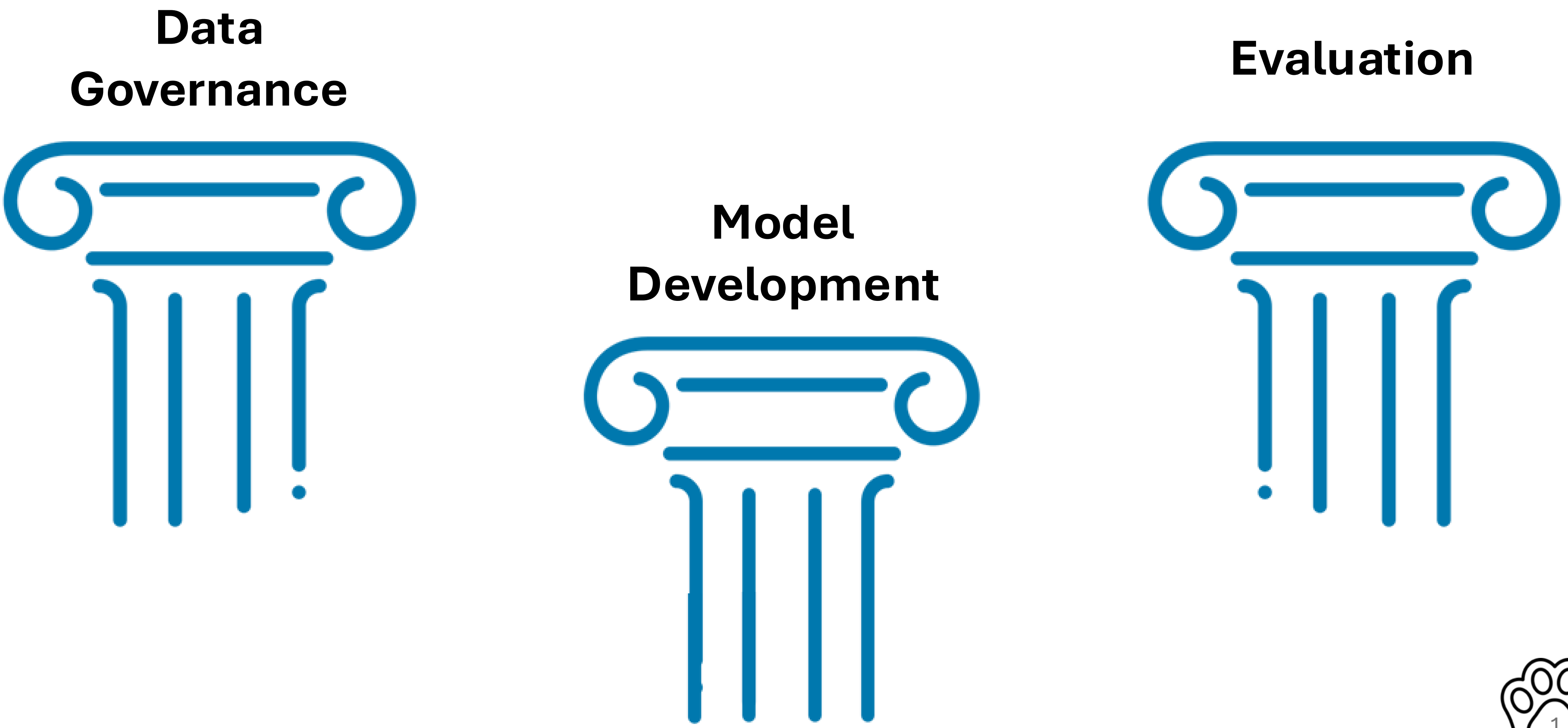
Sean Farrell

Democratising Veterinary EHRs: Balancing Privacy & Open Science for the future of Large Language Model Research in Veterinary Science

BBSRC Durham University University of Liverpool


Data Governance
Model Development
Evaluation

Consult Date	18/05/2025
Breed	Cocker Spaniel
Zip Code	14853
Favourite Colour	Red
Mother's Maiden Name	Basran
Date of Birth	14/01/2025
Dispensed Product(s)	amoxi-clav
Clincial Narrative	OR cookie been v+ since y. spoek w...

