ARIC Manuscript Proposal Form

ARIC Publication Admin Use Only

Publication Committee Review Date: 8/12/2025
ARIC Manuscript Proposal Number#: 4723

1.a. Full Title: Identifying Glycemic Profiles in People with Type 2 Diabetes via Clustering of Continuous Glucose Monitoring Data

b. Abbreviated Title (Length 26 characters): CGM Clustering in T2D

2. Writing Group [please provide a middle initial if available; EX: Adam L Williams]:
Writing group numbers: Jiahuan Helen He, Dan Wang, Mary R. Rooney, Amelia S. Wallace, R. Nisha Aurora, Naresh M. Punjabi, Scott L. Zeger, Mike Fang, Elizabeth Selvin, and others

I, the first author, confirm that all the coauthors have given their approval for this manuscript proposal. **[please confirm with your initials electronically or in writing]**

First author [please provide a middle initial; EX: Adam L Williams]:

Jiahuan Helen		He
First name	Middle initial	Last name

Address:
Phone: 4436351186
E-mail: jhe75@jh.edu

ARIC sponsor to be contacted if there are questions about the manuscript and the first author does not respond or cannot be located. The ARIC sponsor is responsible for **carefully reviewing the proposal and ensuring its soundness prior to submission to the ARIC Publications Committee.** (The ARIC sponsor should be involved enough in ARIC to be able to point the lead author to appropriate ancillary study PIs and to be able to search ARIC manuscript proposals if the lead author doesn't have the access needed to do such a search).

Name: Liz Selvin
E-mail: eselvin@jhu.edu

3. Timeline: Statistical analyses and manuscript writing are planned to be done within 1 year after the approval

4. Rationale:

Continuous glucose monitors (CGMs) are increasingly used in diabetes management for people with type 2 diabetes.[1] These devices provide detailed information on nuanced glucose patterns and can inform personalized diabetes management decisions, including medication timing, meal planning, and physical activity.[3]

CGMs measure interstitial glucose every 1-15 minutes for up to 14 days, allowing patients to observe glucose changes in real time.[2] CGMs enable the detection of postprandial glucose spikes, hypoglycemic episodes, and glycemic variability.

Due to the length of wearing time and frequency of measurements, CGMs produce a prolific amount of data – up to ~20,000 data points for some devices. To obtain interpretable information from these devices, the American Diabetes Association (ADA) incorporated the guidance from the International Consensus on CGM and have recommended 8 glycemic measures for clinical use, including time in range (TIR) 70-180 mg/dL, two time below range (TBR) metrics of <70 mg/dL and <54 mg/dL, two time above range (TAR) metrics of >180 mg/dL and >250 mg/dL, mean glucose, and glucose coefficient of variation (CV), and glucose management indicator (GMI) which is the estimated hemoglobulin A1c value derived from CGM data.[4,5] The 2025 ADA Standards of Care specifically states that individuals with diabetes wearing a CGM should aim for >70% of TIR 70-180 mg/dL and <5% and <1% of TBR <70 mg/dL and TBR <54 mg/dL. should be minimized to be less than 5% and 1%, respectively, to reduce hypoglycemic risk.[4]

While these metrics have been shown to be associated with clinical outcomes, they do not fully take advantage of the rich CGM data. The large amount and high temporal resolution makes CGM data well-suited for data-driven approaches to uncover latent glycemic profiles. One such approach is clustering, which classifies individuals into subgroups based on patterns of input data, grouping those who are more similar to each other than to others, without requiring predefined labels.[6] Several studies[7,8,9,10,11,12] have identified subtypes of type 2 diabetes using clustering methods. Most of these studies have used clinical biomarker and/or genetic data as input features for clustering.[13] Because CGM measures the downstream dynamics of glycemic status, these profiles reflect the differences not only in underlying physiology, but also in treatment regimens and self-management behaviors, and therefore they may provide a more actionable and personalized view of glycemic status and clinical risk. While a few emerging studies have used data from CGMs to define glycemic subgroups in individuals with type 2

diabetes,[14,15] these studies relied on an extensive number of CGM metrics that far exceed what has been recommended in the clinical guideline. This could introduce redundancy and lower clinical interpretability. Furthermore, given the limited number of these studies, there is a need for external validation to assess the reproducibility of the identified clusters.

