JAMIA Open, 4(3), 2021, 1–11
doi: 10.1093/jamiaopen/ooab063
Advance Access Publication Date: 13 July 2021
Research and Applications

Research and Applications

Data-driven identification of temporal glucose patterns in a large cohort of nondiabetic patients with COVID-19 using time-series clustering

Sejal Mistry [1], Ramkiran Gouripeddi[1,2], Julio C. Facelli[1,2]

[1]Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah, USA, [2]Clinical and Translational Science Institute, University of Utah, Salt Lake City, Utah, USA

Corresponding Author: Julio C. Facelli, 412 Wakara Way # 140, Salt Lake City, UT 81408, USA; julio.facelli@utah.edu

Received 10 May 2021; Revised 1 July 2021; Editorial Decision 8 July 2021; Accepted 9 July 2021

ABSTRACT

Objective: Hyperglycemia has emerged as an important clinical manifestation of coronavirus disease 2019 (COVID-19) in diabetic and nondiabetic patients. Whether these glycemic changes are specific to a subgroup of patients and persist following COVID-19 resolution remains to be elucidated. This work aimed to characterize longitudinal random blood glucose in a large cohort of nondiabetic patients diagnosed with COVID-19.

Materials and Methods: De-identified electronic medical records of 7502 patients diagnosed with COVID-19 without prior diagnosis of diabetes between January 1, 2020, and November 18, 2020, were accessed through the TriNetX Research Network. Glucose measurements, diagnostic codes, medication codes, laboratory values, vital signs, and demographics were extracted before, during, and after COVID-19 diagnosis. Unsupervised time-series clustering algorithms were trained to identify distinct clusters of glucose trajectories. Cluster associations were tested for demographic variables, COVID-19 severity, glucose-altering medications, glucose values, and new-onset diabetes diagnoses.

Results: Time-series clustering identified a low-complexity model with 3 clusters and a high-complexity model with 19 clusters as the best-performing models. In both models, cluster membership differed significantly by death status, COVID-19 severity, and glucose levels. Clusters membership in the 19 cluster model also differed significantly by age, sex, and new-onset diabetes mellitus.

Discussion and Conclusion: This work identified distinct longitudinal blood glucose changes associated with subclinical glucose dysfunction in the low-complexity model and increased new-onset diabetes incidence in the high-complexity model. Together, these findings highlight the utility of data-driven techniques to elucidate longitudinal glycemic dysfunction in patients with COVID-19 and provide clinical evidence for further evaluation of the role of COVID-19 in diabetes pathogenesis.

Key words: time-series clusteringCOVID-19diabetes mellitusreal-world data

INTRODUCTION

Coronavirus disease 2019 (COVID-19) is caused by infection with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)[1] and has resulted in over 3 million deaths worldwide as of April 2021.[2] Hyperglycemia, defined as elevated blood glucose, has emerged as an important manifestation of COVID-19 in diabetic and nondiabetic patients.[3–7] Studies have linked hyperglycemia to worse clinical outcomes,[5,8–11] COVID-19 severity,[12–15] exacerbation of diabetes symptoms,[16–19] and mortality,[20,21] but few have explored whether these glycemic changes persist following resolution of COVID-19.

LAY SUMMARY

Hyperglycemia is defined as elevated blood glucose measurements and is common in diabetic patients. Recent findings suggest that patients diagnosed with coronavirus disease 2019 (COVID-19) may experience elevated blood glucose levels during COVID-19 infection. Whether blood glucose levels remain elevated after COVID-19 infection remains poorly understood. This study aimed to identify how patterns of blood glucose levels before, during, and after COVID-19 change. This work analyzed blood glucose levels from 7,502 patients diagnosed with COVID-19 and used machine learning to identify different patterns of blood glucose changes. We found patterns demonstrating an overall increase, overall, decrease, temporary increase, and temporary decrease in blood glucose levels. Some of these patterns were associated with COVID-19 severity and proportion of patients with new-onset diabetes. These findings demonstrate the usefulness of machine learning in understanding glucose changes following COVID-19 diagnosis and indicate that more research is needed to understand if blood glucose monitoring in COVID-19 patients should be routinely performed.

Recent case reports of new-onset diabetes[19,22,23] have increased speculation of long-term glycemic dysfunction following COVID-19 infection. Coronaviruses have previously been associated with dysregulation of glucose metabolism and transient hyperglycemia.[24–26] SARS-CoV-2 is thought to mediate impaired beta-cell function through angiotensin-converting enzyme 2 receptor-induced inflammatory responses in the pancreas.[24,27–34] SARS-CoV-2-associated inflammation may also impair insulin signaling in adipose tissue,[35] resulting in hyperglycemia and insulin resistance.[29] Recent findings suggest a pathogenic role of COVID-19 in inducing diabetes through beta-cell transdifferentiation[36] and direct beta-cell apoptosis.[37] Together, these findings indicate potential mechanisms for persistent COVID-19 induced glucose dysfunction.

Evidence evaluating longitudinal effects of COVID-19 on blood glucose has been difficult to ascertain. Considerable heterogeneity observed in patients who experienced hyperglycemia during COVID-19[38] suggests that there may be subgroups of patients with similar glucose changes. Reported incidences of new-onset diabetes mellitus also varied, as some studies found an increased prevalence during peak periods of the pandemic,[39] others reported no change[40] or a decrease.[41] The CoviDIAB Project was established to define the phenotype of COVID-19-related new-onset diabetes,[42] but does not consider transient or subclinical instances of glucose dysfunction. Previous studies with small sample sizes and short time frames of analysis highlight the need for big data repositories in evaluating long-term effects of COVID-19 and identifying subphenotypes of glucose dysregulation.

This work aims to characterize longitudinal glucose changes in a large cohort of patients with COVID-19. De-identified electronic medical records were utilized to train unsupervised machine learning models to identify glucose trajectories. The goal of this work was to provide a foundational framework for data-driven methods in identifying clinically useful stratifications of blood glucose.

METHODS

In this section, we briefly describe the methods used in this work. For a detailed and reproducible account, see Supplementary Methods.

Study cohort

TriNetX, a health research network whose functionalities are described elsewhere[43,44] (Supplementary Methods), was queried for COVID-19 patients confirmed by ICD-10 codes or laboratory testing indicating the presence of SARS-CoV-2 between January 20, 2020, and November 18, 2020 (Supplementary Table S1). De-identified data were obtained for 256,566 patients with confirmed COVID-19. Patients with at least 1 glucose measurement defined by laboratory codes (Supplementary Table S1) in each of the following timepoints were included: before (2 years to 1 week before COVID-19 diagnosis), during (1 week before to 2 weeks after COVID-19 diagnosis), and after (2 weeks to 1 year after COVID-19 diagnosis). Patients with a history of type 1 or type 2 diabetes defined by ICD-9 or ICD-10 codes prior to COVID-19 diagnosis were excluded (Supplementary Table S1). Using these inclusion and exclusion criteria, a cohort of 7502 patients was assembled (Supplementary Figure S1). The study was deemed exempt by the University of Utah Institutional Review Board.