Consult Date	<<REDACTED>>
Breed	<<REDACTED>>
Zip Code	<<REDACTED>>
Favourite Colour	<<REDACTED>>
Mother's Maiden Name	<<REDACTED>>
Date of Birth	<<REDACTED>>
Dispensed Product(s)	<<REDACTED>>
Clincial Narrative	**OR cookie been v+ since y. spoek w...**

Consult Date	18/05/2025
Breed	Cocker Spaniel
Zip Code	14853
Favourite Colour	Red
Mother's Maiden Name	Basran
Date of Birth	14/01/2025
Dispensed Product(s)	amoxi-clav
Clincial Narrative	OR cookie been v+ since y. spoek w...

OR cookie been v+ since y. spoke to Williams about cause.

ref to Davies for follow up on 24/04. Going to taken by Claire

as O has parkinsons

Consult Date	18/05/2025
Breed	Cocker Spaniel
Zip Code	14853
Favourite Colour	Red
Mother's Maiden Name	Basran
Date of Birth	14/01/2025
Dispensed Product(s)	amoxi-clav
Clincial Narrative	OR cookie been v+ since y. spoek w...

pet name vet name

OR cookie been v+ since y. spoke to Williams about cause.

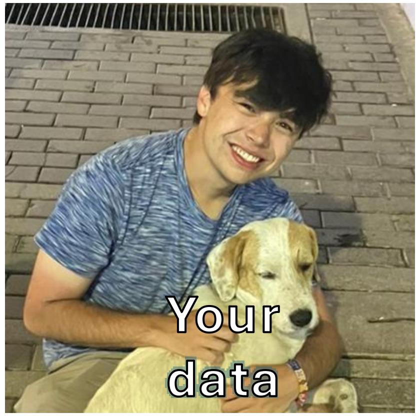
referral name date owner name

ref to Davies for follow up on 24/04. Going to taken by Claire

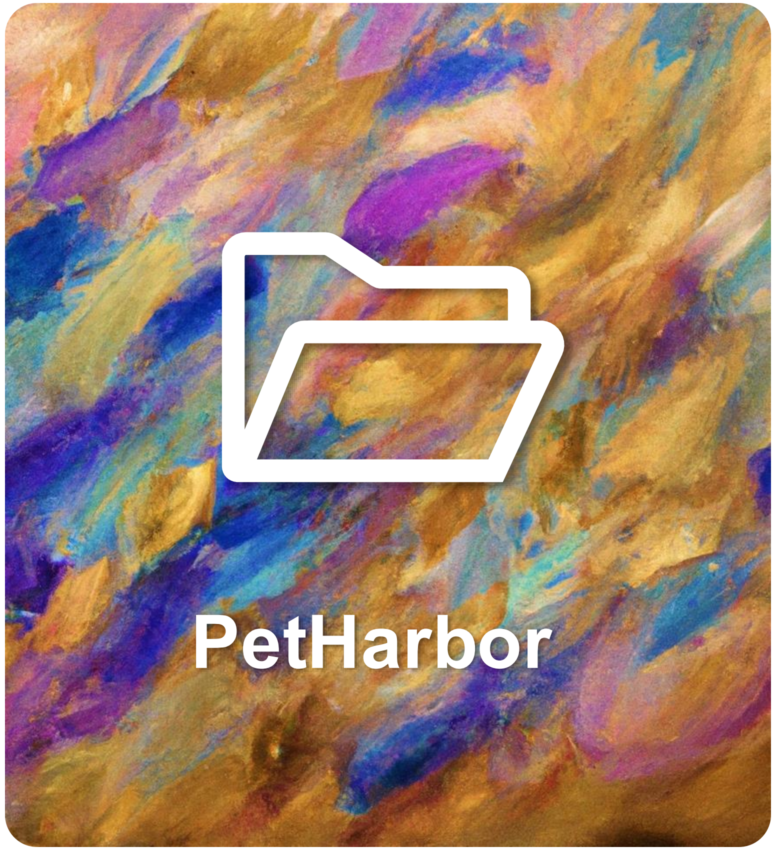
owner personal info

as O has parkinsons

Consult Date	18/05/2025
Breed	Cocker Spaniel
Zip Code	14853
Favourite Colour	Red
Mother's Maiden Name	Basran
Date of Birth	14/01/2025
Dispensed Product(s)	amoxi-clav
Clincial Narrative	OR cookie been v+ since y. spoek w...



- Publish your name and address?
- Give your data to my friend?
- Use it for advertisement?
- Reveal practice specific trade secrets?



Pseudoanonymisation
Automated free-text annonymisation tools

Data Aggregation
Grouping data to minimize unique identifiers

