Cluster Analysis of Cardiovascular Phenotypes in Patients With Type 2 Diabetes and Established Atherosclerotic Cardiovascular Disease: A Potential Approach to Precision Medicine

Abhinav Sharma,[1,2] Yinggan Zheng,[3] Justin A. Ezekowitz,[2,3] Cynthia M. Westerhout,[3] Jacob A. Udell,[4] Shaun G. Goodman,[3,5] Paul W. Armstrong,[3] John B. Buse,[6] Jennifer B. Green,[7] Robert G. Josse,[5] Keith D. Kaufman,[8] Darren K. McGuire,[9] Giuseppe Ambrosio,[10] Lee-Ming Chuang,[11] Renato D. Lopes,[7] Eric D. Peterson,[7] and Rury R. Holman[12]

Diabetes Care 2022;45:204–212 | https://doi.org/10.2337/dc20-2806

OBJECTIVE

Phenotypic heterogeneity among patients with type 2 diabetes mellitus (T2DM) and atherosclerotic cardiovascular disease (ASCVD) is ill defined. We used cluster analysis machine-learning algorithms to identify phenotypes among trial participants with T2DM and ASCVD.

RESEARCH DESIGN AND METHODS

We used data from the Trial Evaluating Cardiovascular Outcomes with Sitagliptin (TECOS) study (n = 14,671), a cardiovascular outcome safety trial comparing sitagliptin with placebo in patients with T2DM and ASCVD (median follow-up 3.0 years). Cluster analysis using 40 baseline variables was conducted, with associations between clusters and the primary composite outcome (cardiovascular death, nonfatal myocardial infarction, nonfatal stroke, or hospitalization for unstable angina) assessed by Cox proportional hazards models. We replicated the results using the Exenatide Study of Cardiovascular Event Lowering (EXSCEL) trial.

RESULTS

Four distinct phenotypes were identified: cluster I included Caucasian men with a high prevalence of coronary artery disease; cluster II included Asian patients with a low BMI; cluster III included women with noncoronary ASCVD disease; and cluster IV included patients with heart failure and kidney dysfunction. The primary outcome occurred, respectively, in 11.6%, 8.6%, 10.3%, and 16.8% of patients in clusters I to IV. The crude difference in cardiovascular risk for the highest versus lowest risk cluster (cluster IV vs. II) was statistically significant (hazard ratio 2.74 [95% CI 2.29–3.29]). Similar phenotypes and outcomes were identified in EXSCEL.

CONCLUSIONS

In patients with T2DM and ASCVD, cluster analysis identified four clinically distinct groups. Further cardiovascular phenotyping is warranted to inform patient care and optimize clinical trial designs.

Despite growing understanding of the pathophysiology of type 2 diabetes mellitus (T2DM), the associated morbidity and mortality remains high. Similar to other

[1]Division of Cardiology, McGill University, Montreal, Quebec, Canada
[2]Mazankowski Alberta Heart Institute, University of Alberta, Edmonton, Alberta, Canada
[3]Canadian VIGOUR Centre, University of Alberta, Edmonton, Alberta, Canada
[4]Peter Munk Cardiac Centre, University Health Network and Women's College Hospital, University of Toronto, Toronto, Ontario, Canada
[5]St. Michael's Hospital, University of Toronto, Toronto, Ontario, Canada
[6]School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC
[7]Duke Clinical Research Institute, Duke University School of Medicine, Durham, NC
[8]Merck & Co., Inc., Kenilworth, NJ
[9]Division of Cardiology, University of Texas Southwestern Medical Center, Dallas, TX
[10]School of Medicine, University of Perugia, Perugia, Italy
[11]Department of Internal Medicine, National Taiwan University Hospital, Taipei, Taiwan
[12]Radcliffe Department of Medicine, University of Oxford, Oxford, U.K.

Corresponding author: Abhinav Sharma, abhinav.sharma@mcgill.ca

Received 17 November 2020 and accepted 30 September 2021

Clinical trial reg. no. NCT00790205, clinicaltrials.gov

This article contains supplementary material online at https://doi.org/10.2337/figshare.16722358.

disease entities (1–3), there is a limited framework and taxonomy to identify the potentially significant biological heterogeneity with regard to cardiovascular (CV) disease risk among patients with T2DM. Such limitations have significant implications for the design of clinical trials and for patient care, as therapeutic interventions with promising preclinical and early phase trial results may not show efficacy in phase 3 trials when tested on a disease state comprising several phenotypic variations (4–7). Recent trials including several thousands of patients with diabetes have identified various glucose-lowering therapies that are safe from a CV perspective and some that improve CV outcomes (7–10). However, it is unclear how to best tailor these therapies to individual patients in routine clinical practice. Having a more precise CV risk classification of patients with T2DM may allow for more discretely targeted populations for future clinical trials and for more effective clinical application in usual care. Recent calls for improved phenotyping of disease to aid in diagnosis, prognosis, and selection of treatments have spurred intensive investigation into precision medicine (4,11).

Cluster analysis, a form of machine learning, has been used as an exploratory technique to analyze molecular data in various disease states and can be used to identify clusters of patients with distinct phenotypes without the need for historical or arbitrary a priori assumptions (2,3,12–15). When using clinical data, cluster analysis coalesces groups of patients together (into "clusters") such that patients in one particular cluster are more similar (e.g., in baseline characteristics, biomarkers, molecular, and/or genetics data) than patients in other clusters (2,3). The resulting clusters may then demonstrate different natural histories, outcomes, and potentially therapeutic responses. While prior biomarker-based analyses have identified potential pathobiological differences between patients with and without diabetes (16), identifying patient clusters on the basis of clinical variables represents an important initial step for future biomarker and genomic phenotyping.

The Trial Evaluating Cardiovascular Outcomes with Sitagliptin (TECOS) identified that among patients with T2DM and established atherosclerotic CV disease (ASCVD), adding sitagliptin to usual care did not increase the risk of major CV events (17,18). The primary aim of the present analysis was to identify whether a cluster analysis algorithm could identify clusters with distinct CV phenotypes among patients with T2DM with prevalent ASCVD using data from TECOS. Secondarily, we aimed to evaluate whether these clusters were associated with different clinical outcomes and whether patients in these clusters had a differential response to sitagliptin. We subsequently externally replicated the identification of similar cluster phenotypes and the association of the clusters with clinical outcomes using data from the Exenatide Study of Cardiovascular Event Lowering (EXSCEL) trial (19).