Preprocessing

Random blood glucose measurements were extracted using LOINC value sets for glucose, and 3 trajectory datasets were created: segment, 3-month, and 6-month (Supplementary Methods). The segment trajectory dataset was created by averaging glucose at each timepoint (before, during, and after). The 3-month trajectory dataset was created by averaging glucose over 3 months' intervals at each timepoint. The 6-month trajectory dataset was created by averaging glucose over 6 months' intervals at each timepoint. All trajectory datasets were scaled using a mean-variance scaler (tslearn 0.4.1 in Python[45]).

Clustering

Unsupervised k-means clustering was performed to identify distinct glucose trajectories (Supplementary Methods). In this work, 2 k-means unsupervised algorithms were trained: time-series k-means (ts k-means) and kernel k-means[46,47] (kernel) (tslearn 0.4.1 in Python[45]). Clustering models were created for each unique combination of trajectory datasets (segment, 3-month, and 6-month) and clustering algorithm (ts k-means and kernel) (Supplementary Figure S2). Clustering was performed for k of 2–20 and performance was evaluated using the Silhouette Score (SS) (scikit-learn 0.24.1 in Python[48,49]) (Supplementary Methods). The model with the best SS score as confirmed by elbow method heuristics was selected for further analysis. Trajectories assigned to each cluster and cluster centers were plotted.

Phenotype analysis

Data extraction

Medications, diagnoses, vital signs, and demographics facts were downloaded for all 7502 patients in the cohort (Supplementary Figure S3). Standardized terminologies and terminology harmonization methods are described in the Supplementary Methods.

luster interpretation

Clusters with similar centers were grouped and features were extracted for each group (Supplementary Methods). Categorical features included the proportion of patients with age older than 65, female sex, death, moderate disease severity, severe disease severity, critical disease severity, hyperglycemic before, during, and after COVID-19, antihyperglycemic agents and insulin, glucose, steroids, and new-onset type 1 or type 2 diabetes (Supplementary Table S2). Continuous features included age, age at death, glucose before, during, and after COVID-19 diagnosis.

Statistical analysis

Continuous features were assessed for normality using D'Agostino's test[50,51] (scipy 1.6.0 in Python[52]) and independence using Kruskal–Wallis *H*-test[53] (Pingouin 0.3.9 in Python[54]). Categorical features were assessed for independence using Chi-square tests with Yates continuity correction (Pingouin 0.3.9 in Python[54]). Categorical variables were also evaluated for Cramer's *V* effect size[55] (Pingouin 0.3.9 in Python[54]). *P*-values for continuous and categorical variables were considered significant at the .05 level. Statistically significant categorical features were further analyzed by scaling the proportions for each feature and visualized using a radar diagram (plotly v4.14.3 in Python[56]).

RESULTS

Cohort characteristics

For the 7502 patients with COVID-19, the mean (standard deviation [SD]) age was 56.25 (19.25) years, 56% ($n = 4197$) were female, 7% ($n = 514$) died, and the mean (SD) death age was 66.72 (17.50) years. From the cohort, 8% ($n = 578$) of patients were classified as having critical, 3% ($n = 244$) as severe, and 4% ($n = 277$) as moderate COVID-19 severity; 7% ($n = 494$) patients were prescribed antihyperglycemic agents and insulin, 12% ($n = 915$) were prescribed glucose, and 14% ($n = 1080$) were prescribed steroids.

The mean (SD) glucose before COVID-19 diagnosis was 6.23 (2.01) mmol/L and 21% (1592) were hyperglycemic. The mean (SD) glucose during COVID-19 diagnosis was 6.06 (1.71) mmol/L and 14% ($n = 1084$) were hyperglycemic. The mean (SD) glucose after COVID-19 diagnosis was 6.42 (1.97) mmol/L and 24% ($n = 1824$) were hyperglycemic; 1.4% ($n = 103$) of patients diagnosed developed new-onset diabetes during or after COVID-19.

Clustering results

A total of 114 models with unique trajectories (segment, 3-month, or 6-month) and algorithms (ts k-means or kernel) were trained on clusters ranging from 2 to 20. The comparison of the best-performing clusters from each model (Supplementary Table S4) clearly

Table 1. Statistical summary of phenotypic features of 3 clusters from the segment and k-means model

Feature		Cluster 1 Risers ($n = 2976$)	Cluster 2 Decliners ($n = 2341$)	Cluster 3 Peakers ($n = 2185$)	*P*-value
Demographics					
Age	Median (IQR) (years)	59 (27)	58 (29)	58 (30)	$P = .082$
	% >65 years (*n*)	36% (1085)	34% (802)	34% (738)	$P = .091$
Sex	% Female (*n*)	56% (1657)	56% (1312)	56% (1228)	$P = .926$
Death	Median (IQR) (years)	70 (22)	71 (21)	70 (24.5)	$P = .798$
	% Death (*n*)	6% (193)	10% (233)	4% (88)	**$P < .001$**
					$V = .092$
COVID-19 severity					
Moderate	% (*n*)	4% (130)	4% (82)	3% (65)	**$P = .027$**
					$V = .031$
Severe	% (*n*)	4% (111)	3% (65)	3% (68)	$P = .137$
Critical	% (*n*)	9% (276)	8% (183)	5% (119)	**$P < .001$**
					$V = .059$
Glucose-altering medications					
Anti-hyperglycemic agents and insulin	% (*n*)	7% (208)	7% (153)	6% (133)	$P = .432$
Glucose	% (*n*)	12% (361)	13% (293)	12% (261)	$P = .833$
Steroids	% (*n*)	15% (449)	15% (342)	13% (289)	$P = .160$
Glucose levels					
Before	Median (IQR) (mmol/L)	5.55 (1.28)	6.60 (1.83)	5.28 (1.05)	**$P < .001$**
	% Hyperglycemic (*n*)	15% (445)	40% (925)	9% (201)	**$P < .001$**
					$V = .313$
During	Median (IQR) (mmol/L)	5.55 (1.05)	5.55 (1.00)	6.16 (1.39)	**$P < .001$**
	% Hyperglycemic (*n*)	10% (290)	10% (223)	26% (561)	**$P < .001$**
					$V = .208$
After	Median (IQR) (mmol/L)	6.66 (1.94)	5.72 (1.33)	5.44 (1.11)	**$P < .001$**
	% Hyperglycemic (*n*)	40% (1191)	17% (395)	10% (219)	**$P < .001$**
					$V = .309$
New-onset diabetes diagnoses					
Diabetes	% (*n*)	1% (40)	1% (34)	1% (29)	$P = .923$

Note: Statistical summary of all features is presented for each cluster. Differences in continuous variables were tested using Kruskal–Wallis *H*-test. Differences between clusters in categorical variables were tested using the Chi-squared test for independence with Yates Continuity Correction and Cramer's *V* for effect size was calculated for statistically significant features. *P*-values were evaluated at the .05 significance level, and the bold text indicates statistical significance.

Abbreviations: COVID-19: coronavirus disease 2019; IQR: interquartile range.