Entity Extraction
NLP Tools

Data Minimisation
Share the minimal viable number of records

Data Access
Multi-reviewer manual reading

Pseudoanonymisation
Automated free-text annonymisation tools

Data Aggregation
Grouping data to minimize unique identifiers

Entity Extraction
NLP Tools

Data Minimisation
Share the minimal viable number of records

Data Access
Multi-reviewer manual reading

HIPAA Safe Harbor	**Examples**
(A) Names	Pet, Owner, Vet Names
(B) Geographic subdivisions	City, Towns, Countries
	Vet practices, hospitals, shelters
(C) Dates	Day/month dates, appointments
(D) Telephone numbers	Client/practice phone numbers
(E) Fax numbers	n/a
(F) Email addresses	Referral/client emails
(G) Social security numbers	n/a
(H) Medical record numbers	n/a
(I) Health plan numbers	Insurance policy numbers
(J) Account numbers	Microchip Numbers
(K) Certificate numbers	MRCVS clinician codes
(L) Vehicle identifiers	n/a
(M) Device identifiers	n/a
(N) URLs	Website urls
(O) IP addresses	n/a
(P) Biometric identifiers	n/a
(Q) Photographic images	n/a
(R) Other identifiers	Passport numbers

Pseudoanonymisation
Automated free-text annonymisation tools

Data Aggregation
Grouping data to minimize unique identifiers

Entity Extraction
NLP Tools

Data Minimisation
Share the minimal viable number of records

Data Access
Multi-reviewer manual reading

Cookie presented with a fever

↓ De-identification

<<pet name>> presented with a fever

↓ Pseudoanonoymisation

Charlie presented with a fever

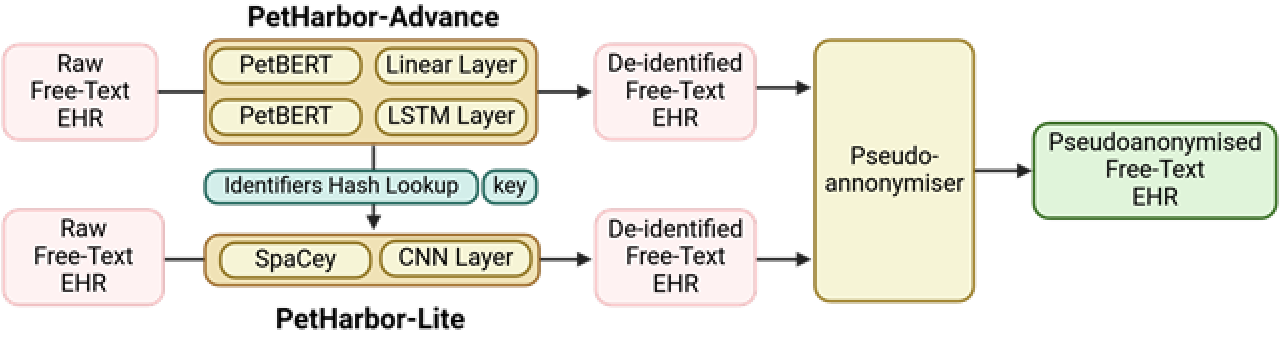
Pseudoanonymisation
Automated free-text annonymisation tools

