A Report on "Multimodal AI Correlates of Glucose Spikes in People with Normal Glucose Regulation, Pre-diabetes and Type 2 Diabetes" by Carletti et al. (2025)

Reviewer 2

February 04, 2026

v1

isitcredible.com

Disclaimer

This report was generated by large language models, overseen by a human editor. It represents the honest opinion of The Catalogue of Errors Ltd, but its accuracy should be verified by a qualified expert. Comments can be made here. Any errors in the report will be corrected in future revisions.

I am wiser than this person; for it is likely that neither of us knows anything fine and good, but he thinks he knows something when he does not know it, whereas I, just as I do not know, do not think I know, either. I seem, then, to be wiser than him in this small way, at least: that what I do not know, I do not think I know, either.

Plato, *The Apology of Socrates*, 21d

To err is human. All human knowledge is fallible and therefore uncertain. It follows that we must distinguish sharply between truth and certainty. That to err is human means not only that we must constantly struggle against error, but also that, even when we have taken the greatest care, we cannot be completely certain that we have not made a mistake.

Karl Popper, 'Knowledge and the Shaping of Reality'

Overview

Citation: Carletti, M., Pandit, J., Gadaleta, M., Chiang, D., Delgado, F., Quartuccio, K., Fernandez, B., Garay, J. A. R., Torkamani, A., Miotto, R., Rossman, H., Berk, B., Baca-Motes, K., Kheterpal, V., Segal, E., Topol, E. J., Ramos, E., and Quer, G. (2025). Multimodal AI Correlates of Glucose Spikes in People with Normal Glucose Regulation, Pre-diabetes and Type 2 Diabetes. *Nature Medicine*. Vol. 31, pp. 3121–3127.

URL: https://doi.org/10.1038/s41591-025-03849-7

Abstract Summary: Type 2 diabetes ($T2D$) is a complex disease monitored poorly by episodic assays like $HbA1c$. This prospective cohort study analyzed multimodal data (including CGM, genetics, and microbiome) from $1,137$ participants (347 deeply phenotyped) across normoglycemic, prediabetic, and $T2D$ states, finding significant differences in glucose spike metrics and demonstrating that a multimodal approach improves $T2D$ risk stratification beyond $HbA1c$ alone.

Key Methodology: Prospective, site-less clinical trial (PROGRESS cohort) collecting multimodal data (CGM, EHR, Fitbit, food logging, $HbA1c$, genomics, gut microbiome) from 347 deeply phenotyped individuals; Spearman's rank correlation analysis; Multimodal binary classification model (XGBoost) for $T2D$ risk assessment, validated on an independent cohort (HPP).

Research Question: How do multimodal data (diet, genetics, exercise, sleep, gut microbiome) correlate with and determine abnormal glucose spikes across different diabetes states (normoglycemia, prediabetes, $T2D$), and can this data be leveraged to define multimodal glycemic risk profiles that improve $T2D$ prevention, diagnosis, and treatment?

Summary

Is It Credible?

Carletti et al. present a prospective, decentralized clinical trial designed to evaluate whether multimodal artificial intelligence can characterize glucose homeostasis better than traditional episodic assays. By collecting data from continuous glucose monitors (CGM), wearable activity trackers, food logs, and biological samples (microbiome, genomics) from 347 individuals, the authors aim to capture the "full complexity" of Type 2 Diabetes (T2D) (p. 3121). The study's central proposition is that this "multimodal glycemic risk profile" offers a more granular and informative assessment of an individual's metabolic health than HbA1c alone, particularly for stratifying risk among prediabetic individuals. The authors report a high degree of accuracy, with their model achieving an area under the curve (AUC) of 0.96 in the primary cohort and 0.90 in an external validation cohort (p. 3123).

The study is methodologically ambitious and succeeds in demonstrating the feasibility of decentralized, sensor-based clinical trials. The ability to recruit a diverse cohort and collect complex physiological data remotely is a significant contribution to the literature on digital health. However, the credibility of the specific biological and clinical claims is tempered by several structural limitations. The most pressing issue is the misalignment between the study's cross-sectional design and its prognostic framing. The authors frequently describe their model as a tool for "prediction of T2D" and for identifying individuals "at risk of progressing into the pathological stage" (p. 3123). Yet, the model is trained on a snapshot of data to classify *current* disease status, not future progression. The "risk profile" generated is effectively a similarity score to the T2D phenotype found in the training set. While this is a valuable metric, it does not inherently validate the prediction of future disease onset, a distinction that requires longitudinal follow-up.

Furthermore, the biological specificity of the model is challenged by the composition of the T2D cohort. A majority of these participants (64 out of 94) were taking antihyperglycemic medications (p. 3123). Consequently, the machine learning model may be detecting the physiological signature of medication use—or a "medicated T2D" phenotype—rather than the unperturbed pathophysiology of the disease. The authors acknowledge that medication use "might potentially result in underestimated differences" (p. 3126) and attempt a sub-analysis on unmedicated individuals, but the small sample size limits the statistical power of this check. This confounding factor complicates the interpretation of the feature importance analysis, as the model's reliance on specific glucose spike metrics may be influenced by drug mechanisms rather than natural disease progression.