Table 2. Statistical summary of phenotypic features of 19 clusters from the segment and k-means model

Feature		Steady risers ($n = 891$)	Delayed risers ($n = 1583$)	Early risers ($n = 690$)	Steady decliners ($n = 346$)	Delayed decliners ($n = 578$)	Early decliners ($n = 1189$)	Peakers ($n = 1392$)	Valleyers ($n = 833$)	*P*-value
Demographics										
Age	Median (IQR) (years)	58 (31)	59 (27)	58 (31)	54 (32)	56 (32)	58 (28)	58 (28)	62 (26)	**$P < .001$**
	% > 65 years (*n*)	36% (318)	36% (570)	34% (233)	31% (106)	31% (180)	35% (416)	33% (465)	40% (337)	**$P = .004$**
										$V = .052$
Sex	% Female (*n*)	59% (530)	54% (855)	56% (383)	60% (208)	60% (345)	55% (653)	56% (773)	54% (450)	**$P = .042$**
										$V = .044$
Death	Median (IQR) (years)	67 (22)	70 (25)	76 (21)	68 (16)	73 (24)	71 (25)	65 (23)	70 (16)	$P = .578$
	% Death (*n*)	4% (33)	6% (99)	3% (24)	6% (22)	5% (30)	11% (130)	4% (53)	15% (123)	**$P < .001$**
										$V = .147$
COVID-19 severity										
Moderate	% (*n*)	4% (33)	5% (73)	2% (15)	2% (15)	3% (16)	3% (40)	3% (47)	6% (46)	**$P = .004$**
										$V = .053$
Severe	% (*n*)	4% (34)	4% (61)	4% (26)	3% (10)	3% (16)	3% (33)	3% (39)	3% (25)	$P = .563$
Critical	% (*n*)	8% (74)	10% (157)	5% (37)	4% (15)	4% (25)	7% (86)	6% (84)	12% (100)	**$P < .001$**
										$V = .089$
Medications										
Anti-hyperglycemic agents or insulin	% (*n*)	7% (62)	7% (107)	5% (36)	6% (22)	5% (27)	6% (72)	7% (97)	9% (71)	$P = .106$
Glucose	% (*n*)	13% (113)	12% (185)	10% (69)	12% (42)	12% (70)	12% (139)	13% (177)	14% (120)	$P = .321$
Steroids	% (*n*)	16% (142)	15% (233)	13% (89)	16% (56)	12% (72)	14% (165)	14% (193)	16% (130)	$P = .395$
Glucose levels										
Before	Median (IQR) (mmol/L)	5.16 (0.89)	5.72 (1.22)	5.05 (0.94)	6.33 (1.44)	5.94 (1.17)	6.88 (2.05)	5.39 (1.05)	6.55 (1.94)	**$P < .001$**
	% Hyperglycemic (*n*)	5% (44)	16% (261)	6% (42)	29% (102)	16% (94)	47% (559)	10% (141)	39% (328)	**$P < .001$**
										$V = .370$
During	Median (IQR) (mmol/L)	5.72 (1.00)	5.44 (1.00)	5.94 (1.17)	5.72 (0.89)	5.88 (1.11)	5.49 (1.00)	6.33 (1.61)	5.33 (1.00)	**$P < .001$**
	% Hyperglycemic (*n*)	11% (96)	8% (129)	19% (134)	12% (41)	15% (87)	8% (96)	31% (426)	8% (65)	**$P < .001$**
										$V = .243$
After	Median (IQR) (mmol/L)	6.27 (1.44)	6.88 (2.11)	5.83 (1.11)	5.22 (0.89)	5.11 (0.89)	5.77 (1.17)	5.38 (1.00)	6.60 (2.05)	**$P < .001$**
	% Hyperglycemic (*n*)	29% (259)	46% (736)	17% (115)	7% (25)	5% (30)	15% (175)	9% (132)	40% (333)	**$P < .001$**
										$V = .359$
New-onset diabetes diagnoses										
Diabetes	% (*n*)	1% (10)	2% (25)	1% (4)	0% (1)	1% (3)	2% (22)	2% (23)	2% (15)	**$P = .048$**
										$V = .043$

Note: Statistical summary of all features is presented for each cluster. Differences in continuous variables were tested using Kruskal–Wallis *H*-test. Differences between clusters in categorical variables were tested using the Chi-squared test for independence with Yates Continuity Correction and Cramer's *V* for effect size was calculated for statistically significant features. *P*-values were evaluated at the .05 significance level, and the bold text indicates statistical significance.

Abbreviations: COVID-19: coronavirus disease 2019; IQR: interquartile range.

shows the superiority of the segment approach. The model with the best SS corresponds to the segment trajectory dataset, ts k-means, and 19 clusters (SS = 0.53). However, elbow method evaluation of the segment trajectory dataset and ts k-means models demonstrated a positive inflection point at 3 clusters (SS = 0.52) (Supplementary Figure S4). Therefore, both models were selected for further analysis as exemplars of low-complexity with $k = 3$ and high-complexity clustering with $k = 19$.

Three clusters model

Cluster visualization

Cluster trajectories and centers from the 3 cluster model, segment trajectory dataset, and ts k-means model are depicted in Figure 1. Three patterns were visually identified according to the cluster center trajectory. Cluster 1 was labeled as "Risers" because its center demonstrated an increasing pattern in scaled glucose during and after COVID-19 diagnosis (Figure 1A, top). Cluster 2 was labeled as "Decliners" because its center demonstrated a decreasing pattern in scaled glucose during and after COVID-19 diagnosis (Figure 1B, top). Cluster 3 was labeled as "Peakers" because its center demonstrated a peak in scaled glucose only during COVID-19 diagnosis (Figure 1C, top).

Phenotype results

Summary statistics for all features were calculated for each cluster (Table 1).

The proportion of patients who died was significantly different between clusters ($P < .001$, $V = .092$), with the "Decliners" containing a greater proportion of patients who died compared to the "Risers" and "Peakers" (10%, $n = 233$; 6%, $n = 193$; and 4%, $n = 88$, respectively). Moderate and critical COVID-19 severity were also significantly different between clusters ($P = .027$, $V = .031$ and $P < .001$, $V = .059$, respectively), with the "Peakers" containing the lowest proportion of moderate and critical cases (3%, $n = 65$ and 5% $n = 119$, respectively).

Glucose levels ($P < .001$ for all 3 comparisons) and the proportion of patients with hyperglycemic glucose measurements before, during, and after COVID-19 diagnosis ($P < .001$ for all 3 comparisons, $V = .313$, .208, and .309, respectively) were significantly different between clusters. The "Decliners" demonstrated the highest median and greatest proportion of hyperglycemia before COVID-19 diagnosis (median (interquartile range [IQR]) = 6.60 (1.83) mmol/L; 40%, $n = 925$). The "Peakers" demonstrated the highest median and greatest proportion of hyperglycemia during COVID-19 diagnosis (median (IQR) = 6.16 (1.39) mmol/L; 26%, $n = 561$), while the "Risers" demonstrated the highest median and greatest proportion of hyperglycemia after COVID-19 diagnosis (median (IQR) = 6.66 (1.94) mmol/L; 40%, $n = 1191$).

