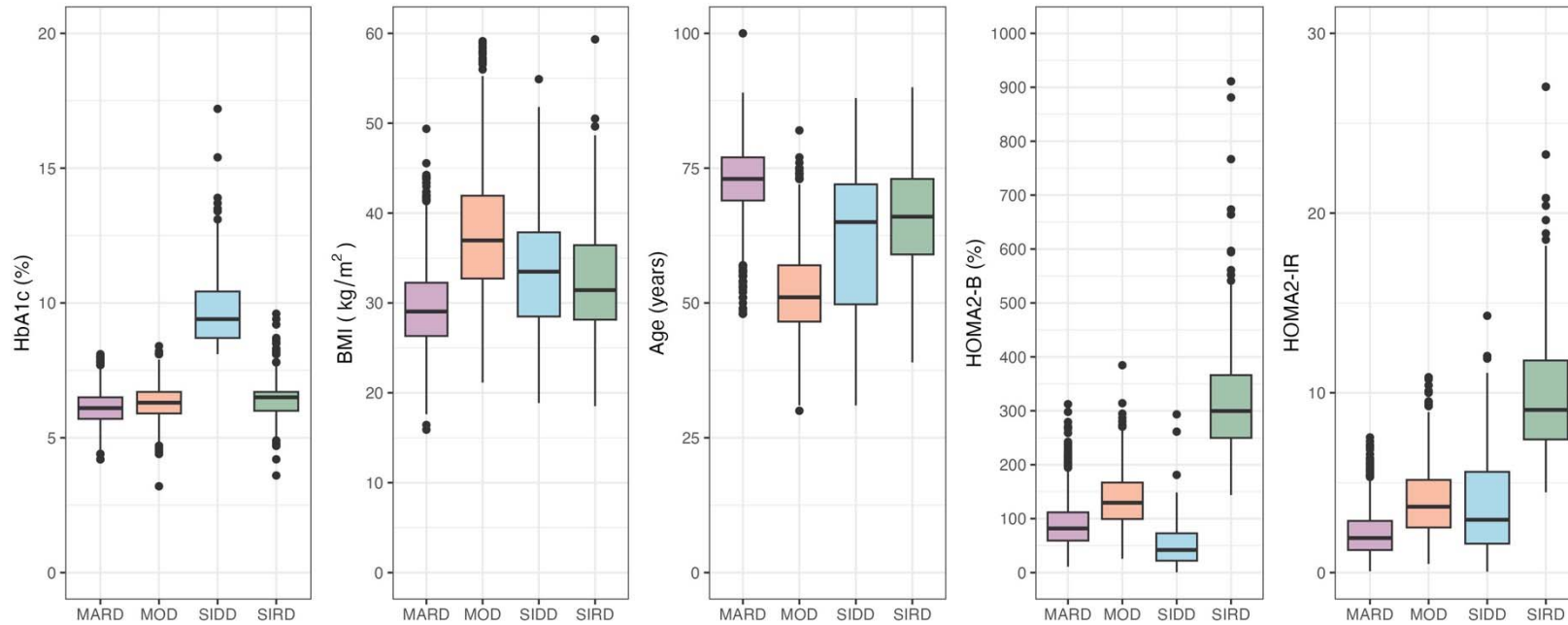
**Figure 1. Study design for translating subphenotypes from cohort studies to electronic health records**



\*Represents subphenotypes replicated using prediction models in both cohort and EHR data.