To this end, the primary aim of our study is to identify distinct glycemic profiles through clustering on a parsimonious set of CGM metrics with high clinical interpretability in people with type 2 diabetes from the Hyperglycemic Profiles in Obstructive Sleep Apnea (HYPNOS) randomized clinical trial.[16] We will externally validate identified profiles using data from ARIC Generation 2 (Gen2) cohort. Because CGM captures short-term glycemic variability and hypoglycemic episodes which are not reflected by HbA1c, the gold standard for glycemic status assessment[4,17], our secondary aim is to examine whether the CGM glycemic profiles can provide information beyond HbA1c.

5. Main Hypothesis/Study Aims:

1) To identify distinct glycemic profiles in individuals with type 2 diabetes via clustering on CGM data from the HYPNOS trial and validate the generalizability and reproducibility of these profiles in ARIC Gen2 cohort

2) To evaluate whether CGM-derived glycemic profiles could offer information beyond HbA1c

6. Design and analysis - please address the following aspects:

a) **Inclusion/exclusion**

We will include participants from ARIC Gen2 with type 2 diabetes and not using insulin who have at least 5 days of CGM data. We exclude participants using insulin to make the external validation cohort (ARIC Gen2) more similar to the study cohort (HYPNOS).

b) **Study design**

This will be a cross-sectional study involving data from the baseline visit of ARIC Gen2, which will be used as an external validation cohort.

c) **Outcome and other variables of interest with specific reference to the time of their collection**

CGM Metrics in ARIC Gen2

We will use CGM data collected via Abbott Freestyle Libre Pro and Abbott Freestyle Libre3 sensors in ARIC Gen2. From the 5-14 days of CGM data in each participant, we will derive 26 CGM features that have high clinical translatability and

reflect the overall glucose level and glycemic variability (**Table 1**). Eight are recommended by the ADA Standards of Care: time in range (TIR 70-180 mg/dL), time below range (TBR <70 mg/dL and <54 mg/dL), time above range (TAR >180 mg/dL and >250 mg/dL), mean glucose, GMI, and glucose CV. In addition, TIR 70-140 mg/dL, TBR <60 mg/dL, TAR >140 mg/dL will be included to capture more granular levels of hypo- and hyperglycemic severity. To reflect diurnal variation, we will also include daytime (6am-23pm) and nighttime (23pm-2nd day 6am)-specific metrics for TIR, TBR, and TAR. Standard deviation (SD) will also be included as a compliment to CV to reflect glycemic variability as clinicians are often more familiar with SD and may use it as a key secondary glycemic variability metric. [5]

Table 1. CGM features selected for clustering analysis

ADA Recommended CGM metrics:			
1.TIR 70-180 mg/dL	2.TBR <70 mg/dL	3.TBR <54 mg/dL	4.TAR >180 mg/dL
5.TAR >250 mg/dL	6.Mean glucose	7.GMI	8.Glucose CV
Additional CGM metrics for more levels of hypo-/hyperglycemic severity, secondary metrics for glycemic variability (standard deviation), and diurnal variation:			
9.TIR 70-140 mg/dL	10.TBR <60 mg/dL	11.TAR >140 mg/dL	12.Standard deviation
13.Daytime TIR 70-180 mg/dL	14.Nighttime TIR 70-180 mg/dL	15.Daytime TBR <70 mg/dL	16.Nighttime TBR <70 mg/dL
17.Daytime TBR <60 mg/dL	18.Nighttime TBR <60 mg/dL	19.Daytime TBR <54 mg/dL	20.Nighttime TBR <54 mg/dL
21.Daytime TAR >140 mg/dL	22.Nighttime TAR >140 mg/dL	23.Daytime TAR >180 mg/dL	24.Nighttime TAR >180 mg/dL
25.Daytime TAR >250 mg/dL	26.Nighttime TAR >250 mg/dL		

CGM: Continuous Glucose Monitoring; ADA: American Diabetes Association; TIR: Time in range; TBR: Time below range; TAR: Time above range; GMI: Glucose management indicator (estimated A1c from CGM data); CV: Coefficient of variation.