RESEARCH DESIGN AND METHODS

Data Source

TECOS was a double-blind, multinational, placebo-controlled CV safety study evaluating the long-term effect of adding sitagliptin, a dipeptidyl peptidase-4 inhibitor, to usual care in patients with T2DM and established ASCVD. The design and results have been reported (17,18). Briefly, TECOS enrolled 14,671 patients who were randomized to the addition of sitagliptin or placebo to their existing glucose-lowering therapy in the context of usual care. Eligible patients were at least 50 years of age with T2DM and established ASCVD and had glycated hemoglobin (HbA_{1c}) values between 6.5 and 8.0% on treatment with stable doses of one or two oral glucose-lowering agents or stable treatment with insulin with or without metformin. Patients were excluded from enrollment if their estimated glomerular filtration rate (eGFR) was $<$30 mL/min/1.73 m^2 or if they had two or more episodes of severe hypoglycemia in the preceding year. Median follow-up was 3 years.

TECOS was designed and run independently by the Duke Clinical Research Institute and the University of Oxford Diabetes Trials Unit in an academic collaboration with the sponsor and funder, Merck Sharp & Dohme Corp., a subsidiary of Merck & Co., Inc. TECOS was approved by the ethics committee for each participating site and monitored by an independent data and safety monitoring board. All participants provided written informed consent for trial participation. Database management was performed by the Duke Clinical Research Institute.

Statistical analysis for the current study was conducted by the Canadian VIGOUR Centre (at the University of Alberta in Edmonton, Alberta, Canada). The authors take responsibility for the manuscript's integrity and had control and authority over its preparation and the decision to publish. The sponsor was able to review the manuscript and provide nonbinding feedback.

End Points

For this study, we have used the TECOS four-point primary composite outcome of time to first event of CV death, nonfatal myocardial infarction (MI), nonfatal stroke, or hospitalization for unstable angina. We also explored the three-point secondary composite outcome of time to first event of CV death, nonfatal MI, or nonfatal stroke; the components of these composites individually; and hospitalization for heart failure (HF). All of these CV events had previously been confirmed by blinded central adjudication.

Cluster Analysis

The statistical methodology of cluster analysis has been described previously (20,21). Multiple steps were included in the process of identifying patient clusters. The first step was variable clustering. Forty candidate baseline variables were selected for variable clustering (Supplementary Appendix 1). Using a criterion of second eigenvalue <1, variables were aggregated into several clusters; this second eigenvalue reflects a vector variable that provides a statistical threshold to split the population groups that maximizes correlation within groups and minimizes correlations between groups. This step was performed separately on continuous and binary variables, resulting in 4 clusters of continuous variables and 10 clusters of categorical variables. A summary score for each individual patient was calculated based on each of the 14 identified variable clusters. Then, the 14 summary scores were standardized to have a mean of zero and SD of 1. From the standardized summary scores, the

second step, patient clustering, was conducted using the Ward minimum variance method of clustering (Supplementary Appendix 2). As the primary analysis of the TECOS trial comparing sitagliptin to placebo stratified results by region, we used a similar approach in evaluating the association of cluster membership to clinical outcomes (18).

Given that the determination of the numbers of clusters was not prespecified, we used the cubic-clustering criterion (CCC) combined with visual examination of the tree diagram. The CCC can be used to estimate the number of clusters based on minimizing the within-cluster sum of squares. The tree diagram was generated to display the semipartial R^2 obtained at each iteration of the patient clustering process. According to CCC and the tree diagram, we examined both five-cluster and four-cluster models. The four-cluster model retained a semipartial R^2 of 0.05 and formed much clearer patterns of patient clusters than the five-cluster model. Therefore, the four-cluster model is presented in this study, and the five-cluster model is shown in Supplementary Appendix 3.

External Replication

An external replication of the cluster analysis was conducted using data from EXSCEL. Briefly, EXSCEL was a multicenter, double-blind study that randomized 14,752 patients with T2DM who had established ASCVD (~70%) or multiple CV risk factors (~30%) to receive subcutaneous injections of once-weekly extended-release exenatide (EQW) or matching placebo (19). Patients were followed for a median of 3.2 years. Once-weekly extended-release exenatide was noninferior to placebo for the primary three-point composite outcome of CV death, nonfatal MI, or nonfatal stroke.

The same clustering technique used in TECOS was applied to the EXSCEL data. The same 40 candidate baseline variables were selected for the variable reduction process, resulting in 6 clusters for continuous variables and 12 clusters for categorical variables. A semipartial R^2 of 0.05, as used in the main analysis, was used to determine the final number of clusters; four patient clusters were identified.

Characteristics and Outcomes Comparisons

For each cluster identified, baseline characteristics for continuous variables are reported as means (SDs) or medians (25th and 75th percentiles) and for categorical variables as percentages. Kaplan–Meier survival analyses were then used to examine the survival functions in the individual clusters, with differences among clusters tested by the log-rank test.

The relative associations between cluster membership and clinical outcomes were assessed using Cox proportional hazards regression stratified by region defined as Asian Pacific and other, Eastern Europe, Latin America, North America, and Western Europe. The proportional hazards assumption was evaluated graphically using the standardized score process and the supremum test. No violations were found. Hazard ratios (HRs) and 95% CIs are reported using cluster II as the reference cluster (as it had the lowest rate of events for the primary outcome). We calculated the median validated thrombolysis in MI (TIMI) risk score for secondary prevention across clusters (22).

To test whether sitagliptin modified the relationship of cluster membership and clinical outcomes, the interaction between study treatment and cluster membership was examined using Cox proportional hazards regression.

The analysis was done on an imputed data set; imputed values were obtained by the Markov chain Monte Carlo method or regression methods using SAS PROC MI. All analyses were based on the intention-to-treat population, with two-sided $P < 0.05$ considered statistically significant. Statistical analyses were performed using SAS software (version 9.4; SAS Institute, Cary, NC).

RESULTS

From TECOS, four distinct clusters associated with CV outcomes were identified based on the summary scores calculated in variable cluster analysis without knowledge of outcomes (Supplementary Appendix 2).

Baseline Characteristics

Baseline characteristics by cluster membership are shown in Tables 1 and 2. Cluster I (40.9% of TECOS participants) had a mean age of 66.2 years and comprised primarily Caucasian men (90.4% Caucasian, 77% men). This cluster had the highest proportion with atherosclerotic coronary heart disease (96.3%) but the lowest proportion with prior cerebrovascular disease (13.1%) and peripheral arterial disease (6.7%). Cluster I also had the lowest mean total cholesterol (160 mg/dL), lowest mean LDL-cholesterol (86 mg/dL), and the highest statin use (85.9%). In addition, this cluster had the highest proportion of prior smokers (48.4%).