Data Aggregation
Grouping data to minimize unique identifiers

Entity Extraction
NLP Tools

Data Minimisation
Share the minimal viable number of records

Data Access
Multi-reviewer manual reading



ARCHITECTURE	PRECISION	RECALL	TIME
Advance	0.97	0.96	33s (GPU) * 3m8s (CPU) **
Lite	0.92	0.84	49s (CPU) **

* 1x Nvidia A4000
** 4 core CPU

Pseudoanonymisation
Automated free-text annonymisation tools

Data Aggregation
Grouping data to minimize unique identifiers

Entity Extraction
NLP Tools

Data Minimisation
Share the minimal viable number of records

Data Access
Multi-reviewer manual reading

```python
from petharbor.advance import Anonymiser

# Initialize the anonymizer
petharbor = Anonymiser()

# Anonymize single text
anonymized_text = petharbor.anonymise("Cookie presented
to Jackson's on 25th May 2025 before travel to Hungary.
Issued passport (GB52354324)")

# Output: <<NAME>> presented to <<ORG>> on <<TIME>>
before travel to <<LOCATION>>. Issued passport
(<<MISC>>)
```

ARCHITECTURE	PRECISION	RECALL	TIME
Advance	0.97	0.96	33s (GPU) 3m8s (CPU)
Lite	0.92	0.84	49s (CPU)

* 1x Nvidia A4000
** 4 core CPU

Pseudoanonymisation

Automated free-text annonymisation tools

Data Aggregation

Grouping data to minimize unique identifiers

Entity Extraction

NLP Tools

Data Minimisation

Share the minimal viable number of records

Data Access

Multi-reviewer manual reading

42 Wallaby Way → **Sydney**

2 years old → **2-4 years old**

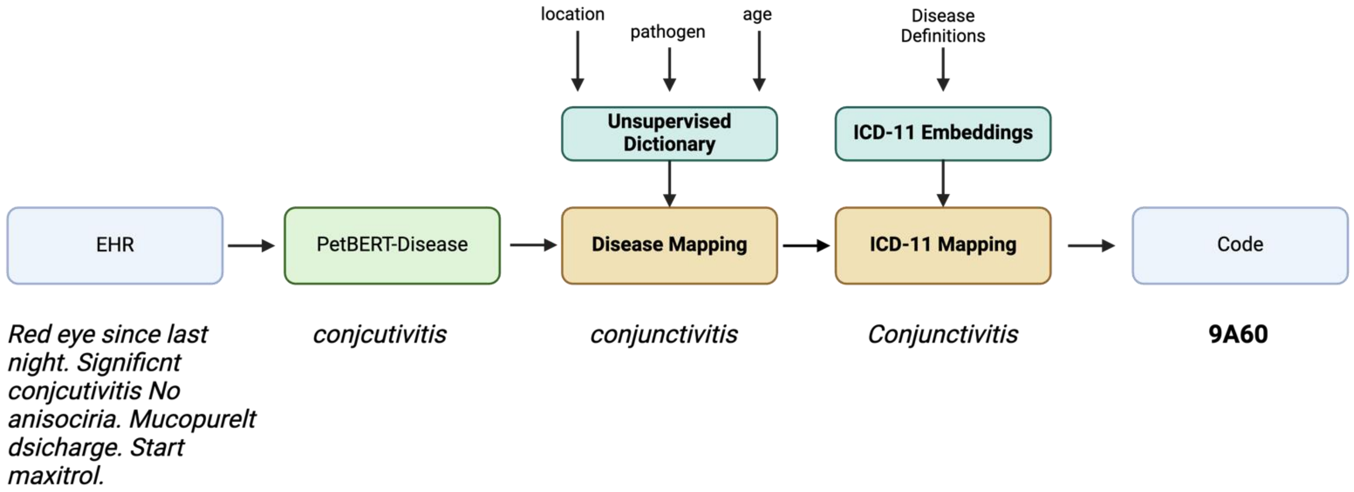
Pseudoanonymisation
Automated free-text annonymisation tools