The "multimodal" nature of the model also warrants scrutiny regarding the contribution of its components. While the narrative emphasizes the integration of diverse data streams including microbiome and genomics, the supplementary analysis reveals that the predictive performance is driven almost entirely by the wearable sensor data (CGM and Fitbit). The addition of microbiome, genomics, food intake, or electronic health record variables did not yield statistically significant improvements in the binary classification task (Supplementary Information, p. 1). This suggests that while the study validates the power of continuous physiological monitoring, the necessity of the more expensive and complex 'omics' data for this specific classification task is not established by the results.

Finally, the reliability of the lifestyle correlations is undermined by data quality issues inherent to self-reporting. The authors transparently admit that accurate reporting of food intake was "challenging" for participants (p. 3126). This likely explains the counter-intuitive finding that higher carbohydrate intake was associated with *faster* spike resolution (p. 3123), a correlation that contradicts established physiological understanding and suggests significant noise or bias in the dietary logs. Despite these limitations, the external validation in the Human Phenotype Project (HPP) co-

hort remains a strong point. Achieving an AUC of 0.90 despite systematic differences in CGM devices (Dexcom vs. FreeStyle Libre) and sampling rates suggests that the core signal captured by the model—likely driven by glycemic variability—is robust and generalizable (pp. 3128–3129).

The Bottom Line

Carletti et al. convincingly demonstrate that wearable sensors (CGM and activity trackers) can classify diabetes status with high accuracy, outperforming standard demographic models. However, the claim that this approach predicts *future* progression is not supported by the cross-sectional design, and the biological "risk profiles" are likely confounded by the high prevalence of medication use in the diabetic cohort. The study is a strong proof-of-concept for remote digital monitoring, but the added value of integrating expensive microbiome and genomic data for this specific diagnostic purpose appears negligible.

Potential Issues

Model misalignment in risk assessment: The study develops a machine learning model to classify individuals as either normoglycemic or having Type 2 Diabetes (T2D) based on cross-sectional data. However, the article consistently frames this diagnostic classification tool as a prognostic instrument for risk assessment. For instance, the authors claim the model improves "the identification of prediabetic individuals at risk of progressing into the pathological stage of the disease" and assesses "an individual's potential progression to T2D" (p. 3123). This language implies a predictive capacity for future events, but the study's cross-sectional design, which captures a single snapshot in time, cannot validate such claims. The "risk profile" generated by the model is a measure of an individual's similarity to the current T2D phenotype observed in the training data, not a validated predictor of future disease onset. While the authors acknowledge in the discussion that the method will only "potentially serve as a foundation for future longitudinal studies," the prognostic framing in the abstract and results may overstate the model's demonstrated capabilities (p. 3126).

Confounding from antihyperglycemic medication: The machine learning model was trained to distinguish between normoglycemic individuals and those with T2D, yet a substantial portion of the T2D cohort (64 out of 94 participants) were taking antihyperglycemic medications (p. 3123). These medications are designed to alter glucose metabolism, meaning the model may be learning to identify a "medicated T2D" phenotype rather than the unperturbed biological signature of the disease. This introduces a significant confound that could affect the model's feature importance and its applicability to unmedicated individuals. The authors attempt to address this by performing a sub-analysis comparing medicated and unmedicated T2D individuals (n=30), finding no statistically significant differences in glucose metrics. However, this comparison is likely underpowered due to the small sample size of the unmedi-

cated group. The authors acknowledge this limitation in the discussion, noting that medication use "might potentially result in underestimated differences in glucose spike metrics between diabetics and non-diabetics," but this potential confound remains a central challenge to the interpretation of the model's findings (p. 3126).

Systematic mismatches in external validation data: The study's claim of external validation in the Human Phenotype Project (HPP) cohort is potentially weakened by systematic differences in data collection methodologies compared to the primary PROGRESS cohort. First, the core continuous glucose monitoring (CGM) data was collected with different devices: the PROGRESS cohort used Dexcom G6 monitors (5-minute sampling), while the HPP cohort used FreeStyle Libre Pro devices (15-minute sampling) (pp. 3128–3129). This difference in temporal resolution could affect the ability to detect the rapid glucose excursions that define the study's "spike" metrics. Second, the benchmark comparator, HbA1c, was measured with lower precision in the validation cohort. In the PROGRESS cohort, HbA1c was measured from a contemporaneous blood sample, whereas in the HPP cohort, it was extracted from electronic health records and could be up to 90 days removed from the CGM monitoring period (pp. 3128–3129). The authors are transparent about these limitations, noting that different CGM devices "might lead to biases" (p. 3126) and that the HbA1c time mismatch "can potentially cause inaccuracies in the analysis" (p. 3129). Nonetheless, these domain shifts introduce non-random measurement differences that may compromise the robustness of the validation.