Radar visualization

Significantly different scaled categorical features were displayed in a radar diagram for each cluster and features with a value greater than 0.5 are reported. "Risers" were characterized by hyperglycemia after COVID-19 diagnosis and moderate and critical COVID-19 severity (Figure 1a, bottom). "Decliners" were characterized by the greatest proportion of patients who died and hyperglycemia before COVID-19 diagnosis (Figure 1B, bottom). "Peakers" were characterized by hyperglycemia during COVID-19 diagnosis (Figure 1C, bottom).

Nineteen clusters model

Cluster visualization

Cluster trajectories and centers from the 19 cluster model, segment trajectory dataset, and ts k-means model are depicted in Figure 2, from which 8 general patterns were visually identified according to cluster centers.

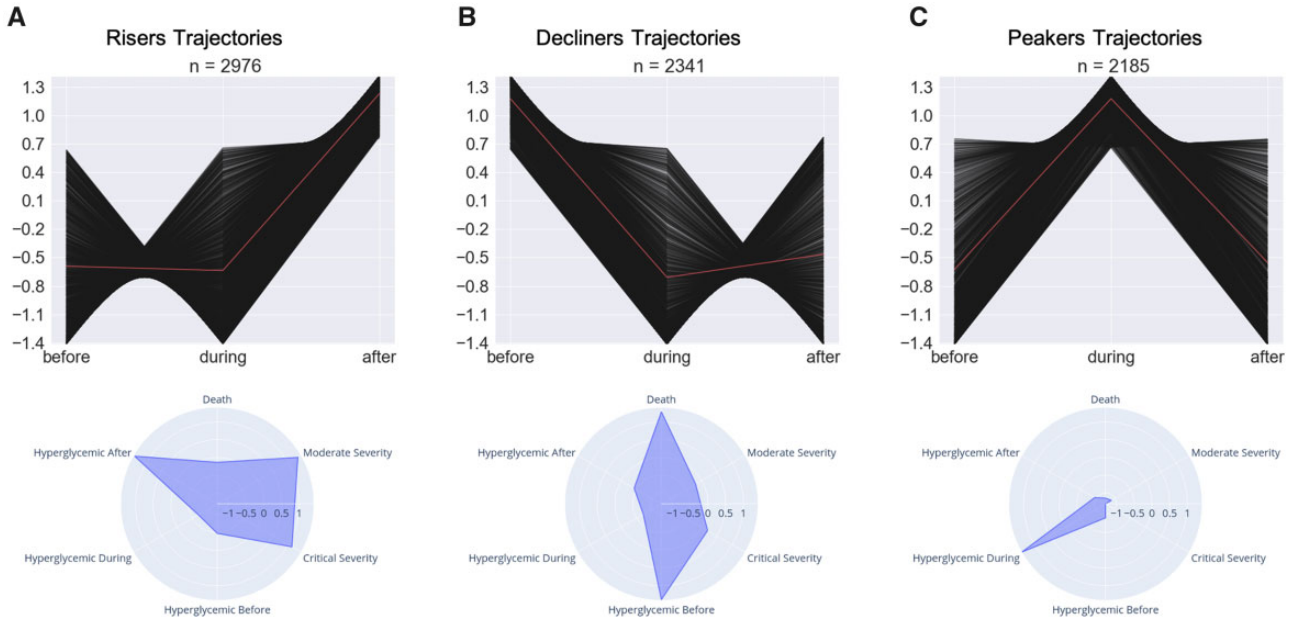

Figure 1. Glucose trajectories and radar plots of phenotypic features of 3 clusters from the segment and k-means model. (A) Glucose trajectory of the 'Risers' cluster with black lines representing individual trajectories and the red line representing cluster centers (top) and radar plot of the scaled statistically significant features (bottom). (B) Glucose trajectory of the 'Decliners' cluster with black lines representing individual trajectories and the red line representing cluster centers (top) and radar plot of the scaled statistically significant features (bottom). (C) Glucose trajectory of the 'Peakers' cluster with black lines representing individual trajectories and the red line representing cluster centers (top) and radar plot of the scaled statistically significant features (bottom).

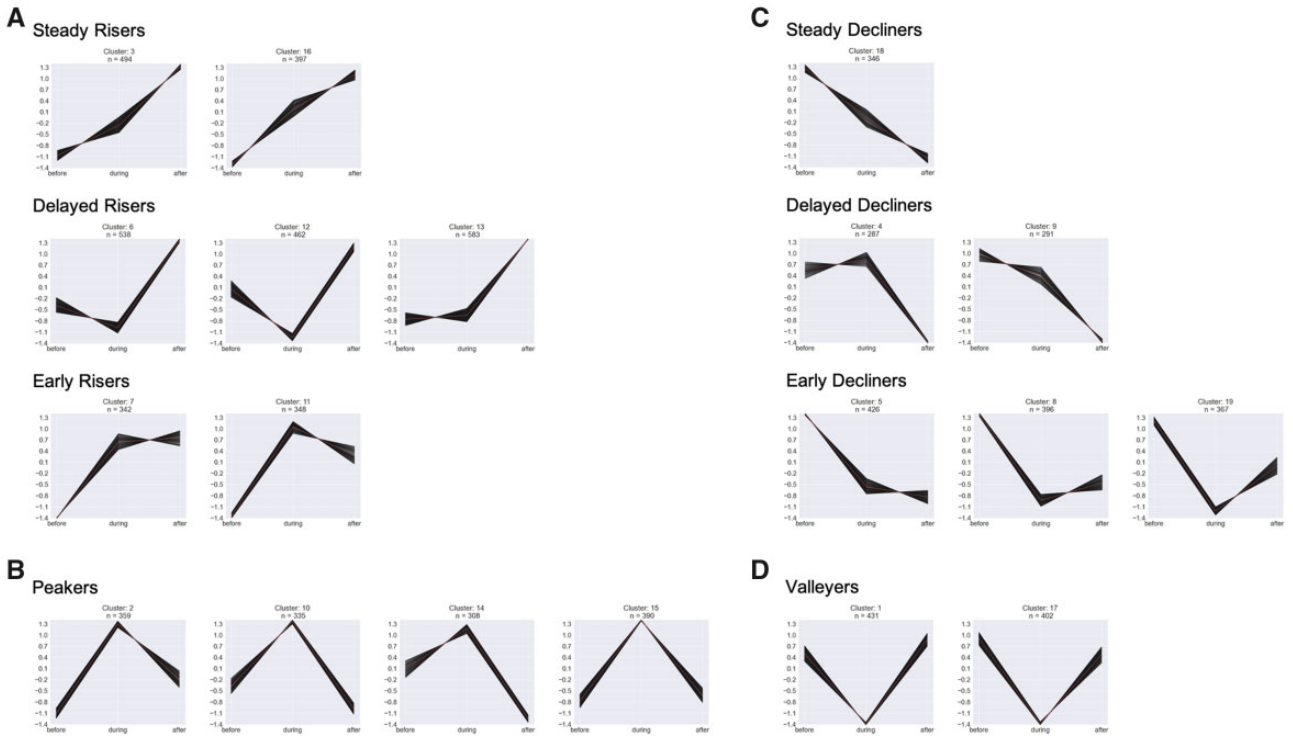

Figure 2. Glucose trajectories of the 19 clusters from the segment and k-means model. (A) Glucose trajectory of the 'Risers' cluster with black lines representing individual trajectories and the red line representing cluster centers. (B) Glucose trajectory of the 'Peakers' cluster with black lines representing individual trajectories and the red line representing cluster centers. (C) Glucose trajectory of the 'Decliners' cluster with black lines representing individual trajectories and the red line representing cluster centers. (D) Glucose trajectory of the 'Valleyers' cluster with black lines representing individual trajectories and the red line representing cluster centers.