Data Aggregation
Grouping data to minimize unique identifiers

Entity Extraction
NLP Tools

Data Minimisation
Share the minimal viable number of records

Data Access
Multi-reviewer manual reading



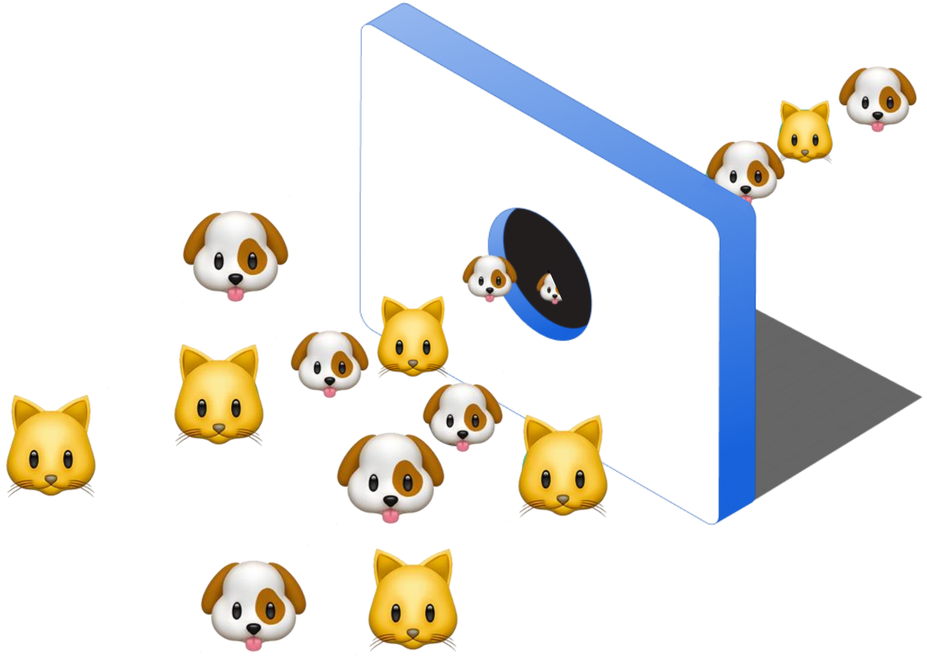
Pseudoanonymisation
Automated free-text annonymisation tools