Cluster II (23.7% of TECOS participants) had the lowest mean age (63.2 years) and the highest proportion of Asian patients (85.6%). This cluster had the lowest mean BMI (26.4 kg/m^2) and blood pressure (133/77 mmHg). This cluster had the shortest mean duration of diabetes (11.0 years) and the lowest median urinary albumin-to-creatinine ratio (UACR; 8.0 mg/g). These patients also had the second lowest mean LDL-cholesterol (91 mg/dL) and the lowest mean triglycerides (144 mg/dL).

Cluster III (18.2% of TECOS participants) had the highest proportion of females (43.1%), the longest mean duration of diabetes (12.2 years), and the highest proportion with prior diabetic neuropathy (37.6%). This cluster had the lowest proportion of patients with ASCVD (21.3%) but the highest proportion of patients with prior cerebrovascular disease (55.8%) or prior peripheral arterial disease (40.6%). This cluster also had the highest proportion of patients who were current smokers (13.2%). These patients also had the highest mean LDL-cholesterol (99 mg/dL) and the highest mean HDL-cholesterol (46 mg/dL) but the lowest proportion of statin use at baseline (67.9%).

Cluster IV (17.1% of TECOS participants) had the highest mean age (66.8 years) and was composed primarily of Caucasian males (84.2% White, 64.4% male). This cluster had the highest median UACR (17.7 mg/g) and the lowest eGFR (71.7 mL/min/1.73 m^2). In addition, this cluster had the second highest proportion of patients with prior coronary artery disease (83.5%). Almost every patient in this cluster had a prior history of HF (99%).

Table 1—Baseline characteristics by cluster in TECOS

	Cluster I	Cluster II	Cluster III	Cluster IV
N	6,001 (40.9)	3,490 (23.7)	2,672 (18.2)	2,508 (17.1)
Age (years)[a]	66.2 (7.8)	63.2 (7.8)	65.7 (7.9)	66.8 (8.1)
Women	1,380 (23.0)	870 (24.9)	1,155 (43.2)	892 (35.6)
Race				
White	5,424 (90.4)	219 (6.3)	2,202 (82.4)	2,112 (84.2)
Black	161 (2.7)	80 (2.3)	163 (6.1)	43 (1.7)
Asian	69 (1.1)	2,986 (85.6)	84 (3.1)	126 (5.0)
Other	347 (5.8)	205 (5.9)	223 (8.3)	227 (9.1)
Region				
Asia Pacific and other	1,118 (18.6)	2,928 (83.9)	304 (11.4)	215 (8.6)
Eastern Europe	1,335 (22.2)	103 (3.0)	1,084 (40.6)	1,443 (57.5)
Latin America	528 (8.8)	232 (6.6)	439 (16.4)	272 (10.8)
North America	1,766 (29.4)	142 (4.1)	390 (14.6)	296 (11.8)
Western Europe	1,254 (20.9)	85 (2.4)	455 (17.0)	282 (11.2)
Ethnicity				
Hispanic/Latino	719 (12.0)	263 (7.5)	494 (18.5)	322 (12.8)
Duration of diabetes (years)[b]	11.8 (8.1)	11.0 (7.7)	12.2 (8.5)	11.4 (8.3)
HbA_{1c} (%)	7.2 (0.5)	7.3 (0.5)	7.2 (0.5)	7.2 (0.5)
HbA_{1c} (mmol/mol)	55.5 (6.7)	56.7 (6.8)	55.7 (7.4)	55.5 (6.9)
BMI (kg/m^2)	31.3 (5.5)	26.4 (4.0)	31.1 (5.5)	31.7 (5.6)
Systolic blood pressure (mmHg)	135 (17)	133 (17)	139 (17)	136 (17)
Diastolic blood pressure (mmHg)	76 (11)	77 (10)	78 (10)	79 (10)
eGFR (mL/min/1.73 m^2)[c]	75.3 (20.6)	76.4 (20.6)	74.9 (22.3)	71.7 (21.4)
eGFR $<$50 mL/min/1.73 m^2[c]	497 (8.3)	263 (7.6)	261 (9.9)	350 (14.1)
UACR (mg/g)	10.8 (4.4, 31.8)	8.0 (3.3, 30.1)	12.4 (3.5, 40.0)	17.7 (4.6, 53.9)
Total cholesterol (mg/dL)	160 (43)	162 (42)	178 (49)	173 (49)
LDL-cholesterol (mg/dL)	86 (76)	91 (36)	99 (41)	97 (40)
HDL-cholesterol (mg/dL)	43 (12)	43 (11)	46 (14)	44 (13)
Triglycerides (mg/dL)	178 (122)	144 (71)	163 (81)	166 (83)
Triglycerides (mmol/L)	2.0 (1.4)	1.6 (0.8)	1.8 (0.9)	1.9 (0.9)
Prior atherosclerotic coronary disease	5,778 (96.3)	2,420 (69.3)	570 (21.3)	2,095 (83.5)
MI	3,320 (55.3)	1,079 (30.9)	319 (11.9)	1,537 (61.3)
$\geq$50% coronary stenosis	4,226 (70.4)	1,897 (54.4)	238 (8.9)	1,326 (52.9)
Prior PCI	3,327 (56.4)	1,190 (34.6)	154 (5.8)	1,043 (42.1)
CABG	2,031 (33.8)	786 (22.5)	161 (6.0)	686 (27.4)
Prior cerebrovascular disease	787 (13.1)	682 (19.5)	1,490 (55.8)	629 (25.1)
Prior peripheral arterial disease	403 (6.7)	639 (18.3)	1,085 (40.6)	306 (12.2)
Prior HF	57 (0.9)	12 (0.3)	67 (2.5)	2,507 (100.0)
NYHA class 3 or higher	22 (38.6)	6 (50.0)	29 (43.3)	316 (12.6)
Cigarette smoking				
Current smoker	768 (12.8)	288 (8.3)	352 (13.2)	270 (10.8)
Prior smoker	2905 (48.4)	982 (28.1)	962 (36.0)	995 (39.7)
Never smoked	2,328 (38.8)	2,220 (63.6)	1,358 (50.8)	1,243 (49.6)
Diabetic neuropathy	1,051 (17.5)	547 (15.7)	1,006 (37.6)	750 (29.9)
Retinopathy	619 (10.3)	314 (9.0)	512 (19.2)	419 (16.7)
TIMI risk score for secondary prevention	3 (2, 3)	2 (2, 3)	3 (2, 4)	4 (3, 4)

Data for continuous variables are mean (SD) or median (Q1, Q3), and categorical variables are *n* (%). UACR data available for only 5,148 patients. SI conversion factors: UACR (mg/g to g/mol), multiply by 0.1131; total cholesterol, LDL-cholesterol, and HDL-cholesterol (mg/dL to mmol/L), multiply by 0.0259; and triglycerides (mg/dL to mmol/L), multiply by 0.0113. CABG, coronary artery bypass graft; NYHA, Hew York Heart Association; PCI, percutaneous coronary intervention. [a]Age missing among patients in Lithuania as birth date could not be provided. [b]Duration = (year of randomization − year of diagnosis) + 1. [c]MDRD formula used to calculate eGFR. Site-reported values are presented.