The first set of clusters were labeled as "Risers" because all cluster means demonstrated an overall increase in scaled glucose values and varied by timing of that increase (Figure 2A). Clusters 3 and 16 were included into the "Steady Risers" because their centers demonstrated a steady increase in scaled glucose during and after COVID-19 diagnosis. Clusters 6, 12, and 13 were included into the "Delayed Risers" because their centers demonstrated a delayed increase in scaled glucose that occurred after COVID-19 diagnosis. Clusters 7 and 11 were included into the "Early Risers" because their centers demonstrated an early increase in scaled glucose during and persisted after COVID-19 diagnosis. The next set of clusters were labeled as "Peakers" because their centers demonstrated a peak in scaled glucose during COVID-19 diagnosis that subsequently returned to baseline (Figure 2B). Clusters 2, 10, 14, and 15 were included into the "Peakers."

The next set of clusters were labeled as "Decliners" because their centers demonstrated an overall decrease in scaled glucose values and varied by timing of that decrease (Figure 2C). Cluster 18 was included in "Steady Decliners" because its center demonstrated a steady decline in scaled glucose during and after COVID-19 diagnosis. Clusters 4 and 9 were included in "Delayed Decliners" because their centers demonstrated a delayed decline in scaled glucose after COVID-19 diagnosis. Clusters 5, 8, and 19 were included into the "Early Decliners" because their centers demonstrated an early decline in scaled glucose that occurred during and persisted following COVID-19 diagnosis. The final set of clusters were labeled as "Valleyers" because their centers demonstrated a drop in scaled glucose during COVID-19 diagnosis that returned to baseline afterward (Figure 2D). Clusters 1 and 17 were included into the "Valleyers."

Phenotype results

Summary statistics for all features were calculated for each cluster (Table 2).

Age ($P < .001$) and proportion of patients greater than 65 years ($P = .004$, $V = .052$) were significantly different among groups, with "Valleyers" having the oldest and greatest proportion of patients older than 65 (median (IQR) = 62 (26) years; 40%, $n = 337$). The proportion of females were also significantly different among groups, though the effect size was small ($P = .042$, $V = .044$). The proportion of patients who died was significantly differ among groups ($P < .001$, $V = .147$), with the greatest proportion of death observed for the "Valleyers" (15%, $n = 123$). Moderate and critical COVID-19 severity were significantly different among groups ($P = .004$, $V = .053$ and $P < .001$, $V = .089$, respectively). "Valleyers" and "Delayed Risers" demonstrated the greatest proportion of patients with moderate (6%, $n = 46$ and 5%, $n = 73$, respectively) and critical severity (12%, $n = 100$, and 10%, $n = 157$, respectively).

Glucose levels ($P < .001$ for all 3 comparisons) and the proportion of patients with hyperglycemia before, during, and after COVID-19 diagnosis ($P < .001$ for all 3 comparisons, $V = .370$, .242, and .359, respectively) were significantly different between groups. The greatest median glucose and highest proportion of hyperglycemia were attributed to "Early Decliners" before COVID-19 diagnosis (median (IQR) = 6.88 (2.05) mmol/L; 47%, $n = 559$, respectively), "Peakers" during COVID-19 diagnosis (median (IQR) = 6.33 (1.61) mmol/L; 31%, $n = 426$, respectively), and "Delayed Risers" after COVID-19 diagnosis (median (IQR) = 6.88 (2.11) mmol/L; 46%, $n = 736$, respectively). Diagnosis of diabetes mellitus

was also significantly different between groups ($P = .048$, $V = .043$), with "Delayed Risers" containing the greatest number of patients with new-onset Diabetes (2%, $n = 25$).

Radar visualization

Significantly different scaled categorical features were displayed in a radar diagram for each group (Figure 3) and features with a value greater than 0.5 are reported. "Steady Risers" were characterized by female sex and hyperglycemia after COVID-19 diagnosis, "Delayed Risers" were characterized by patients aged 65 or older, moderate and critical COVID-19 severity, hyperglycemia after COVID-19 diagnosis, and new-onset diabetes mellitus, and "Early Risers" were characterized by hyperglycemia during COVID-19 diagnosis (Figure 3A). "Peakers" were characterized by hyperglycemia during COVID-19 and new-onset diabetes mellitus (Figure 3B). "Steady Decliners" were characterized by female sex and hyperglycemia before COVID-19 diagnosis, "Delayed Decliners" were characterized by female sex, and "Early Decliners" were characterized by death, hyperglycemia before COVID-19 diagnosis, and new-onset diabetes mellitus (Figure 3C). "Valleyers" were characterized by patients aged 65 or older, death, moderate and critical COVID-19 severity, hyperglycemia before and after COVID-19, and new-onset diabetes mellitus (Figure 3D).

DISCUSSION

This study provides the first data-driven assessment of random blood glucose measurements in a large, multicenter cohort of patients diagnosed with COVID-19 in the United States. Time-series clustering identified 2 optimal models for characterizing glucose trajectories: a low-complexity model with 3 clusters and a high-complexity model with 19 clusters. Cluster membership for both models was associated with proportions of patients with moderate and critical COVID-19 severity, mortality, and hyperglycemia at each timepoint, and mean glucose values at each timepoint (Tables 1 and 2). The 19 cluster model was additionally associated with the proportion of patients with age older than 65 years, female sex, and new-onset diabetes mellitus diagnosis, and mean age at death (Table 2). These results highlight the utility of using unsupervised clustering to learn novel patterns of glucose changes and call for longitudinal glucose monitoring in patients with COVID-19.

The simpler 3 cluster model consisted of "Risers," "Decliners," and "Peakers." Findings from the "Risers" suggest that there is a group of individuals who experience moderate or critical COVID-19 with associated increases in blood glucose after resolution of disease. Though this does not indicate an increase in new-onset diabetes mellitus, this work shows that subclinical glucose dysfunction can occur in patients following COVID-19. The "Decliners" demonstrated that elevated glucose before COVID-19 diagnosis is associated with death. These findings support previous evidence that hyperglycemia before COVID-19 indicates a poor prognosis.[3,57–59] Finally, findings from the "Peakers" suggest that a group of patients experience elevated blood glucose during COVID-19 diagnosis that is not characterized by moderate and critical COVID-19 severity, supporting previous evidence that coronaviruses can induce transient hyperglycemia.[24] The 19 cluster model consisted of 8 groups that captured a wide variety of subtle changes in glucose trajectories that remained masked in the 3 cluster model. The "Early Decliners," "Valleyers," "Delayed Risers," and "Peakers" were characterized by a sharp peak in diabetes diagnosis. Of these 4 groups, only the "During Decliners" and "Delayed Risers" were associated with moderate or critical COVID-19 severity. Together, these findings suggest that glucose trajectories may be associated with increased diabetes incidence following SARS-CoV-2 infection in both symptomatic and asymptomatic patients.