Demographics, HbA1c, and Medication Use in ARIC Gen2

Age, sex, race-center were collected via self-report at baseline.

HbA1c was measured using Tosoh G8 HPLC analyzer with a coefficient of variation <2%.

Diabetes medication use was determined using the ARIC Gen2 (overlapping with ARIC Visit 10) Medication Survey. Participants brought all medications taken in the prior four weeks, which were recorded by trained study staff through barcode scanning or manual transcription following standardized protocols.

d) **Summary of data analysis**

Clustering in HYPNOS

We will apply robust and sparse k-means clustering (RSKC) algorithm[18] with iterative refinement to identify distinct glycemic profiles using the 26 CGM features in HYPNOS. The robustness parameter was set to zero, meaning no observations were trimmed as outliers. To leverage the sparsity of the algorithm, we imposed an L1 constraint on feature weights, allowing down-weighting less informative features and prioritizing those most relevant for distinguishing clusters. Given the high correlation among the 26 CGM features, the sparsity constraint mitigated redundancy from features carrying overlapping information. The number of clusters and the sparsity parameter (L1) were jointly optimized using a grid search procedure that selected the combination yielding the highest silhouette score, with iterative refinement applied when a single dominant cluster emerged.

To assess the stability of the clusters, we will repeat the full two-round RSKC procedure with random sampling of 80% of the original data 50 times and comparing the resulting cluster assignments. Pairwise adjusted Rand index values will be calculated for each layer to quantify the consistency across runs, with higher value indicating greater stability.

To characterize the clusters, we used the key CGM metrics that were recommended by ADA guideline (i.e., TIR 70-180 mg/dL, TBR <70 mg/dL, TBR <54 mg/dL, TAR >180 mg/dL, TAR >250 mg/dL, mean glucose, GMI, and CV) to label them into clinically interpretable glycemic profiles. To assess whether the identified glycemic profiles captured information in addition to HbA1c, we will display the distribution of HbA1c across glycemic profiles in box plots.

We will compare the demographics, impacts of fatigue on daily activities, clinical characteristics, and medication use across glycemic profiles using ANOVA for continuous variables and Pearson's chi-squared test for categorical variables.

External Validation – ARIC Gen2

To externally validate the clusters derived from HYPNOS CGM data, we will apply a result-based validation approach described in Ullmann et al. (2021).[19] We will first standardize the CGM features of ARIC Gen2 using the mean and standard deviation from the HYPNOS dataset. Each ARIC Gen2 participant will then be assigned to the HYPNOS cluster with the closest centroid (i.e., the average values of CGM features across all individuals in that cluster), and these are the projected clusters. To assess agreement, we will compare the projected cluster assignments with de novo clusters generated independently within ARIC Gen2 using the same RSKC method. A heatmap of overlapping proportions will be used to visualize concordance between the projected and the de novo clusters, and the adjusted Rand index will be calculated to quantify the level of agreement.

All analyses will be performed in R version 4.2.2.

e) **Any anticipated methodologic limitations or challenges if present** No

f) **Will the author need Limited data to complete the proposed manuscript?**

☐ Yes, Limited data is needed. (Provide a brief (2-3 sentences) justification for requesting PHF data)

☒ No, De-identified data will be sufficient.