Table 2—Baseline glucose-lowering and cardiac-related medication use by cluster in TECOS

	Cluster I	Cluster II	Cluster III	Cluster IV
N	6,001 (40.9)	3,490 (23.7)	2,672 (18.2)	2,508 (17.1)
Metformin	4,964 (82.7)	3,062 (87.7)	2,115 (79.2)	1,825 (72.8)
Sulfonylurea	2,291 (38.2)	2,162 (61.9)	1,114 (41.7)	1,078 (43.0)
Thiazolidinedione	220 (3.7)	86 (2.5)	53 (2.0)	37 (1.5)
Insulin	1,573 (26.2)	359 (10.3)	770 (28.8)	706 (28.1)
More than two agents above	3,494 (58.2)	2,498 (71.6)	1,486 (55.6)	1,342 (53.5)
β-Blocker	4,335 (72.2)	1,765 (50.6)	1,248 (46.7)	1,974 (78.7)
ACE inhibitor or ARB	4,896 (81.6)	2,332 (66.8)	2,187 (81.8)	2,140 (85.3)
Calcium channel blocker	2,013 (33.5)	1,030 (29.5)	1,045 (39.1)	873 (34.8)
Diuretic	2,425 (40.4)	796 (22.8)	1,320 (49.4)	1,479 (59.0)
Thiazide	1,478 (60.9)	523 (65.7)	839 (63.6)	624 (42.2)
Aspirin	5,049 (84.1)	2,803 (80.3)	1,773 (66.4)	1,893 (75.5)
Other antiplatelet	1,263 (21.0)	1,143 (32.8)	332 (12.4)	449 (17.9)
Statin	5,157 (85.9)	2,813 (80.6)	1,813 (67.9)	1,936 (77.2)
Ezetimibe	461 (7.7)	106 (3.0)	119 (4.5)	75 (3.0)
Nitrates	1,343 (22.4)	657 (18.8)	220 (8.2)	593 (23.6)

Data are *n* (%). ARB, angiotensin receptor blocker.

Clinical Outcomes by Cluster

The highest incidence of the four-point primary composite outcome occurred in cluster IV (16.8%) and the lowest in cluster II (8.6%), with an HR of 2.74 (Figs. 1 and 2 and Supplementary Table 1). The incidence of the primary outcome was numerically similar for clusters I and III (11.6% and 10.3%, respectively). Compared with cluster II, the risk of the primary outcome was significantly increased in cluster I (HR 1.57 [95% CI 1.34–1.85]; $P < 0.001$), cluster III (HR 1.55 [95% CI 1.28–1.88]; $P < 0.001$), and cluster IV (HR 2.74 [95% CI 2.29–3.29]; $P < 0.001$) (Fig. 2).

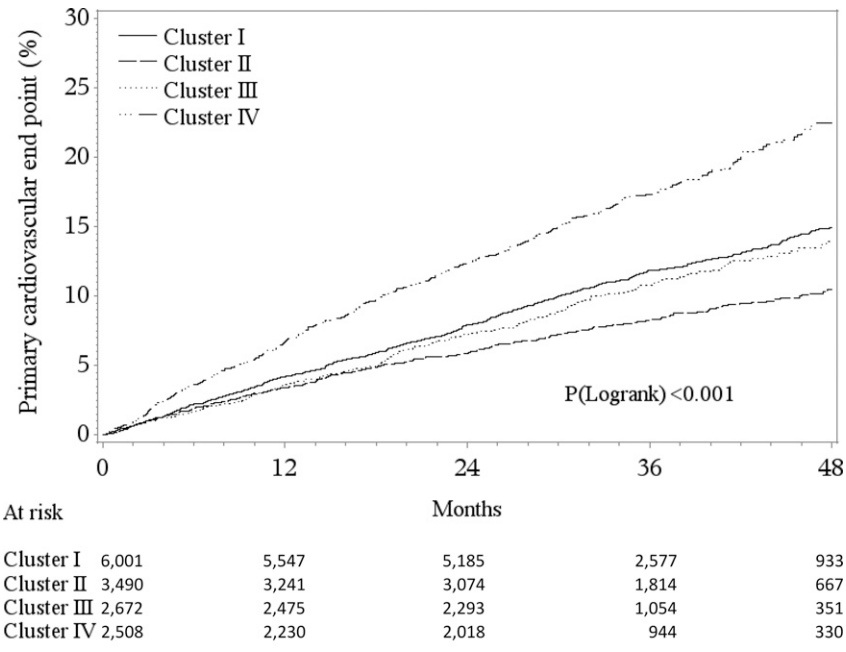

Figure 1—Kaplan-Meier estimated cumulative incidence of CV death, nonfatal MI, nonfatal stroke, or hospitalization for unstable angina end point by cluster.

These relationships were similar for the three-point secondary composite outcome, with cluster IV having the highest event rate (15.1%) and cluster II the lowest event rate (7.6%) (Fig. 2 and Supplementary Table 1). When individual end points were evaluated, cluster IV had the highest risk of CV death and HF hospitalization (compared with cluster II) (Supplementary Fig. 1). For fatal or nonfatal MI, clusters I and IV had a similarly higher risk compared with cluster II. For unstable angina hospitalization, compared with cluster II, only clusters I and IV had a significantly higher risk (Supplementary Fig. 1).

Clinical Outcomes by Sitagliptin Assignment and Cluster Membership

There were no differential treatment effects for the four-point primary composite outcome according to sitagliptin assignment across the clusters despite the different risks for this outcome by cluster membership (interaction $P = 0.9$) or for the three-point secondary composite outcome (interaction $P = 0.7$) (Supplementary Fig. 2). The TIMI risk score for secondary prevention was highest among cluster IV, the cluster with the highest overall event risk (Table 1).