Data Aggregation
Grouping data to minimize unique identifiers

Entity Extraction
NLP Tools

Data Minimisation
Share the minimal viable number of records

Data Access
Multi-reviewer manual reading



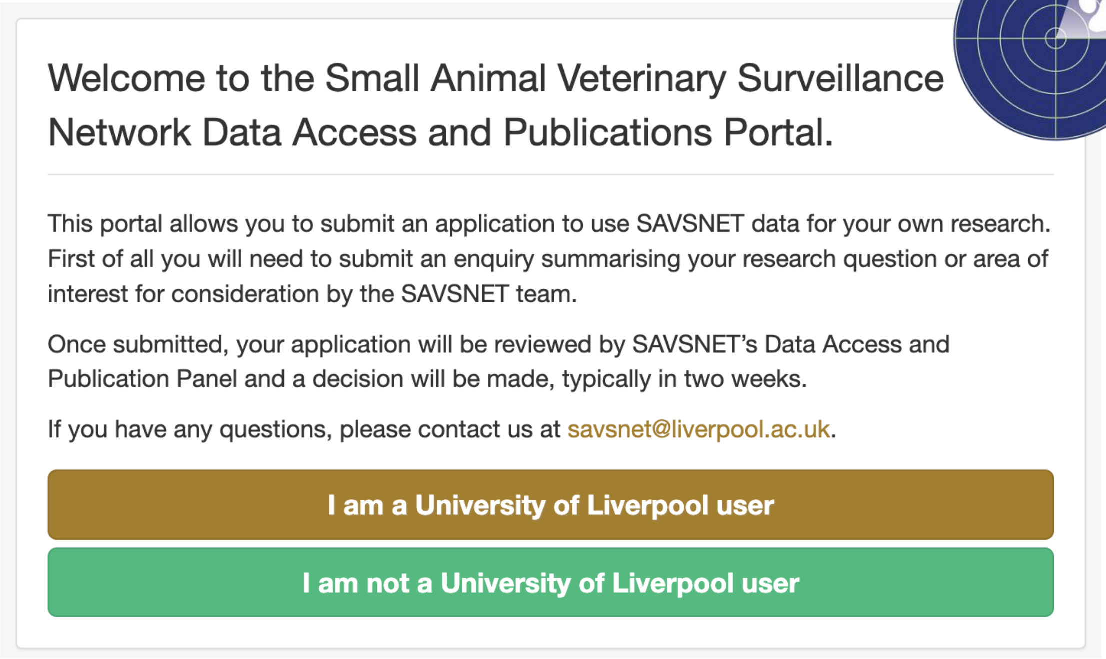
Pseudoanonymisation
Automated free-text annonymisation tools

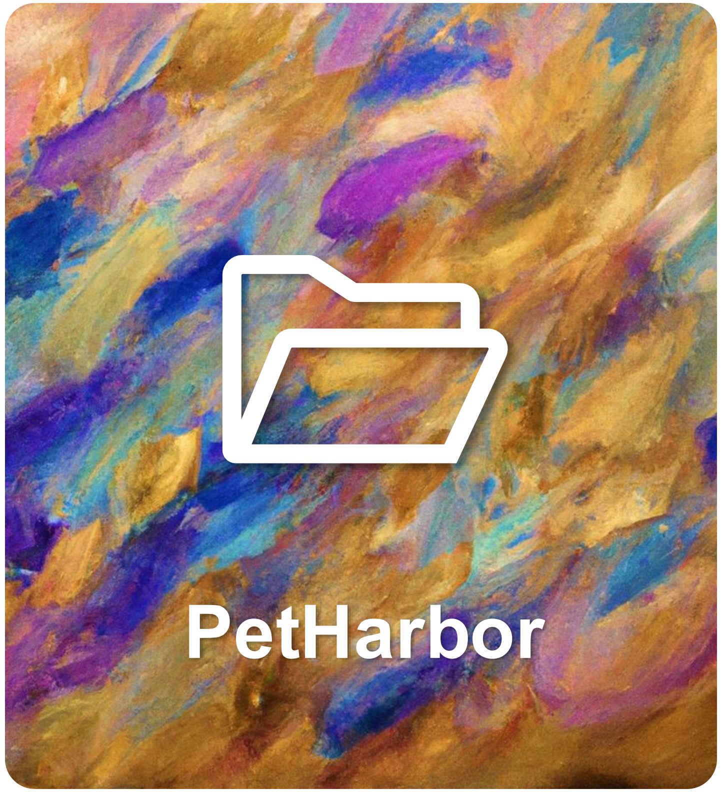
Data Aggregation
Grouping data to minimize unique identifiers

