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 11, 2026

v2

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?

Editor's Note

Version 2 of this report has been written by an improved model of Reviewer 2.

Summary

Is It Credible?

Carletti et al. present a "multimodal AI" approach to characterize glucose spikes and stratify Type 2 Diabetes (T2D) risk. By collecting data from 347 participants using continuous glucose monitors (CGM), activity trackers, and self-collected biosamples (microbiome, genomics), the authors trained a machine learning model to distinguish individuals with T2D from those with normoglycemia. They claim this model generates "multimodal glycemic risk profiles" that are "more informative than HbA1c" and can identify prediabetic individuals "at risk of progressing" to disease (p. 3123). While the article demonstrates the logistical feasibility of decentralized, remote phenotyping, the predictive claims attached to the AI model appear overstated given the study's design and results.

The core claim that the model provides a prognostic "risk of progressing" is not fully supported by the evidence. The model was trained on cross-sectional data to classify *current* disease status—distinguishing existing T2D patients from normoglycemic controls. When the authors apply this classifier to prediabetic individuals, they interpret the resulting probability score as a measure of future risk. While it is a reasonable hypothesis that prediabetic individuals who phenotypically resemble T2D patients are more likely to progress, this remains a hypothesis. The study lacks the longitudinal follow-up required to validate this prognostic capability. Consequently, the "glycemic risk profile" is effectively a similarity score to the current T2D phenotype, not a validated predictor of future disease onset.

Furthermore, the "multimodal" aspect of the model—specifically the integration of complex and costly 'omics' data—appears to offer diminishing returns relative to the study's framing. The authors emphasize the combination of microbiome, genomics, and food intake data. However, supplementary analyses reveal that adding micro-

biome, genomics, or food data individually to base demographic variables yielded "no statistically significant improvements" in model performance (Supplementary Results, p. 24). While the aggregate model using *all* data streams did achieve the highest performance (AUC 0.97), representing a statistically significant improvement over the base model (AUC 0.85), the primary performance boost came from glucose spike metrics derived from CGM and, to a lesser extent, wearable activity data. This suggests that while the full multimodal approach is technically superior, the significant cost and burden of collecting stool and saliva samples may not be justified by the marginal predictive gain over a "CGM-enhanced" model.

The robustness of the findings is further challenged by potential confounding and data quality issues. A significant majority of the T2D training group (64 of 94) were taking antihyperglycemic medications (p. 3123). As a result, the model may be learning to detect the physiological signatures of medication use rather than the unadulterated biological signature of T2D. Additionally, the study suffered from high attrition; only 347 of the 1,137 enrolled participants provided sufficient data for the final analysis (p. 3122). This drop-off suggests that the protocol may be too burdensome for a general population, introducing potential selection bias toward highly compliant individuals. The reliance on self-reported diet data, which the authors acknowledge was challenging for participants to report accurately (p. 3126), also weakens the reliability of the reported correlations between lifestyle factors and glucose spikes, though some expected physiological correlations (e.g., carbohydrate intake and spike resolution) were still observed.

Finally, while the study includes an external validation in the Human Phenotype Project (HPP), systematic differences between the cohorts complicate the comparison. The validation cohort used different CGM devices with lower temporal resolution (15-minute vs. 5-minute sampling) and relied on HbA1c data that could be temporally mismatched by up to 90 days within a 180-day window (p. 3129). While the authors attempted to harmonize these datasets, such domain shifts introduce

noise that makes it difficult to confirm whether the "risk profiles" are truly generalizing or simply capturing broad glycemic instability. Ultimately, the study succeeds in quantifying glucose spike metrics across diabetes states, but the utility of the AI model as a superior prognostic tool remains unproven.

The Bottom Line

Carletti et al. successfully demonstrate the feasibility of collecting deep phenotypic data remotely, but the claim that their AI model identifies individuals "at risk of progressing" to diabetes is a hypothesis rather than a proven conclusion. The model is built on cross-sectional data and essentially measures similarity to current diabetes patients, not future risk. Furthermore, while the aggregate "multimodal" model achieved the best statistical performance, the expensive genomic and microbiome data added marginal predictive value over standard continuous glucose monitoring, suggesting the complexity of the full approach may not be cost-effective for this specific classification task.

Potential Issues

Model framing and cross-sectional design: The study develops a machine learning model to classify individuals based on cross-sectional data, but the article's framing may imply a prognostic capability that the design cannot validate. The model is presented as a tool for creating "multimodal glycemic risk profiles" that could improve "the identification of prediabetic individuals at risk of progressing" and assess "an individual's potential progression to T2D" (p. 3123). This language suggests a capacity to predict future events. However, a cross-sectional design can only capture an individual's similarity to the current T2D phenotype observed in the training data, not a validated measure of future disease onset. The authors are transparent about this limitation in the discussion, clarifying that the study "represents the initial step" and that the method will only "potentially serve as a foundation for future longitudinal studies" for validation (p. 3126). Nonetheless, the prognostic framing used in the results section could be interpreted as overstating the model's demonstrated capabilities.