External Replication in the EXSCEL Trial

Using data from EXSCEL, four clusters were also identified with similar baseline demographic profiles as TECOS (Supplementary Tables 2 and 3). Furthermore, the tree diagram identified similar cluster grouping in EXCEL and TECOS (Supplementary Appendix 2). Aligned with the TECOS data, the cluster with the greatest prevalence of patients from Asia (cluster II) had the lowest risk of the three-point primary major adverse CV event outcome (Supplementary Table 4). In comparison, the cluster that had all patients with baseline HF (cluster IV) had the highest risk of three-point major adverse CV event, CV death, and HF hospitalization (Supplementary Table 4).

CONCLUSIONS

We evaluated the ability of a cluster analysis algorithm to identify phenotypically distinct clusters with regard to CV

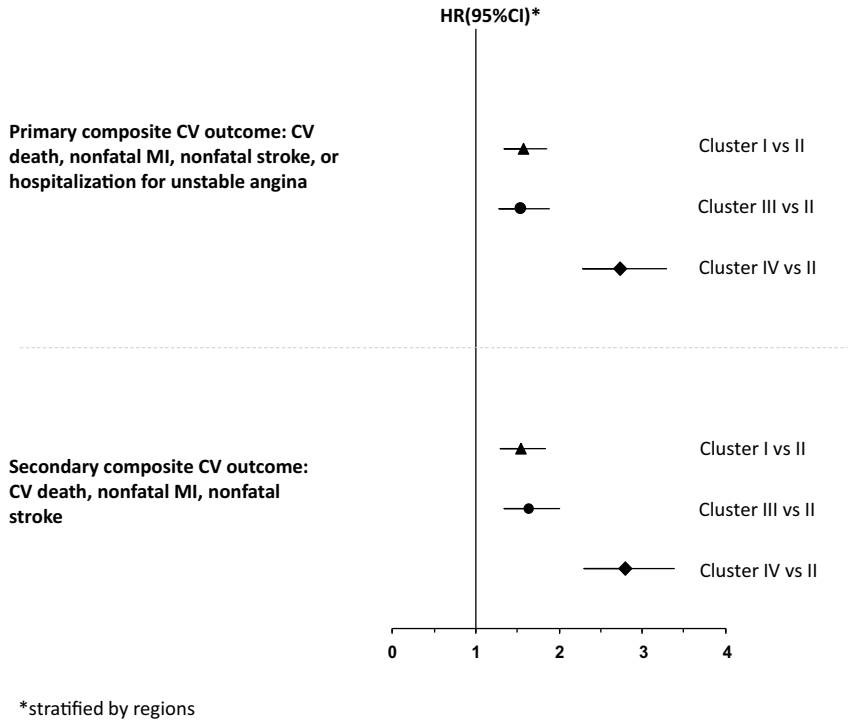

Figure 2—Association between cluster and clinical outcomes (cluster II as reference category).

risk among patients with T2DM at high CV risk. Using data from TECOS, we identified the following major findings: *1*) four distinct clusters had clinically different phenotypes; *2*) these clusters were associated with different risks of CV outcomes; *3*) sitagliptin did not modify the association between clinical outcomes and cluster groups, even among the highest risk groups, nor was there heterogeneity of the effect of sitagliptin on CV outcomes across the clusters; and *4*) the profiles of the clusters and the association of the clusters with clinical outcomes were externally replicated using data from EXSCEL. Clustering appears to occur primarily around concomitant comorbidities such as macrovascular disease, microvascular disease, HF, and kidney dysfunction. Our results demonstrate that an agnostic, data-driven approach is able to identify clinically distinct phenotypes of patients with different trajectories of risk. Furthermore, these findings highlight the clinical heterogeneity in CV risk that exists among patients with T2DM who have prevalent ASCVD or multiple CV risk factors.

The identification of clustering of clinically distinct groups associated with variable CV risk highlights the significant phenotypic heterogeneity among patients with T2DM and ASCVD. Our data identified several distinct clusters of patients based on macrovascular disease, microvascular disease, and metabolic profiles. Cluster I reflects a group of patients primarily with coronary artery disease, while cluster III suggests a noncoronary ASCVD cohort of patients. Cluster II represents a group of Asian patients with low BMI, while cluster IV suggests a group with primarily HF and kidney dysfunction. Differences were also seen with the distribution of microvascular disease. Cluster III had the highest prevalence of diabetic neuropathy, while cluster II had the lowest; cluster IV had the highest UACR, while cluster II had the lowest. In addition to the distribution of macrovascular and microvascular disease, distinctive patterns of LDL-cholesterol, HDL-cholesterol, and triglycerides were identified across the clusters. These results build on previous reports of cluster analyses that have identified patterns of comorbidities and lipids among patients with diabetes associated with CV risk (14,21,23–25). Subclassification of T2DM based on distribution of vascular disease, comorbidities, and metabolic profiles may lead to an improved understanding of the underlying biology of patients with T2DM and ASCVD.

Traditional modeling has demonstrated mixed results in identifying patient cohorts at different risks of CV outcomes across a spectrum of disease states, including diabetes (12,26–29). Machine-learning algorithms represent a tool for analysis of large data sets to aid in the identification of comparatively high- and low-risk patients (11). Extending results from other populations without T2DM (13,14,30), the present analyses demonstrate that among patients with T2DM and ASCVD, cluster analysis identified patients at high and low risk of CV outcomes. These results likely reflect the prevalence of HF in cluster IV, which is associated with significantly increased risk of CV outcomes among patients with diabetes (31); furthermore, a lower risk of CV outcomes has been seen in Asian patients (32). Although use in populations outside of the clinical trial setting remains to be examined, these results suggest that cluster analysis may play a complementary role to traditional modeling in identifying patients at different risk of CV events. Yet, compared with traditional risk modeling, such as the UK Prospective Diabetes Study (UKPDS) Risk Engine and Outcomes Model, challenges are present using cluster analysis for robust prognostication for CV events (33,34). The principal challenge is the need to access a large number of baseline variables for a given patient to identify cluster membership, which would enable the subsequent identification of future CV risk. Future work will be needed to determine whether such agnostic data-driven strategies can improve prognostication above traditional modeling.