Please note, Limited dataset access is strict and rarely provided. Limited data includes identifiable information such as dates (birthdays, visit dates, etc.). CMS, Genomic, Geocoded, Proteomics/Somalogic, and other -omic data all fall under the limited data category. De-identified data does not include dates. All dates are date adjusted to "Days since Visit 1".

7. Will the data be used by a for-profit organization in this manuscript? ☐ Yes* ☒ No

8. Will the DNA data be used in this manuscript? ☐ Yes* ☒ No

* Please carefully review the instructions and documentation on the ARIC website at: https://aric.cscc.unc.edu/aric9/researchers/Obtain_Submit_Data [ARIC website > Researchers > Obtain/Submit Data] for instructions on proper data use.

9. The lead author or the ARIC sponsor of this manuscript proposal has reviewed the list of existing ARIC Study manuscript proposals and has found no overlap between this proposal and previously approved manuscript proposals either published or still in active status. ARIC Investigators have access to the publications lists under the Study Members Area of the website at: https://aric.cscc.unc.edu/aric9/proposalsearch [ARIC Website >Publications >Proposal Search]

☒ Yes ☐ No

10. What are the most related manuscript proposals in ARIC (authors are encouraged to contact lead authors of these proposals for comments on the new proposal or collaboration)?

4360

11.a. Is this manuscript proposal associated with any ARIC ancillary studies or does it use current [or ongoing] ancillary study data (this includes ACHIEVE)?

☒ Yes ☒ No→ **Skip to question 12**

11.b. If yes to 11.a., is the proposal

☒ A. primarily the result of an ancillary study

☐ B. primarily based on ARIC data with ancillary data playing a minor role (usually control variables)

11.c. If yes to 11.a., list number* 2020.26

*ancillary studies are listed by number
https://aric.cscc.unc.edu/aric9/researchers/ancillary_studies/approved_ancillary_studies [ARIC Website >Ancillary Studies >Approved Ancillary Studies]

12a. Manuscript preparation is expected to be completed in one to three years. If a manuscript is not submitted for ARIC review at the end of the 3-years from the date of the approval, the manuscript proposal will expire.

12b. The NIH instituted a Public Access Policy in April, 2008 which ensures that the public has access to the published results of NIH funded research. It is **your responsibility to upload manuscripts to PubMed Central** whenever the journal does not and be in compliance with this policy. The NIH-NHLBI Public Access Brochure about the public access policy from http://publicaccess.nih.gov/ is posted in https://aric.cscc.unc.edu/aric9/publications/policies_forms_and_guidelines [ARIC Website > Publications >Publication Policies, Forms, and Guidelines]. http://publicaccess.nih.gov/submit_process_journals.htm shows you which journals automatically upload articles to PubMed central.

References:

1. Shields S, Thomas R, Durham J, Moran J, Clary J, Ciemins EL. Continuous glucose monitoring among adults with type 2 diabetes receiving noninsulin or basal insulin therapy in primary care. *Sci Rep*. 2024;14(1):31990. doi:10.1038/s41598-024-83548-4

2. Bergenstal RM. Understanding Continuous Glucose Monitoring Data. In: *Role of Continuous Glucose Monitoring in Diabetes Treatment*. American Diabetes Association; 2018. Accessed July 28, 2024. http://www.ncbi.nlm.nih.gov/books/NBK538967/

3. Klonoff DC, Ahn D, Drincic A. Continuous glucose monitoring: A review of the technology and clinical use. *Diabetes Research and Clinical Practice*. 2017;133:178-192. doi:10.1016/j.diabres.2017.08.005

4. American Diabetes Association Professional Practice Committee. 6. Glycemic Goals and Hypoglycemia: Standards of Care in Diabetes—2025. *Diabetes Care*. 2024;48(Supplement_1):S128-S145. doi:10.2337/dc25-S006

5. Battelino T, Danne T, Bergenstal RM, et al. Clinical Targets for Continuous Glucose Monitoring Data Interpretation: Recommendations From the International Consensus on Time in Range. *Diabetes Care*. 2019;42(8):1593-1603. doi:10.2337/dci19-0028

