nature research

Corresponding author(s): Kang Zhang

# Reporting Summary

Nature Research 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 Research policies, see Authors & Referees and the Editorial Policy Checklist.

## Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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 |
| ☐ | ☒ An indication of 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<br>*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 statistics 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<br>*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 |
| ☒ | ☐ Clearly defined error bars<br>*State explicitly what error bars represent (e.g. SD, SE, CI)* |

*Our web collection on statistics for biologists may be useful.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | The electronic health records used in this study was developed by a Chinese vendor named Zesing Electronic Medical Records. |
| Data analysis | Mecab (URL: https://github.com/taku910/mecab) was used for tokenization. word2vec from the python Tensorflow package was used to embed the 4363 tokens with 100 features, to represent the semantics and similarities of the word in the high dimensional space. Python seaborn package was used to generate hierarchical clustering. |

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/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

We have made available the Jupyter notebook that we used in constructing and validating the hierarchical logistic regression models: https://s3.cn-

north-1.amazonaws.com.cn/ped.emr/Data/hierachical_logistic_regression.ipynb. To protect patient confidentiality, we have deposited de-identified aggregated patient data in a secured and patient confidentiality compliant cloud in China in concordance with data security regulations. Data access can be requested by writing to the corresponding authors. All data access requests will be reviewed and (if successful) granted by the Data Access Committee

# Field-specific reporting

Please select 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/authors/policies/ReportingSummary-flat.pdf

# Life sciences study design

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

| | |
|---|---|
| Sample size | Electronic health records were collected from 1,362,559 outpatient patient visits from the Guangzhou Women and Children's Medical Center. (See Method). It has been well documented that machine learning techniques improve with a greater amount of input data, so the large volume of encounters here contributed to the robustness of the diagnostic system (See Discussion). |
| Data exclusions | We did not exclude any data |
| Replication | No experimental replication was attempted. |
| Randomization | The data was split into a training cohort, consisting of 70% of the total visit records, and a testing cohort, comprised of the remaining 30%. (See Method) |
| Blinding | Blinding is not applicable since only electronic health records, in which patient sensitive information was removed during the initial extraction of EHR data and EHR were de-identified, were used to evaluate performance of AI system v.s. human physicians. |

# Reporting for specific materials, systems and methods

## Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ Unique biological materials |
| ☒ | ☐ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology |
| ☒ | ☐ Animals and other organisms |
| ☐ | ☒ Human research participants |

## Methods

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

## Human research participants

Policy information about studies involving human research participants

| | |
|---|---|
| Population characteristics | Electronic health records of 1,362,559 outpatient visits from 567,498 pediatric patients from the Guangzhou Women and Children's Medical Center were collected. These records encompassed physician encounters for pediatric patients presenting to this institution from January 2016 to July 2017. The median age was 2.35 years (range: 0 to 18, 95% confidence interval: 0.2 to 9.7 years old), and 40.11% were female. 11,926 patient visit records from an independent cohort of pediatric patients from Zhengcheng Women and Children's Hospital (Guangdong Province, China) were used for a comparison study between our AI system and human physicians. |
| Recruitment | Electronic health records of 1,362,559 outpatient visits from 567,498 pediatric patients and 11,926 patient visit records were collected from Guangzhou Women and Children's Medical Center and Zhengcheng Women and Children's Hospital. The study was approved by the Guangzhou Women and Children's Medical Center and Zhengcheng Women and Children's Hospital institutional review board and ethics committee and complied with the Declaration of Helsinki. Consents were obtained from all participants at the initial hospital visit. Patient sensitive information was removed during the initial extraction of EHR data and EHR were de-identified. A data use agreement was composed and upheld by all institutions involved in the data collection and analysis. Data were stored in a fully HIPAA-compliant manner. |