nature research

Corresponding author(s): Joseph R Ledsam

Last updated by author(s): Oct 26, 2020

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.

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	Data collection was performed by independent members of the VA National Data Center without involvement from research team members. Collection was performed using the Vista EHR system and associated databases.
Data analysis	The networks used the TensorFlow library with custom extensions. Analysis was performed with custom code written in Python 2.7. Please see the manuscript methods section for more detail.

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. 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

The clinical data used for the training, validation and test sets were collected at the US Department of Veterans Affairs and transferred to a secure data centre with strict access controls in de-identified format. Data were used with both local and national permissions. They are not publicly available and restrictions apply to their use. The de-identified dataset, or a test subset, may be available from the US Department of Veterans Affairs subject to local and national ethical approvals.

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	The dataset consisted of all eligible patients during a five year period across the entire VA healthcare system in the USA. The test population was a random selection of 10% of these, totaling 70,681 individual patients and 252,492 unique admissions (please refer to methods section for more details on how test populations were selected). A sample size requirement of 179 patients would be required to detect sensitivity and specificity at 0.05 marginal error and 95% confidence. The total number of test patients exceeded this requirement by two orders of magnitude.
Data exclusions	We excluded patients below the age of 18 and above the age of 90 in accordance with HIPAA Safe Harbor criteria, and patients without any serum creatinine recorded in EHR. (See paper methods for more detail.) To protect patient privacy sites with fewer than 250 admissions during the five year time period were also excluded; four of the 1,243 health care facilities from which the VA is composed were excluded based on this criteria. All exclusion criteria were established prior to beginning the work.
Replication	All 70,681 patients in the test set were randomly selected and were not correlated in any way. The experiments can be interpreted as 70,681 replicas of the model applied to a single patient over a fifteen year period.
Randomization	The data were randomly divided into training (80% of observations), validation (5%), calibration (5%) and testing (10%) sets. All data for a single patient was assigned to exactly one of these splits. (See paper methods for more detail.)
Blinding	When assigning patients randomly to test, validation and training groups investigators were blinded to patient covariates and all features in the EHR not required to perform the research (e.g., creatinine was required to label AKI as a ground truth). Patient recruitment was conducted by independent members of the VA National Data Center; research team members were blinded to this recruitment.

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
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

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	The data included all VA patients aged between 18 and 90 admitted for secondary care to medical or surgical services between 10/1/2011 to 9/30/2015, with laboratory data that included serum creatinine recorded in EHR and with at least one year of EHR data prior to admission data. The test set was a randomly selected 10% of all admissions included in the work. Average age was 62.3. Males represented 93.6% of the test population. Average number of inpatient admissions was 3.6; average admission duration was 9.6 days. AKI occurred in 13.4% of admissions. These figures were consistent with the population of the VA as a whole.
Recruitment	The data was recruited from the US Department of Veterans Affairs (VA). The VA is composed of 1,243 health care facilities, including 172 VA Medical Centers and 1,062 outpatient sites of care. Aggregating data from one or more of these facilities are 130 data centres, of which 114 had data for inpatient admissions used in this study. Four sites were excluded due to small numbers of patients: fewer than 250 admissions during the fifteen year time period. No other patients were excluded based on location, and no other exclusion criteria were applied. The final dataset consisted of the records for all 703,782 patients that met

inclusion and exclusion criteria.

Ethics oversight

This work, and the collection of data on implied consent, received Tennessee Valley Healthcare System Institutional Review Board (IRB) approval from the US Department of Veterans Affairs. De-identification was performed in line with the Health Insurance Portability and Accountability Act (HIPAA), and validated by the US Department of Veterans Affairs Central Database and Information Governance departments. Only de-identified retrospective data was used for research, without the active involvement of patients.

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