6. Loftus TJ, Shickel B, Balch JA, et al. Phenotype clustering in health care: A narrative review for clinicians. *Front Artif Intell*. 2022;5. doi:10.3389/frai.2022.842306

7. Ahlqvist E, Storm P, Käräjämäki A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *Lancet Diabetes Endocrinol*. 2018;6(5):361-369. doi:10.1016/S2213-8587(18)30051-2

8. Wagner R, Heni M, Tabák AG, et al. Pathophysiology-based subphenotyping of individuals at elevated risk for type 2 diabetes. *Nat Med*. 2021;27(1):49-57. doi:10.1038/s41591-020-1116-9

9. Udler MS, Kim J, Grotthuss M von, et al. Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. *PLOS Medicine*. 2018;15(9):e1002654. doi:10.1371/journal.pmed.1002654

10. Wesolowska-Andersen A, Brorsson CA, Bizzotto R, et al. Four groups of type 2 diabetes contribute to the etiological and clinical heterogeneity in newly diagnosed individuals: An IMI DIRECT study. *Cell Rep Med*. 2022;3(1):100477. doi:10.1016/j.xcrm.2021.100477

11. Suzuki K, Hatzikotoulas K, Southam L, et al. Genetic drivers of heterogeneity in type 2 diabetes pathophysiology. *Nature*. 2024;627(8003):347-357. doi:10.1038/s41586-024-07019-6

12. Nair ATN, Wesolowska-Andersen A, Brorsson C, et al. Heterogeneity in phenotype, disease progression and drug response in type 2 diabetes. *Nat Med*. 2022;28(5):982-988. doi:10.1038/s41591-022-01790-7

13. Misra S, Wagner R, Ozkan B, et al. Precision subclassification of type 2 diabetes: a systematic review. *Commun Med (Lond)*. 2023;3:138. doi:10.1038/s43856-023-00360-3

14. Tao R, Yu X, Lu J, et al. Multilevel clustering approach driven by continuous glucose monitoring data for further classification of type 2 diabetes. *BMJ Open Diabetes Res Care*. 2021;9(1):e001869. doi:10.1136/bmjdrc-2020-001869

15. Healey E, Morato C, Murillo J, Kohane I. Heterogeneity of continuous glucose monitoring features and their clinical associations in a type 2 diabetes population. *Diabetes, Obesity and Metabolism*. 2025;27(7):3957-3966. doi:10.1111/dom.16432

16. Rooney MR, Aurora RN, Wang D, Selvin E, Punjabi NM. Rationale and design of the Hyperglycemic Profiles in Obstructive Sleep Apnea (HYPNOS) trial. *Contemporary Clinical Trials*. 2021;101:106248. doi:10.1016/j.cct.2020.106248

17. Chehregosha H, Khamseh ME, Malek M, Hosseinpanah F, Ismail-Beigi F. A View Beyond HbA1c: Role of Continuous Glucose Monitoring. *Diabetes Ther*. 2019;10(3):853-863. doi:10.1007/s13300-019-0619-1

18. Kondo Y, Salibian-Barrera M, Zamar R. RSKC: An R Package for a Robust and Sparse K-Means Clustering Algorithm. *Journal of Statistical Software*. 2016;72:1-26. doi:10.18637/jss.v072.i05

19. Ullmann T, Hennig C, Boulesteix AL. Validation of cluster analysis results on validation data: A systematic framework. *WIREs Data Mining and Knowledge Discovery*. 2022;12(3):e1444. doi:10.1002/widm.1444

To access publications materials, click "Log in" at the top-right corner of the ARIC website, then navigate to Publications >Publication Policies, Forms, and Guidelines in the top menu bar.

Click "Forgot Password" if you are experiencing issues with logging in. For assistance with publications, please contact the Publication Administrator at ARICPUB@unc.edu.