Despite similarities across the clusters in TECOS and EXSCEL, there are some differences that warrant further discussion. Baseline sulfonylurea use was highest in cluster II across TECOS and EXSCEL, but the prevalence was higher in TECOS compared with EXSCEL (61.9% and 41.8%, respectively). Similarly, baseline insulin use was highest in cluster III across TECOS and EXSCEL, but the prevalence was higher in TECOS than EXSCEL (28.8% and 51.3%, respectively). The differences in prevalence seen across clusters likely reflect variations in baseline characteristics and practice patterns of sites enrolling across trials. In TECOS, baseline use of sulfonylurea and insulin was 45.3% and 23.2%, respectively; in

EXSCEL, baseline use of sulfonylurea and insulin was 36.6% and 46.3%, respectively. In addition, there are demographic differences between the trials. In TECOS, the total prevalence of Hispanic/Latino participants was 12.3%, with the maximum number seen in cluster III (18.5%). In EXSCEL, the total prevalence of Hispanic/Latino participants was 20.5%, with the maximum number seen in cluster II (59.9%). Such differences may have arisen due to the difference in enrollment across regions. Hence, our results also provide further insights into some of the challenges of cluster analysis, namely that baseline characteristics, differences in practice patterns, and regional differences in enrollment may influence cluster characteristics.

In the TECOS study, sitagliptin had no impact on the risk of adverse CV events (19,35). Traditional subgroup analyses of interventions focus on the presence or absence of an individual risk factor and rarely capture the complexity and heterogeneous nature of the patient population in clinical practice. Previous cluster analyses in populations without diabetes have identified differential treatment effects of some interventions (13,14,30,36). Cluster analysis may represent a methodology to extend upon and complement traditional subgroup analyses by grouping patients into phenotypically and clinically distinct groups and assessing the risk of CV events across these groups.

Implications of Cluster Analysis Results on Patient Care and Clinical Trials Planning

These results may have direct implications for patient care. Patients who resemble cluster I may benefit from intensification of therapies that may reduce vascular risk given the high rates of fatal and nonfatal MIs (37). Patients in cluster IV may benefit from intensification of therapies that reduce HF outcomes. For instance, sodium–glucose cotransporter 2 inhibitors that have demonstrated efficacy in reducing the risk of HF hospitalization in patients with T2DM and CV disease may be considered given the beneficial HF outcomes (8). Other glucose-lowering therapies that may have attenuated benefit in patients with T2DM and HF can be avoided (38). Tailoring therapies to risk profiles represents a fundamental focus of precision medicine, and data-driven approaches to identifying patient phenotypes may represent one approach in identifying these profiles. While the clusters demonstrated varying degrees of risk as identified by the TIMI risk score for secondary prevention, the use of cluster analysis as a tool for risk stratification would require further validation against other risk stratification tools and in other patient populations.

The present results may also have important implications for the design of clinical trials among patients with T2DM. The wide ASCVD risk heterogeneity among patients observed in recent clinical trials of patients with T2DM may contribute to the apparent lack of efficacy of seemingly promising therapies (38). Therapeutic interventions that may be effective in more precisely targeted patient populations may not show benefit in study populations with significant phenotypic variations in etiology, clinical features, and risk of outcomes (13). Cluster analysis may provide a strategy to identify patient populations who should be enrolled within clinical trials and as a complement to traditional subgroup analyses of clinical trials.

To operationalize the use of clusters, specific calculators could be created within electronic health record systems that could assign cluster membership based on available characteristics. Future studies will be needed to define whether implementation of therapeutic decision-making based on cluster membership would change outcomes. Furthermore, whether identification of a few cluster variables is sufficient to define the entire cluster for prognosis or response to therapies warrants further evaluation.

Limitations

These analyses are subject to the limitations of a post hoc evaluation. TECOS enrolled a population with ASCVD, and EXSCEL predominantly had participants with ASCVD (75.1% of total trial population). Hence, our results are not generalizable to populations predominantly without preexisting ASCVD. The use of a selected clinical trial population may not allow for generalizability to a nontrial population The phenotypes ascribed to different clusters are primarily descriptive; however, similar approaches have been used in other cluster analyses (13,14,30,37). The clustering algorithms may have different results depending on the variables used and the quality and completeness of the available data; however, in the context of a CV safety trial, extensive baseline demographics and clinical characteristics, combined with adjudication-confirmed clinical outcomes available for analyses, represent high-quality data. The choice of a stopping rule of the cluster algorithm at five clusters is somewhat arbitrary; although an increased number of clusters would allow for more discrete phenotypes, the smaller number of patients (and therefore events) per cluster would limit evaluation of differential clinical outcomes and treatment differences across clusters. The relatively short follow-up in the clinical trials may not have allowed for findings of greater differences in event rates between the clusters; studies of longer follow-up in diverse populations outside of clinical trials would enable a further evaluation of the utility of clustering analyses. The clustering algorithm did not identify groups that had a reduced risk of CV outcomes with sitagliptin; however, none of the clusters had an increased risk of CV outcomes associated with sitagliptin. These results provide further reassurances of the CV safety of sitagliptin in patients with T2DM and established ASCVD. The replication of our results in EXSCEL, which enrolled patients with T2DM who had ASCVD or multiple CV risk factors, further demonstrates the feasibility of using data-driven approaches to identify phenotypes of patients with T2DM at high risk of CV disease. Furthermore, while our TECOS results were replicated in EXSCEL, the ability to classify patients with T2DM into specific clusters for prognostication and to enable specific medical action requires future validation in prospective studies. Our analysis did not include medication in the clustering, as medication use is often dictated by local practice patterns, guidelines, access to therapies, and resources of countries. Hence, baseline or postbaseline medication use and associations with outcomes may be confounded.

In conclusion, we found that a data-driven algorithm used among patients with T2DM and established ASCVD

identified clusters with unique phenotypes and different risks of CV outcomes. The pathophysiologic mechanisms leading to the development of these phenotypes and the different risk of CV outcomes seen between clusters remain to be confirmed in future studies. The use of clustering algorithms as a tool for risk stratification warrants further evaluation. Identifying patient phenotypes using machine-learning algorithms in order to target specific therapies represents a potential approach to precision medicine and warrants further evaluation.

Funding. The TECOS trial was funded by Merck Sharp & Dohme, a subsidiary of Merck & Co., Inc. The EXSCEL trial was funded by Amylin Pharmaceuticals, a wholly owned subsidiary of AstraZeneca. A.S. received support from the Fonds de Recherche Santé Quebec Junior 1 Clinical Research Scholars program, Alberta Innovates Health Solutions, and the European Society of Cardiology Young Investigator Award. J.B.B. is supported by grants from the National Institutes of Health (UL1TR002489, U01DK098246, UC4DK108612, and U54DK118612), Patient-Centered Outcomes Research Institute, and American Diabetes Association.