Potential confounding from antihyperglycemic medication: The machine learning model was trained to distinguish between normoglycemic individuals and those with Type 2 Diabetes (T2D), yet a substantial portion of the T2D cohort (64 out of 94 participants) were taking antihyperglycemic medications (p. 3123). As these medications are designed to alter glucose metabolism, the model may be learning to identify a "medicated T2D" phenotype rather than the unperturbed biological signature of the disease. The authors attempt to address this by performing a sub-analysis comparing glucose spike metrics between medicated (n=64) and unmedicated (n=30) T2D individuals, finding no statistically significant differences in these specific metrics (p. 3123). While this finding supports their decision to pool the groups, the comparison is likely underpowered due to the small sample size of the unmedicated group. The authors acknowledge this limitation, noting in the discussion that med-

ication use "might potentially result in underestimated differences in glucose spike metrics between diabetics and non-diabetics" (p. 3126), leaving this as a potential challenge to the interpretation of the model's findings.

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. First, the core continuous glucose monitoring (CGM) data was collected with different devices: the primary PROGRESS cohort used Dexcom G6 monitors (5-minute sampling, p. 3128), while the HPP cohort used FreeStyle Libre Pro devices (15-minute raw sampling, p. 3129). Although the authors harmonized the data by interpolating the HPP data to 1-minute epochs (p. 3129), the initial difference in temporal resolution could affect the detection of rapid glucose excursions. 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 (within a 180-day window). 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" (p. 3129). These domain shifts introduce non-random measurement differences that may compromise the robustness of the validation.

Marginal contribution of most 'omics' data streams to model performance: The article's central narrative emphasizes the superiority of a "multimodal" approach, but a supplementary analysis suggests that much of the model's predictive power is derived from wearable sensor data. 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" when each was added individually (Supplementary Results, p. 24). While the final model incorporating *all* data streams did achieve the highest performance (AUC 0.97), which was a statistically significant improvement over the base model (AUC 0.85), the limited marginal contribution of individual 'omics' modalities tempers the broader claims about the necessity of integrating these complex and costly data streams for this specific classification task.

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. While 412 participants shared CGM data, the reduction to 347 was due to strict data completeness requirements for the multimodal analysis (p. 3122, Extended Data Fig. 1). This significant drop-off indicates 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 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. For instance, the authors report a statistically significant negative correlation between carbohydrate intake and spike resolution time (p. 3123). While this interpretation is physiologically plausible and suggests some signal in the data, reporting findings based

on data acknowledged to be unreliable calls for caution regarding the validity of conclusions related to diet.

Omission of socioeconomic confounders in correlation analysis: The study investigates 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) in the statistical models. SES is known to be associated with many of the variables under investigation, including diet, physical activity, and gut microbiome composition. While the authors collected data on education and rurality as part of their definition for Underrepresented in Biomedical Research (UBR) participants (p. 3122) and provisioned devices to mitigate access bias, they did not include SES as a covariate in the correlation models. This omission means that some of the reported associations, particularly those related to lifestyle, could be influenced by unmeasured socioeconomic differences.

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. 3122) relies on a broad definition of "Underrepresented in Biomedical Research" that includes age over 65 and rural residence. While the authors are transparent about this 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). Second, the article refers to the HPP cohort as "independent" (p. 3123). While the cohorts are indeed distinct populations (US vs. Israel) and datasets, the significant overlap in authorship and institutional affiliations between the two projects, which the authors disclose in the competing interests section (p. 3125), is a nuance worth considering when assessing the rigor of the validation.

Future Research

Longitudinal validation of risk scores: Future work must validate the prognostic utility of the glycemic risk profiles by following prediabetic individuals over time. A longitudinal design is required to determine whether individuals with high "risk profile" scores actually convert to Type 2 Diabetes at higher rates than those with lower scores, thereby justifying the claim of improved risk stratification.

Evaluation of medication-naive cohorts: To ensure the model is detecting the biological signature of diabetes rather than the effects of treatment, future studies should train and test the model on medication-naive T2D populations. This would isolate the disease phenotype from the confounding influence of antihyperglycemic drugs, which were prevalent in the current study's T2D cohort.

Cost-benefit analysis of data modalities: Given the marginal performance gains observed from adding microbiome and genomic data individually, future research should rigorously evaluate the cost-effectiveness of these modalities. Studies should determine whether the improvement in classification accuracy justifies the significant financial cost and participant burden associated with collecting and processing these complex biological samples compared to using CGM and demographic data alone.

isitcredible.com