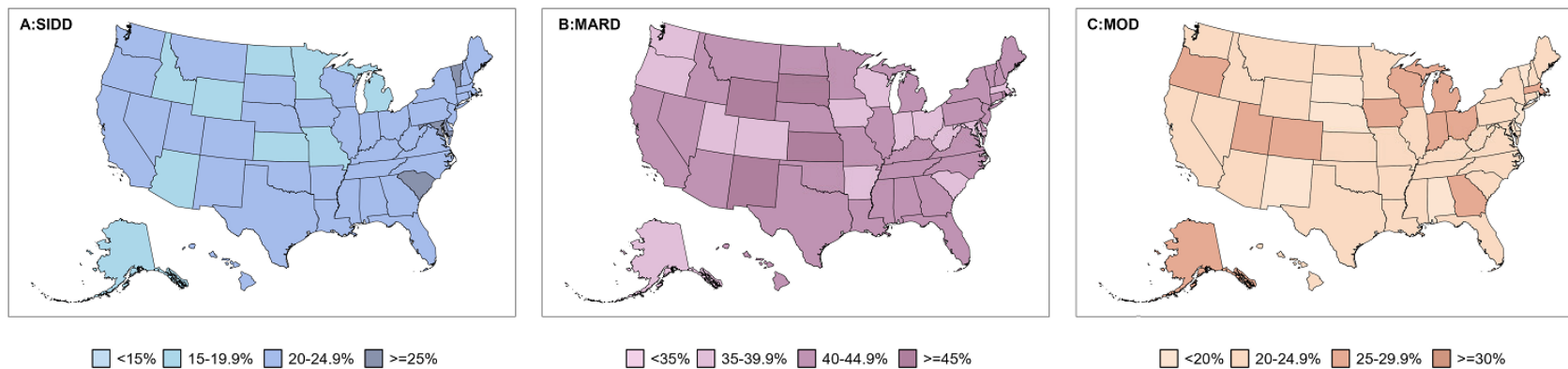


**Figure 2. Distribution of key variables by subphenotypes of newly diagnosed type 2 diabetes in pooled cohort studies**



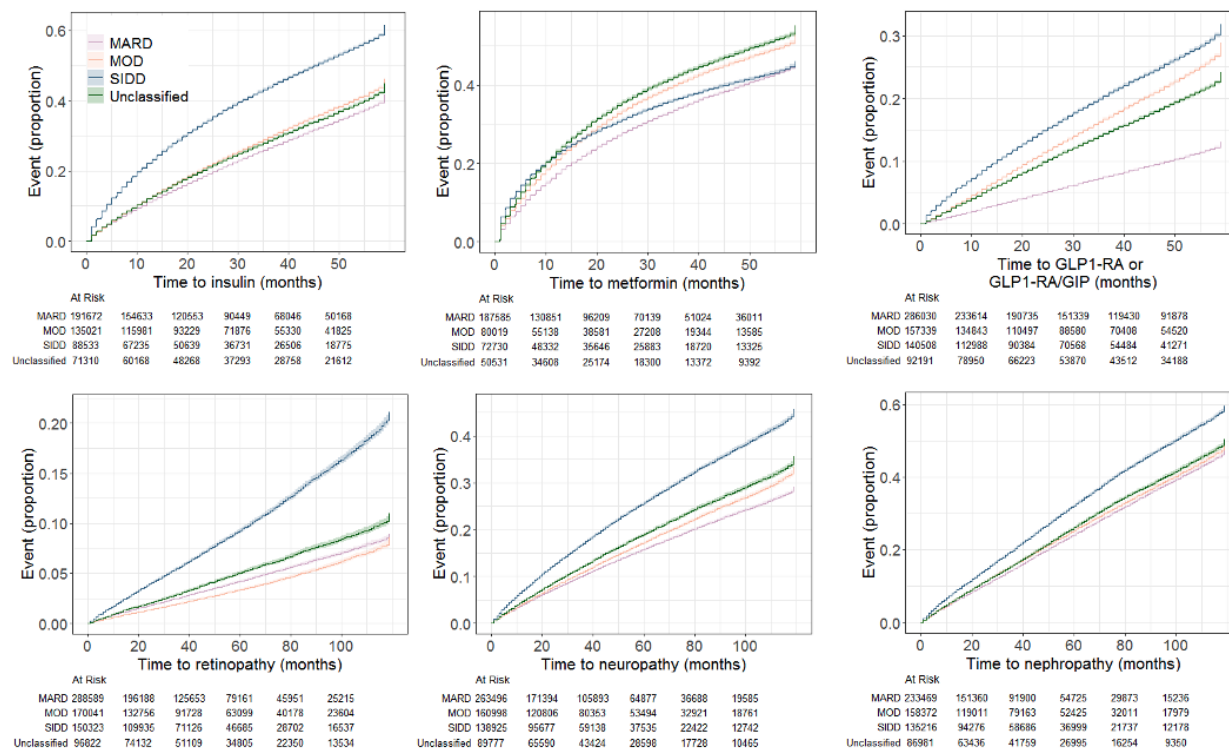
Distribution of HbA1c (%), BMI (kg/m<sup>2</sup>), age (years), HOMA2-B (%), HOMA2-IR for de-novo clusters at diagnosis. K-means clustering was not done separately for males and females based on findings from ANDIS study that suggested similarities in cluster centroids by sex.

**Figure 3. Geographic distribution of type 2 diabetes subphenotypes in Epic Cosmos**



Estimates are based on a sample of 727,076 observations. Panel A: SIDD, Panel B: MARD, Panel C: MOD;

**Figure 4. Time to pharmacological prescriptions and microvascular complications after diagnosis of type 2 diabetes in Epic Cosmos**



All estimates are sex and age-adjusted cumulative incidence curves for (A) time to insulin, (B) time to metformin, (C) time to GLP1-RA or GLP1-RA/GIP (incretin mimetics), (D) time to retinopathy, (E) time to neuropathy and (F) time to nephropathy—more detailed information provided in **Supplementary Table 14**.