Duality of Interest. A.S. reports receiving support from Roche Diagnostics, Boehringer-Ingelheim, Novartis, and Takeda. J.A.E. has received research grant support and/or personal fees from Amgen, AstraZeneca, Bayer, Merck, Novartis, and Servier. J.A.U. has received support from Novartis (clinical trial site); received consultancy from Amgen, Boehringer Ingelheim, Eli Lilly and Company, Janssen, Merck, Novartis, Sanofi Pasteur; received honoraria from the Canadian Cardiovascular Society, Novartis (registry steering committee), and Servier. S.G.G. has received research grant support and/or personal fees from Amgen, AstraZeneca, Bayer, Boehringer Ingelheim, Bristol-Myers Squibb, Daiichi Sankyo, Eli Lilly and Company, Fenix Group International, Ferring Pharmaceuticals, GlaxoSmithKline, Janssen/Johnson & Johnson, Matrizyme, Merck, Novartis, Pfizer, Regeneron Pharmaceuticals, Sanofi, Servier, and Tenax Therapeutics. P.W.A. has received grants, personal fees, and nonfinancial support from Merck and grants from AstraZeneca. J.B.B.'s contracted consulting fees and travel support for contracted activities are paid to the University of North Carolina by Adocia, AstraZeneca, Dance Biopharm, Dexcom, Eli Lilly and Company, Fractyl Health, GI Dynamics, Intarcia Therapeutics, Lexicon Pharmaceuticals, MannKind Corporation, Metavention, NovaTarg Therapeutics, Novo Nordisk, Orexigen Therapeutics, PhaseBio, Sanofi, Senseonics, vTv Therapeutics Inc, and Zafgen; he reports grant support from AstraZeneca, Eli Lilly and Company, Intarcia Therapeutics, Johnson & Johnson, Lexicon Pharmaceuticals, Medtronic, Novo Nordisk, Sanofi, Theracos, Tolerion, and vTv Therapeutics Inc; he is a consultant to Cirius Therapeutics Inc, CSL Behring, Mellitus Health, Neurimmune AG, Pendulum Therapeutics, and Stability Health; and holds stock/options in Mellitus Health, Pendulum Therapeutics, PhaseBio, and Stability Health. J.B.G. has received grants from AstraZeneca, Boehringer Ingelheim, GlaxoSmithKline, and Sanofi and personal fees from AstraZeneca, Merck, Boehringer-Ingelheim, and Novo Nordisk. R.G.J. has received grants or personal fees from Amgen, AstraZeneca, Boehringer Ingelheim, Eli Lilly and Company, Janssen, and Merck. K.D.K. is an employee of Merck & Co., Inc., the manufacturer of sitagliptin, and owns stock and stock options in Merck & Co., Inc. D.K.M. has provided clinical trial leadership for Lilly USA, LLC, AstraZeneca, Sanofi Aventis, Janssen, Boehringer Ingelheim, Merck & Co, Novo Nordisk, Lexicon Pharmaceuticals, Eisai Co., Ltd., GlaxoSmithKline, and Esperion Therapeutics and consultancy for Afimmune, AstraZeneca, Sanofi Aventis, Lilly USA, LLC, Boehringer Ingelheim, Merck & Co, Pfizer, Novo Nordisk, Applied Therapeutics, and Metavant Sciences. G.A. has served on an advisory board and speakers bureau for Merck. R.D.L. has received research support from Bristol-Myers Squibb and GlaxoSmithKline; and personal fees from Bayer, Boehringer Ingelheim, Bristol-Myers Squibb, GlaxoSmithKline, Pfizer, and Portola Pharmaceuticals. E.D.P. has received grants from Janssen, Merck, Sanofi, AstraZeneca, Genentech, and Amgen; and has consulting associations with Janssen, Bayer, Merck, and Sanofi. R.R.H. has received grants and personal fees from Merck; grants from Bayer, AstraZeneca, and Bristol-Myers Squibb; personal fees from Amgen, Bayer, Intarcia Therapeutics, Novartis, and Novo Nordisk; and other support from GlaxoSmithKline, Janssen, and Takeda Pharmaceutical Company. No other potential conflicts of interest relevant to this article were reported.

Author Contributions. A.S. conceived of the analysis, developed the analysis plan, and wrote the initial drafts of the manuscript. Y.Z. conducted the statistical analysis. J.A.E. provided critical review and edited the manuscript. C.M.W. supervised the statistical analysis plan development and statistical analysis. J.A.U. provided critical review and edited the manuscript. S.G.G. provided critical review and edited the manuscript. P.W.A. provided critical review and edited the manuscript. J.B.B. provided critical review and edited the manuscript. J.B.G. provided critical review and edited the manuscript. R.G.J. provided critical review and edited the manuscript. K.D.K. provided critical review and edited the manuscript. D.K.M. provided critical review and edited the manuscript. G.A. provided critical review and edited the manuscript. L.-M.C. provided critical review and edited the manuscript. R.D.L. provided critical review and edited the manuscript. E.D.P. provided critical review and edited the manuscript. R.R.H. provided critical review and edited the manuscript. A.S. and R.R.H. are the guarantors of this work and, as such, had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Prior Presentation. This study was presented at the European Society of Cardiology Congress, Munich, Germany, 25–29 August 2018.

References

1. De Keulenaer GW, Brutsaert DL. Systolic and diastolic heart failure: different phenotypes of the same disease? Eur J Heart Fail 2007;9:136–143
2. Auro K, Joensuu A, Fischer K, et al. A metabolic view on menopause and ageing. Nat Commun 2014;5:4708
3. Creixell P, Reimand J, Haider S, et al.; Mutation Consequences and Pathway Analysis Working Group of the International Cancer Genome Consortium. Pathway and network analysis of cancer genomes. Nat Methods 2015;12:615–621
4. National Research Council (US). *Committee on A Framework for Developing a New Taxonomy of Disease. Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. 2011*. Washington, DC, National Academies Press, 2011
5. Tromp J, Voors AA, Sharma A, et al. Distinct pathological pathways in patients with heart failure and diabetes. JACC Heart Fail 2020;8:234–242
6. Braunwald E. The war against heart failure: the Lancet lecture. Lancet 2015;385:812–824
7. Sharma A, Cooper LB, Fiuzat M, et al. Glucose-lowering therapies to treat patients with heart failure and diabetes mellitus. JACC Heart Fail 2018;6:813–822
8. Zelniker TA, Wiviott SD, Raz I, et al. SGLT2 inhibitors for primary and secondary prevention of cardiovascular and renal outcomes in type 2 diabetes: a systematic review and meta-analysis of cardiovascular outcome trials. Lancet 2019;393:31–39
9. Kristensen SL, Rørth R, Jhund PS, et al. Cardiovascular, mortality, and kidney outcomes with GLP-1 receptor agonists in patients with type 2 diabetes: a systematic review and meta-analysis of cardiovascular outcome trials. Lancet Diabetes Endocrinol 2019;7:776–785
10. Sharma A, Pagidipati NJ, Califf RM, et al. Impact of regulatory guidance on evaluating cardiovascular risk of new glucose-lowering therapies to treat type 2 diabetes mellitus: lessons learned and future directions. Circulation 2020;141:843–862
11. Ashley EA. The precision medicine initiative: a new national effort. JAMA 2015;313:2119–2120
12. Henao R, Murray J, Ginsburg G, Carin L, Lucas JE. Patient clustering with uncoded text in electronic medical records. AMIA Annu Symp Proc 2013;2013:592–599
13. Ahmad T, Pencina MJ, Schulte PJ, et al. Clinical implications of chronic heart failure phenotypes defined by cluster analysis. J Am Coll Cardiol 2014;64:1765–1774
14. Shah SJ, Katz DH, Selvaraj S, et al. Phenomapping for novel classification of heart failure with preserved ejection fraction. Circulation 2015;131:269–279
15. Lindman BR, Dávila-Román VG, Mann DL, et al. Cardiovascular phenotype in HFpEF patients with or without diabetes: a RELAX trial