This work highlights the potential for data-driven techniques to elucidate longitudinal glycemic dysfunction in patients with COVID-19. Unlike previous studies characterizing glucose in COVID-19,[3,9,20,60] this work utilizes electronic medical records from a large cohort of patients across the United States. Presentation of a low- and high-complexity model emphasizes the heterogeneity in glucose curves of patients with COVID-19. The presented analytical approach can serve as a framework for future data-driven analyzes in longitudinal glycemic evaluation of COVID-19 patients. In addition, the presented analysis demonstrates the utility of real-world clinical data in longitudinal research analyses. This work also captures COVID-19 cases from a broader range of pandemic, as previous studies are limited by short time frames that often coincide with peak periods of the pandemic.[5,8] There is a dearth of knowledge on the long-term outcomes of patients with COVID-19 because longitudinal studies of the clinical sequela are limited, and this study contributes to the longitudinal understanding of glycemic changes following COVID-19 diagnosis.

To the best of our knowledge, this is the first work assessing the role of COVID-19 in the etiology of diabetes using real-world clinical data. Many viruses, including enterovirus, Coxsackievirus B, rotavirus, mumps virus, cytomegalovirus, and rubella, have been associated with diabetes pathogenesis in epidemiological studies.[61,62] Similarly, these results provide hypotheses for further investigation into why certain individuals may develop diabetes following a COVID-19 infection. Though this work follows individuals for a short period of time surrounding COVID-19 infection, these results highlight the need for longer-term follow-ups and the development of trajectories for other diabetes-related parameters, as evidenced by previous studies investigating body mass index and glycosylated hemoglobin.[63–66] Initiatives such as the National COVID Cohort Collaborative (N3C) provide opportunities for addresses longer-term follow-up and diabetes-related clinical consequences of COVID-19.[67] N3C leverages the vast amount of electronic medical record collected through the course of the pandemic and aims to facilitate novel scientific discoveries to aid in responding to the pandemic.

There are several limitations of this work. First, it is possible that fewer individuals with asymptomatic or mild disease sought medical care at large healthcare organizations that are represented by TriNetX. Due to inherent limitations in electronic medical record data, distinguishing asymptomatic and mild COVID-19 was not possible. In addition, recent work suggests that asymptomatic COVID-19 cases are more prevalent than previously thought.[68] Therefore, these results may represent glycemic changes in a more severe spectrum of COVID-19 cases. Next, this study only included individuals with glucose measurements before, during, and after COVID-19, resulting in a significant reduction of patients from the original cohort (Supplementary Figure S1). Duration of COVID-19 is highly variable in patients and our stratification may oversimply the COVID-19 time course. However, this stratification of glucose timepoints was necessary because electronic medical record data are highly variable and identifying a consistent time axis for analysis was not possible. This was partially accounted for by comparing the performance of multiple time intervals, and our findings call for longitudinal cohort studies in patients diagnosed with COVID-19 with consistently and uniformly collected glucose measurements. In addition, these inclusion criteria may favor patients who receive care

Figure 3. Radar plot of phenotypic features of the 19 clusters from the segment and k-means model. (A) Radar plot of the scaled statistically significant features of the 'Risers' cluster. (B) Radar plot of the scaled statistically significant features of the 'Peakers' cluster. (C) Radar plot of the scaled statistically significant features of the 'Decliners' cluster. (D) Radar plot of the scaled statistically significant features of the 'Valleyers' cluster.

continuously from a single medical center. Further work is needed to validate the impact of COVID-19 on longitudinal glucose changes in populations who do not receive consistent care. Additionally, there were a variable number of glucose measurements for each patient included in generating the segment trajectory dataset (Supplementary Methods). A subset of patients only contained single random glucose levels per timepoint, which may under-represent overall glucose levels for the given patients. Of the 7502 patients in-

cluded in this study, 7230 patients had more than 1 glucose measurement per timepoint, suggesting that the bias due to the number of random blood glucose measurements may be minimal. Given that the average time to COVID-19 symptom onset is 5 days[69] and viral detection 2 weeks after symptom onset was minimal,[70] we defined 3 weeks to encompass COVID-19 duration. However, the time to recovery for COVID-19 patients is variable and more robust methods may be needed to capture disease duration. Random blood glucose measurements were selected as they are appropriate estimates for glycemic control in noninsulin-dependent patients.[71] However, their use is imprecise, and further studies should be conducted using fasting blood glucose measurements in well-defined cohorts recruited for specific studies. While our cluster groups were derived using subjective methods, we plan to validate these phenotypes using external data-centric methods[72] with availability of longer post-COVID records for these patients in TriNetX and N3C.

CONCLUSIONS

This work utilizes time-series unsupervised clustering techniques to identify distinct groups of glucose trajectories. We provide a detailed analysis of 2 models corresponding to 3 and 19 clusters. Both models provide insights on how cluster features can be associated with clinical phenotypes and expected outcomes. This work demonstrates that unsupervised learning models of different complexities can be used to provide a clinically useful stratification of glucose trajectories in patients diagnosed with COVID-19. Further subtyping and longer follow-up can identify subphenotypes of patients who are more susceptible to glycemic dysregulation following COVID-19 infection and inform appropriate point-of-care guidelines for COVID-19 management.

FUNDING

This work was supported by the National Library of Medicine (T15LM5920634 to SM) and the National Center for Advancing Translational Sciences (UL1TR002538). Computational resources were provided by the University of Utah Center for High Performance Computing, partially funded by the NIH Shared Instrumentation Grant (1S10OD021644-01A1). De-identified data were provided by TriNetX, a global health research network compliant with Section 164.514(b)(1) of the HIPAA Privacy Rule.

AUTHOR CONTRIBUTIONS

All authors conceived the work, SM and RG secured the dataset, and SM performed all analysis and drafted the initial draft. All authors contributed and approved the final draft.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY STATEMENT

The data underlying this article was accessed from the TriNetX Research Network and was provided by TriNetX under license/by permission. Data will be shared on request to the corresponding author with the permission of TriNetX. The query criteria utilized in this research are provided in Supplementary Table S1 and specifics will be shared at reasonable request to the corresponding author.