High risk of selection bias and limited generalizability: The study's final analysis is based on 347 participants from an initial enrollment of 1,137, representing a 69% attrition rate (p. 3122; Extended Data Fig. 1). The primary reason for exclusion was insufficient CGM data, indicating that a large portion of the initial cohort did not adhere to the demanding monitoring protocol. The authors checked for selection bias by comparing the included and excluded groups on age and sex and found no significant differences (p. 3122). However, this check did not extend to other potentially

important variables like BMI, socioeconomic status, or digital literacy. It is plausible that participants who successfully completed the protocol are systematically different from those who did not, potentially being more health-conscious, motivated, or technologically adept. This "compliant user bias" may limit the generalizability of the findings to the broader population.

Reliance on self-reported data of acknowledged low quality: The analysis of lifestyle factors relies on self-reported data, particularly for food intake, which the authors concede was of questionable accuracy. They state that "accurate reporting of food intake for participants in real-world conditions... proved challenging (in adherence and accuracy) for many" (p. 3126). Despite this significant data quality issue, dietary variables are included in the correlation analyses and contribute to a key finding: a statistically significant negative correlation between carbohydrate intake and spike resolution (p. 3123). This suggests, counter-intuitively, that higher carbohydrate intake is associated with faster glucose absorption. Reporting a key finding based on data acknowledged to be unreliable calls into question the validity of conclusions related to diet and potentially other self-reported lifestyle factors.

Marginal contribution of most 'omics' data streams to model performance: The article's central narrative emphasizes the superiority of a "multimodal" approach that integrates diverse data streams. However, a supplementary analysis of the model's performance suggests that much of the predictive power is derived from wearable sensor data, with limited contribution from other modalities. The analysis shows that while CGM and Fitbit data provided statistically significant improvements over a base model of demographic variables, "there were no statistically significant improvements observed when adding microbiome variables... genomics variables... food intake variables... or EHR variables" individually (Supplementary Information, p. 1). This finding suggests that the high performance of the final model is driven primarily by continuous physiological monitoring, which may temper the broader claims about the necessity of integrating complex and costly 'omics' data for this

specific classification task.

Omission of socioeconomic confounders in correlation analysis: The study investigates the associations between glucose spike metrics and various lifestyle and biological factors while controlling for age, sex, and polygenic risk score (p. 3123). However, the analysis does not adjust for potential confounding by socioeconomic status (SES), such as education level or income. SES is known to be strongly associated with many of the variables under investigation, including diet, physical activity, and gut microbiome composition. The authors collected data on education and rurality as part of their UBR definition but did not include these as covariates in the correlation models (p. 3122). The omission of SES controls means that some of the reported associations, particularly those related to lifestyle factors, could be influenced by unmeasured socioeconomic differences among participants.

Presentation and transparency issues: Several aspects of the study's presentation could be clarified. First, the claim of a "diverse cohort with 48.1% of participants self-identified as UBR" (p. 3125) relies on a broad definition of "Underrepresented in Biomedical Research" that includes age over 65 and rural residence. While the authors are transparent in Table 1, this framing may obscure the low representation of key racial and ethnic minority groups, such as Black (3.2%) and Hispanic (3.5%) participants (p. 3123, Table 1). Second, the article refers to the HPP cohort as "independent" (p. 3123), which is true in the sense of being a separately collected dataset from a different population. However, the significant overlap in authorship and institutional affiliations between the two projects is a nuance worth considering when assessing the rigor of the validation, a fact the authors disclose in the competing interests section (p. 3131). Finally, a minor clerical error exists between the main text, which states 412 participants shared CGM data, and Extended Data Fig. 1, which indicates this number was 406 (p. 3122; Extended Data Fig. 1).

Future Research

Longitudinal validation of risk scores: To substantiate the claim that the multimodal model predicts disease progression, future research must move beyond cross-sectional classification. A longitudinal study following drug-naive prediabetic individuals over several years is required to determine if the "glycemic risk profile" at baseline actually correlates with the time-to-onset of T2D. This would transform the model from a sophisticated diagnostic tool into a genuine prognostic instrument.

Deconfounding medication effects: Future work should prioritize training classification models exclusively on unmedicated T2D populations or individuals with new-onset diabetes prior to treatment initiation. This would ensure that the learned features represent the underlying pathology of the disease rather than the pharmacodynamics of antihyperglycemic agents. If recruiting a large unmedicated cohort is not feasible, future analyses could employ causal inference methods to adjust for the treatment effect, provided the sample size is sufficient.

Cost-benefit analysis of data modalities: Given that the supplementary results indicated no significant performance gain from microbiome, genomics, or food logs, future studies should rigorously evaluate the cost-effectiveness of the multimodal approach. Research should aim to identify the "minimum viable dataset"—likely a combination of CGM and actigraphy—that achieves comparable accuracy to the full multimodal suite. This would have significant implications for the scalability and clinical implementation of such screening tools.

isitcredible.com