ancillary study. J Am Coll Cardiol 2014;64:541–549
16. Sharma A, Demissei BG, Tromp J, et al. A network analysis to compare biomarker profiles in patients with and without diabetes mellitus in acute heart failure. Eur J Heart Fail 2017;19:1310–1320
17. Green JB, Bethel MA, Paul SK, et al. Rationale, design, and organization of a randomized, controlled Trial Evaluating Cardiovascular Outcomes with Sitagliptin (TECOS) in patients with type 2 diabetes and established cardiovascular disease. Am Heart J 2013;166:983–989.e7
18. Green JB, Bethel MA, Armstrong PW, et al.; TECOS Study Group. Effect of sitagliptin on cardiovascular outcomes in type 2 diabetes. N Engl J Med 2015;373:232–242
19. Holman RR, Bethel MA, Mentz RJ, et al.; EXSCEL Study Group. Effects of once-weekly exenatide on cardiovascular outcomes in type 2 diabetes. N Engl J Med 2017;377:1228–1239
20. Jolliffe I. *Principal Component Analysis*. New York, Wiley, 2005
21. Agarwal S, Jacobs DR Jr, Vaidya D, et al. Metabolic syndrome derived from principal component analysis and incident cardiovascular events: the Multi Ethnic Study of Atherosclerosis (MESA) and Health, Aging, and Body Composition (Health ABC). Cardiol Res Pract 2012;2012:919425
22. Bergmark BA, Bhatt DL, Braunwald E, et al. Risk assessment in patients with diabetes with the TIMI risk score for atherothrombotic disease. Diabetes Care 2018;41:577–585
23. Frazier-Wood AC, Glasser S, Garvey WT, et al. A clustering analysis of lipoprotein diameters in the metabolic syndrome. Lipids Health Dis 2011;10:237
24. Tajik P, Meijer R, Duivenvoorden R, et al. Asymmetrical distribution of atherosclerosis in the carotid artery: identical patterns across age, race, and gender. Eur J Prev Cardiol 2012;19:687–697
25. Lansky AJ, Ng VG, Maehara A, et al. Gender and the extent of coronary atherosclerosis, plaque composition, and clinical outcomes in acute coronary syndromes. JACC Cardiovasc Imaging 2012;5(Suppl.):S62–S72
26. Yang X, So WY, Tong PC, et al.; Hong Kong Diabetes Registry. Development and validation of an all-cause mortality risk score in type 2 diabetes. Arch Intern Med 2008;168:451–457
27. Noble D, Mathur R, Dent T, Meads C, Greenhalgh T. Risk models and scores for type 2 diabetes: systematic review. BMJ 2011;343:d7163
28. Coiro S, Girerd N, Sharma A, et al. Association of diabetes and kidney function according to age and systolic function with the incidence of sudden cardiac death and non-sudden cardiac death in myocardial infarction survivors with heart failure. Eur J Heart Fail 2019;21:1248–1258
29. Sharma A, Vaduganathan M, Ferreira JP, et al. Clinical and biomarker predictors of expanded heart failure outcomes in patients with type 2 diabetes after a recent acute coronary syndrome: insights from the EXAMINE trial. J Am Heart Assoc 2020;9:e012797
30. Kao DP, Lewsey JD, Anand IS, et al. Characterization of subgroups of heart failure patients with preserved ejection fraction with possible implications for prognosis and treatment response. Eur J Heart Fail 2015;17:925–935
31. Cavender MA, Steg PG, Smith SC Jr, et al.; REACH Registry Investigators. Impact of diabetes mellitus on hospitalization for heart failure, cardiovascular events, and death: outcomes at 4 years from the Reduction of Atherothrombosis for Continued Health (REACH) Registry. Circulation 2015;132:923–931
32. Harumi Higuchi Dos Santos M, Sharma A, Sun JL, et al. International variation in outcomes among people with cardiovascular disease or cardiovascular risk factors and impaired glucose tolerance: insights from the NAVIGATOR Trial. J Am Heart Assoc 2017;6:e003892
33. Stevens RJ, Kothari V, Adler AI; United Kingdom Prospective Diabetes Study (UKPDS) Group. The UKPDS risk engine: a model for the risk of coronary heart disease in type II diabetes (UKPDS 56). Clin Sci (Lond) 2001;101:671–679
34. Sharma A, Sun JL, Lokhnygina Y, et al. Patient phenotypes, cardiovascular risk, and ezetimibe treatment in patients after acute coronary syndromes (from IMPROVE-IT). Am J Cardiol 2019;123:1193–1201
35. McGuire DK, Van de Werf F, Armstrong PW, et al.; Trial Evaluating Cardiovascular Outcomes With Sitagliptin (TECOS) Study Group. Association between sitagliptin use and heart failure hospitalization and related outcomes in type 2 diabetes mellitus. JAMA Cardiol 2016;1:126–135
36. Shah AM, Pfeffer MA. The many faces of heart failure with preserved ejection fraction. Nat Rev Cardiol 2012;9:555–556
37. Marso SP, Daniels GH, Brown-Frandsen K, et al.; LEADER Steering Committee; LEADER Trial Investigators. Liraglutide and cardiovascular outcomes in type 2 diabetes. N Engl J Med 2016;375:311–322
38. Fudim M, White J, Pagidipati NJ, et al. Effect of once-weekly exenatide in patients with type 2 diabetes mellitus with and without heart failure and heart failure-related outcomes: insights from the EXSCEL trial. Circulation 2019;140:1613–1622