REFERENCES

1. Kuiken T, Fouchier RA, Schutten M, *et al.* Newly discovered coronavirus as the primary cause of severe acute respiratory syndrome. *Lancet* 2003; 362 (9380): 263–70.
2. Medicine JHU and COVID-19 Dashboard by the Center for System Science and Engineering. https://coronavirus.jhu.edu/map.html Accessed February 23, 2021.
3. Zhang Y, Li H, Zhang J, *et al.* The clinical characteristics and outcomes of patients with diabetes and secondary hyperglycaemia with coronavirus disease 2019: a single-centre, retrospective, observational study in Wuhan. *Diabetes Obes Metab* 2020; 22 (8): 1443–54.
4. Sachdeva S, Desai R, Gupta U, *et al.* Admission hyperglycemia in non-diabetics predicts mortality and disease severity in COVID-19: a pooled analysis and meta-summary of literature. *SN Compr Clin Med* 2020; 2 (11): 2161–6.
5. Bode B, Garrett V, Messler J, *et al.* Glycemic characteristics and clinical outcomes of COVID-19 patients hospitalized in the United States. *J Diabetes Sci Technol* 2020; 14 (4): 813–21.
6. Ebekozien OA, Noor N, Gallagher MP, *et al.* Type 1 diabetes and COVID-19: preliminary findings from a multicenter surveillance study in the U.S. *Diabetes Care* 2020; 43 (8): e83–5.
7. Iacobellis G, Penaherrera CA, Bermudez LE, *et al.* Admission hyperglycemia and radiological findings of SARS-CoV2 in patients with and without diabetes. *Diabetes Res Clin Pract* 2020; 164: 108185.
8. Fadini GP, Morieri ML, Boscari F, *et al.* Newly-diagnosed diabetes and admission hyperglycemia predict COVID-19 severity by aggravating respiratory deterioration. *Diabetes Res Clin Pract* 2020; 168: 108374.
9. Shen Y, Fan X, Zhang L, *et al.* Thresholds of glycemia and the outcomes of COVID-19 complicated with diabetes: a retrospective exploratory study using continuous glucose monitoring. *Diabetes Care* 2021; 44 (4): 976–82.
10. Atkins JL, Masoli JAH, Delgado J, *et al.* Preexisting comorbidities predicting COVID-19 and mortality in the UK Biobank community cohort. *J Gerontol Ser* 2020; 75: glaa183.
11. Roncon L, Zuin M, Rigatelli G, *et al.* Diabetic patients with COVID-19 infection are at higher risk of ICU admission and poor short-term outcome. *J Clin Virol* 2020; 127: 104354.
12. Li X, Xu S, Yu M, *et al.* Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan. *J Allergy Clin Immunol* 2020; 146 (1): 110–8.
13. Zhu B, Jin S, Wu L, *et al.* J-shaped association between fasting blood glucose levels and COVID-19 severity in patients without diabetes. *Diabetes Res Clin Pract* 2020; 168: 108381.
14. Liu Q, Chen H, Li J, *et al.* Fasting blood glucose predicts the occurrence of critical illness in COVID-19 patients: a multicenter retrospective cohort study. *J Infect* 2020; 81 (3): e20–3.
15. Gregory JM, Slaughter JC, Duffus SH, *et al.* COVID-19 severity is tripled in the diabetes community: a prospective analysis of the pandemic's impact in type 1 and type 2 diabetes. *Diabetes Care* 2021; 44 (2): 526–32.
16. Li J, Wang X, Chen J, *et al.* COVID-19 infection may cause ketosis and ketoacidosis. *Diabetes Obes Metab* 2020; 22 (10): 1935–41.
17. Dżygało K, Nowaczyk J, Szwilling A, *et al.* Increased frequency of severe diabetic ketoacidosis at type 1 diabetes onset among children during COVID-19 pandemic lockdown: an observational cohort study. *Pediatr Endocrinol Diabetes Metab* 2020; 26 (4): 167–75.
18. Orioli L, Hermans MP, Thissen J-P, *et al.* COVID-19 in diabetic patients: related risks and specifics of management. *Ann Endocrinol (Paris)* 2020; 81 (2–3): 101–9.
19. Chee YJ, Ng SJH, Yeoh E. Diabetic ketoacidosis precipitated by Covid-19 in a patient with newly diagnosed diabetes mellitus. *Diabetes Res Clin Pract* 2020; 164: 108166.

