nature portfolio

Corresponding author(s): Edward Ramos, Giorgio Quer

Last updated by author(s): May 1, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	A study-specific smartphone app (MyDataHelps, CareEvolution) was developed to facilitate participants in their self-guided completion of the required tasks.
Data analysis	Software utilized for genetic analysis include: PLINK v1.9; ADMIXTURE v1.3.0. Software utilized for microbiome analysis include: fastp v0.23.4; Bowtie 2 v2.5.2; FLASH v1.2.11; MetaPh1An v4.0.6. Python libraries utilized in the analysis include: Python v3.11.5 for code development; NumPy v1.26.0, pandas v2.1.1 and SciPy v1.11.2 for data manipulation and processing; pingouin v0.5.3 for statistical tests; XGBoost v2.0.3 for machine learning models training; seaborn v0.13.2 and Matplotlib v3.8.0 for visualization.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Participant data with known identifiers removed are available for academic research purposes upon reasonable request and within established restrictions, i.e., not for commercial purposes or for the benefit of for-profit entities. Access to the PROGRESS data is subject to approval by the study sponsor (Tempus AI) and the study principal investigator to ensure alignment with the original study objectives and ethical guidelines. Approved researchers will be granted secure access through a designated data environment; direct downloads or transfer of individual-level raw data are not permitted. Requests for data access should include a summary of the proposed research objectives, analytic plan, and institutional affiliation. For further information or to submit a data access request, please contact the corresponding authors. Responses will be processed within 6 weeks.
Reference data from the HPP study is available at https://github.com/ayya-keshet/CGMap. Access to this dataset, for universities and other research institutions, can be requested at https://humanphenotypeproject.org/.
The 1000 Genomes Project dataset was sourced from https://www.internationalgenome.org/. The GRCh38 human reference genome dataset was sourced from the UCSC Genome Browser (https://genome.ucsc.edu).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	For the PROGRESS study, sex and gender identity information was self-reported. We conducted sex-based analyses.
Reporting on race, ethnicity, or other socially relevant groupings	For the PROGRESS study, ethnicity and education level information was self-reported. The determination of whether an individual resides in a rural or non-rural area was based on the ZIP code provided by the participant.
Population characteristics	For the PROGRESS study, we included 347 individuals in the analysis, of whom 174 (50.1%) were normoglycemic, 79 (22.8%) were prediabetic, and 94 (27.1%) had type 2 diabetes, based on self-reported data. The average age of the selected population was 49.1 ± 15.6 years (mean ± standard deviation), and 66.9% were females.
Recruitment	Study recruitment was initiated in February 2021 via social media outreach and partnerships with healthcare provider systems. Paid advertisements on the website Nextdoor in the greater San Diego area drove interested individuals to a study web page (https://progress.scripps.edu/) where they indicated a desire to join the study as well as a brief capture of their diabetes status. Over one thousand interested individuals from that campaign were then invited to screen for the study beginning in July 2021. Additional social media channels such as Facebook campaigns were used to generate interest nationally. In addition, outreach to patients from our healthcare partners was done via identification of eligible participants based on criteria (e.g., diabetes status) via EHR. Individuals who appeared to meet the inclusion and exclusion criteria were sent an electronic notification through the EHR with an invitation to participate in the study. All recruitment campaigns directed interested individuals to the study website with an overview of the study and a call to action to either join the waiting list or enroll to screen. Lastly, an email-based campaign leveraging a third-party organization, Research Match, was utilized where notifications were sent to participants who had registered in the Research Match database and listed having prediabetes or T2D. By focusing on digital campaigns and emphasizing that participants did not need to attend any in person site visits to fully participate, we were able to recruit a diverse patient population both in age, socioeconomic status, and geographic location. Technologies are a key component of remote, real-world data collection in decentralized clinical trials, though they may cause unintentional selection biases due to difficulties for certain groups of individuals to access many of these technologies, resulting in potential discrimination against the low-income socioeconomic class or the elderly. In this study, we attempted to mitigate this bias through study-provisioned devices to all participants.
Ethics oversight	The protocol for the PROGRESS study was reviewed and approved by the Scripps Office for the Protection of Research Subjects (IRB-20-7635).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For the PROGRESS study, a 1000-person sample size was initially chosen as a convenience sample to adequately demonstrate the scalability of the direct-to-participant experience, and account for known confounders. Between November 3, 2021, and September 1, 2023, PROGRESS enrolled 1,137 individuals. After excluding the participants whose data was insufficient (see 'Data exclusions'), a total of 347 PROGRESS participants were included in the analysis.
Data exclusions	We excluded: individuals who did not meet the minimum continuous glucose monitor wearing time (at least 16 hours/day and at least 5 days); individuals who did not report their age or sex; individuals whose glycated haemoglobin (HbA1c) test was not available; individuals with type 1 diabetes (based on electronic health records data); individuals taking insulin at the time of enrollment (based on electronic health records data).
Replication	All the proposed metrics in our study are deterministic, ensuring that the results are fully reproducible. The entire processing pipeline has been thoroughly tested, consistently yielding the same results from the raw data to the final outcomes.
Randomization	For the training and testing of machine learning models, participants were deterministically assigned to training or testing cohorts. We took measures to ensure that individuals in the testing cohort were not present in the training cohort, thus maintaining the independence of the samples. The detailed procedure for this allocation process can be found in the manuscript.
Blinding	Blinding was not relevant to the study as we did not test the effect of any treatment.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	ClinicalTrials.gov registration number: NCT04881019.
Study protocol	Access to the study protocol should be requested to the corresponding authors.
Data collection	For the PROGRESS study, data was collected from November 2021 to September 2023. The data collection was performed remotely through study-provided wearable devices (ncluding a continuous glucose monitor and a fitness tracker), study-provided kits for bio samples self-collection, surveys, and mobile apps.
Outcomes	The main outcome of interest was to distinguish differences in glycemic response characteristics in participants with and without type 2 diabetes. This was achieved by computing glucose spike metrics from the collected continuous glucose monitoring data and assessing the differences among different diabetes states (normoglycemia, prediabetes, and type 2 diabetes).

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*