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
☐ 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
☐ 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: We utilized de-identified electronic medical record data from Mount Sinai Health System approved by IRB.

Data analysis: We performed our data analysis using Python and R, and their open-source libraries.

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

The clinical data here were used under license from Mount Sinai Data Warehouse in the current study. As a result, this dataset is not publicly available. Qualified researchers affiliated with the Mount Sinai Health Systems may apply for access to these data through the Mount Sinai Health Systems Institutional Review Board.
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
We developed a pregnancy journey construction algorithm to identify 83,954 patients with 114,312 pregnancies represented in the MSHS EMR systems and to reconstruct 108,557 full pregnancy journeys of 80,021 patients between 2002 and 2019.

Data exclusions
Describe any data exclusions. If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.

Replication
Exact replication is not possible due to the fact that we cannot share the data publicly.

Randomization
Describe how samples/organisms/participants were separated into experimental groups. If allocation was not random, describe how covariates were controlled OR if this is not relevant to your study, explain why.

Blinding
Describe whether the investigators were blinded to group allocation during data collection and/or analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.

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 | Methods |
|----------------------------------|---------|
| n/a | Involved in the study |
| ☒ | Antibodies |
| ☒ | Eukaryotic cell lines |
| ☒ | Palaeontology and archaeology |
| ☒ | Animals and other organisms |
| ☒ | Human research participants |
| ✗ | Clinical data |
| ✗ | Dual use research of concern |
| ☒ | Involved in the study |
| ✗ | ChIP-seq |
| ☒ | Flow cytometry |
| ☒ | MRI-based neuroimaging |

Clinical data

Policy information about clinical studies: guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration
Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.

Study protocol
IRB-17-01245

Data collection
We utilized de-identified electronic medical record from Mount Sinai Health System. We identified 114,312 deliveries with 83,954 unique patients, and reconstructed 108,557 full pregnancy journeys of 80,021 patients between 2002 and 2019.

Outcomes
We implemented a diagnosis phenotyping algorithm to predict probability of patient who will develop Preeclampsia in pregnancy journey.