20. Mazori AY, Bass IR, Chan L, *et al.* Hyperglycemia is associated with increased mortality in critically ill patients with COVID-19. *Endocr Pract* 2021; 27 (2): 95–100.
21. de Jong M, Woodward M, Peters SAE. Diabetes and COVID-19-related mortality in women and men in the UK Biobank: comparisons with influenza/pneumonia and coronary heart disease. *Diabetes Care* 2021; 44 (2): e22–4.
22. Reddy PK, Kuchay MS, Mehta Y, *et al.* Diabetic ketoacidosis precipitated by COVID-19: a report of two cases and review of literature. *Diabetes Metab Syndr* 2020; 14 (5): 1459–62.
23. Gupta A, Madhavan MV, Sehgal K, *et al.* Extrapulmonary manifestations of COVID-19. *Nat Med* 2020; 26 (7): 1017–32.
24. Yang J-K, Lin S-S, Ji X-J, *et al.* Binding of SARS coronavirus to its receptor damages islets and causes acute diabetes. *Acta Diabetol* 2010; 47 (3): 193–9.
25. Fignani D, Licata G, Brusco N, *et al.* SARS-CoV-2 receptor angiotensin I-converting enzyme type 2 (ACE2) is expressed in human pancreatic β-cells and in the human pancreas microvasculature. *Front Endocrinol (Lausanne)* 2020; 11: 596898.
26. Yang JK, Feng Y, Yuan MY, *et al.* Plasma glucose levels and diabetes are independent predictors for mortality and morbidity in patients with SARS. *Diabet Med* 2006; 23 (6): 623–8.
27. Yang L, Han Y, Nilsson-Payant BE, *et al.* A human pluripotent stem cell-based platform to study SARS-CoV-2 tropism and model virus infection in human cells and organoids. *Cell Stem Cell* 2020; 27 (1): 125–36.e7.
28. Hamming I, Timens W, Bulthuis M, *et al.* Tissue distribution of ACE2 protein, the functional receptor for SARS coronavirus. A first step in understanding SARS pathogenesis. *J Pathol* 2004; 203 (2): 631–7.
29. Sathish T, Tapp RJ, Cooper ME, *et al.* Potential metabolic and inflammatory pathways between COVID-19 and new-onset diabetes. *Diabetes Metab* 2021; 47 (2): 101204.
30. Yao X, Li T, He Z, *et al.* A pathological report of three COVID-19 cases by minimal invasive autopsies. *Zhonghua Bing Li Xue Za Zhi* 2020; 49 (5): 411–7.
31. Drucker DJ. Coronavirus infections and type 2 diabetes-shared pathways with therapeutic implications. *Endocr Rev* 2020; 41 (3): 457–70.
32. Pal R, Bhansali A. COVID-19, diabetes mellitus and ACE2: the conundrum. *Diabetes Res Clin Pract* 2020; 162: 108132.
33. Hussain A, Bhowmik B, do Vale Moreira NC. COVID-19 and diabetes: knowledge in progress. *Diabetes Res Clin Pract* 2020; 162: 108142.
34. Singh AK, Gupta R, Ghosh A, Misra A. Diabetes in COVID-19: prevalence, pathophysiology, prognosis and practical considerations. *Diabetes Metab Syndr* 2020; 14 (4): 303–10.
35. Hotamisligil G, Shargill N, Spiegelman B. Adipose expression of tumor necrosis factor-alpha: direct role in obesity-linked insulin resistance. *Science* 1993; 259 (5091): 87–91.
36. Tang X, Uhl S, Zhang T, *et al.*. SARS-CoV-2 infection induces beta cell transdifferentiation. *Cell Metab* 2021; 33 (8): 1577–91.e7.
37. Wu C-T, Lidsky PV, Xiao Y, *et al.* SARS-CoV-2 infects human pancreatic β-cells and elicits β-cell impairment. *Cell Metab* 2021; 33 (8): 1565–76.e5.
38. Misra A, Ghosh A, Gupta R. Heterogeneity in presentation of hyperglycaemia during COVID-19 pandemic: a proposed classification. *Diabetes Metab Syndr* 2021; 15 (1): 403–6.
39. Unsworth R, Wallace S, Oliver NS, *et al.* New-onset type 1 diabetes in children during COVID-19: multicenter regional findings in the U.K. *Diabetes Care* 2020; 43 (11): e170–1.
40. Tittel SR, Rosenbauer J, Kamrath C, *et al.*; DPV Initiative. Did the COVID-19 lockdown affect the incidence of pediatric type 1 diabetes in Germany? *Diabetes Care* 2020; 43 (11): e172–3.
41. Rabbone I, Schiaffini R, Cherubini V, *et al.*; Diabetes Study Group of the Italian Society for Pediatric Endocrinology and Diabetes. Has COVID-19 delayed the diagnosis and worsened the presentation of type 1 diabetes in children? *Diabetes Care* 2020; 43 (11): 2870–2.
42. Rubino F, Amiel SA, Zimmet P, *et al.* New-onset diabetes in Covid-19. *N Engl J Med* 2020; 383 (8): 789–90.
43. Taquet M, Luciano S, Geddes JR, *et al.* Bidirectional associations between COVID-19 and psychiatric disorder: retrospective cohort studies of 62354 COVID-19 cases in the USA. *Lancet Psychiatry* 2021; 8 (2): 130–40.
44. Harrison PJ, Luciano S, Colbourne L. Rates of delirium associated with calcium channel blockers compared to diuretics, renin-angiotensin system agents and beta-blockers: an electronic health records network study. *J Psychopharmacol* 2020; 34 (8): 848–55.
45. Tavenard R, Faouzi J, Vandewiele G, *et al.* Tslearn, a machine learning toolkit for time series data. *J Mach Learn Res* 2020; 21: 1–6.
46. Cuturi M. Fast global alignment kernels. In: 28th International Conference on Machine Learning. 2011; Bellevue, WA.
47. Dhillon IS, Guan Y, Kulis B. Kernel k-means: spectral clustering and normalized cuts. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2004: 551–6; Seattle, WA.
48. Buitinck L, Louppe G, Blondel M, *et al.* API design for machine learning software: experiences from the scikit-learn project. In: ECML PKDD Workshop: Languages for Data Mining and Machine Learning. 2013:108–22.
49. Pedregoda F, Gramfort VG, *et al.* Scikit-learn: machine learning in python. *J Mach Learn Res* 2011; 12: 2825–30.
50. D'Agostino RB. An omnibus test of normality for moderate and large size samples. *Biometrika* 1971; 58: 341–8.
51. D'Agostino R, Pearson ES. Tests for departure from normality. Empirical results for the distributions of b_2 and $\sqrt{b_1}$. *Biometrika* 1973; 60 (3): 613–22.
52. Virtanen P, Gommers R, Oliphant TE, *et al.*; SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 2020; 17 (3): 261–72.
53. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. *J Am Stat Assoc* 1952; 47 (260): 583–621.
54. Vallat R. Pingouin: statistics in Python. *J Open Source Softw* 2018; 3 (31): 1026.
55. Cramér H. *Mathematical Methods of Statistics (PMS-9)*. Princeton, NJ: Princeton University Press; 1999.
56. Plotly Technologies Inc. *Collaborative Data Science*. Montreal, QC: Plotly Technologies Inc.; 2015.
57. Li H, Tian S, Chen T, *et al.* Newly diagnosed diabetes is associated with a higher risk of mortality than known diabetes in hospitalized patients with COVID. *Diabetes Obes Metab* 2020; 22 (10): 1897–906.
58. Liu S, Zhang Q, Wang W, *et al.* Hyperglycemia is a strong predictor of poor prognosis in COVID-19. *Diabetes Res Clin Pract* 2020; 167: 108338.
59. Ceriello A. Hyperglycemia and the worse prognosis of COVID-19. Why a fast blood glucose control should be mandatory. *Diabetes Res Clin Pract* 2020; 163: 108186.
60. Sardu C, D'Onofrio N, Balestrieri ML, *et al.* Outcomes in patients with hyperglycemia affected by COVID-19: can we do more on glycemic control? *Diabetes Care* 2020; 43 (7): 1408–15.
61. Hyöty H, Taylor K. The role of viruses in human diabetes. *Diabetologia* 2002; 45 (10): 1353–61.
62. Filippi CM, von Herrath MG. Viral trigger for type 1. *Diabetes* 2008; 57 (11): 2863–71.
63. Clements MA, Schwandt A, Donaghue KC, *et al.*; Australasian Diabetes Data Network (ADDN) Study Group, the T1D Exchange Clinic Network (T1DX), and the German/Austrian/Luxembourgian Diabetes-Patienten-Verlaufsdokumentation (DPV) Initiative. Five heterogeneous HbA1c trajectories from childhood to adulthood in youth with type 1 diabetes from three different continents: a group-based modeling approach. *Pediatr Diabetes* 2019; 20 (7): 920–31.
64. Miller RG, Orchard TJ, Onengut-Gumuscu S, *et al.* Heterogeneous long-term trajectories of glycaemic control in type 1 diabetes. *Diabetic Med* 2021; 38 (8): e14545.
65. Zhang T, Xu J, Li S, *et al.* Trajectories of childhood BMI and adult diabetes: the Bogalusa Heart Study. *Diabetologia* 2019; 62 (1): 70–7.

66. Nano J, Dhana K, Asllanaj E, *et al.* Trajectories of BMI before diagnosis of type 2 diabetes: the Rotterdam study. *Obesity (Silver Spring)* 2020; 28 (6): 1149–56.
67. Haendel MA, Chute CG, Bennett TD, *et al.* The National COVID Cohort Collaborative (N3C): rationale, design, infrastructure, and deployment. *J Am Med Inform Assn* 2020; 28: ocaa196.
68. Kalish H, Klumpp-Thomas C, Hunsberger S, *et al.* Undiagnosed SARS-CoV-2 seropositivity during the first six months of the COVID-19 pandemic in the United States. *Sci Transl Med* 2021; 13 (601): eabh3826.
69. Lauer SA, Grantz KH, Bi Q, *et al.* The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Ann Intern Med* 2020; 172 (9): 577–82.
70. Walsh KA, Jordan K, Clyne B, *et al.* SARS-CoV-2 detection, viral load and infectivity over the course of an infection. *J Infect* 2020; 81 (3): 357–71.
71. Gill GV, Hardy KJ, Patrick AW, *et al.* Random blood glucose estimation in type 2 diabetes: does it reflect overall glycaemic control? *Diabetic Med* 1994; 11 (7): 705–8.
72. Ma E-Y, Kim J-W, Lee Y, *et al.* Combined unsupervised-supervised machine learning for phenotyping complex diseases with its application to obstructive sleep apnea. *Sci Rep* 2021; 11 (1): 4457.