Entity Extraction
NLP Tools

Data Minimisation
Share the minimal viable number of records

Data Access
Multi-reviewer manual reading





www.nature.com/scientificdata

scientific **data**

Check for updates

OPEN

DATA DESCRIPTOR

MIMIC-IV, a freely accessible electronic health record dataset

Alistair E. W. Johnson[1,2]✉, Lucas Bulgarelli[1], Lu Shen[3], Alvin Gayles[3], Ayad Shammout[3], Steven Horng[3], Tom J. Pollard[1], Sicheng Hao[1], Benjamin Moody[1], Brian Gow[1], Li-wei H. Lehman[1], Leo A. Celi[1,3] & Roger G. Mark[1]

Digital data collection during routine clinical practice is now ubiquitous within hospitals. The data contains valuable information on the care of patients and their response to treatments, offering exciting opportunities for research. Typically, data are stored within archival systems that are not intended to support research. These systems are often inaccessible to researchers and structured for optimal storage, rather than interpretability and analysis. Here we present MIMIC-IV, a publicly available database sourced from the electronic health record of the Beth Israel Deaconess Medical Center. Information available includes patient measurements, orders, diagnoses, procedures, treatments, and deidentified free-text clinical notes. MIMIC-IV is intended to support a wide array of research studies and educational material, helping to reduce barriers to conducting clinical research.

Background

Thanks to the widespread adoption of electronic health record systems, data collected during routine clinical practice is now digitally stored in hospitals across the United States. Despite widespread storage of this data, archiving systems are often not designed to support research, making them difficult to navigate and access. In addition, routinely collected clinical data is often sporadic and noisy, reflecting the processes by which it was generated, where quality of data collection is understandably peripheral to the act of providing high quality care.

The intensive care unit (ICU) is an especially data-rich environment as patients require close monitoring. The typically acute nature of ICU patient illness and the importance of immediate intervention also make the environment of high-interest to researchers. Uniquely, there are a number of publicly available critical care datasets which have enabled research in this area. These projects largely build upon MIMIC, a waveform database with demographics digitally transcribed from paper records for over 90 patients[1]. MIMIC-II followed with a significantly increased sample size and breadth of information due to the clinical information being entirely sourced from various digital information systems[2]. More recently, MIMIC-III was published in 2015 and significantly expanded MIMIC-II, containing data for over 40,000 patients[3]. Outside of the MIMIC projects, a number of other critical care datasets have been made available to the worldwide research community. The eICU Collaborative Research Database (eICU-CRD) v2.0 comprises of 200,859 stays at ICUs and step-down units across 208 hospitals in the continental United States[4]. The AmsterdamUMCdb provides granular information for 23,106 admissions of 20,109 unique individuals admitted to a single academic medical center in the Netherlands[5]. The HiRID database contains high-resolution data for almost 34,000 admissions between 2008–2016 at Bern University Hospital in Switzerland[6,7]. HiRID contains 712 routinely collected physiological variables with one data entry every two minutes. The Pediatric Intensive Care (PIC) database is sourced from The Children's Hospital at Zhejiang University School of Medicine with 12,881 patients and 13,941 ICU stays admitted from 2010–2018[8].

Although the increasing number of datasets publicly available for research is encouraging, a number of areas for improvement remain. Data content varies considerably across datasets, with each having a particular strength. HiRID contains high resolution physiologic variables, eICU-CRD spans hundreds of distinct hospitals, while PIC contains pediatric patients. Clinical practice evolves quickly, requiring continual updating of the resources in order for derivative research to remain relevant. Finally, most datasets comprise of only one modality of information, clinical observations, and omit other important domains such as imaging, free-text, physiologic waveforms, and genomics.

[1]Massachusetts Institute of Technology, Cambridge, MA, USA. [2]The Hospital for Sick Children, Toronto, ON, Canada. [3]Beth Israel Deaconess Medical Center, Boston, MA, USA. ✉e-mail: aewj@mit.edu

SCIENTIFIC DATA | (2023) 10:1 | https://doi.org/10.1038/s41597-022-01899-x 1

www.nature.com/scientificdata

scientific data

Check for updates

OPEN

DATA DESCRIPTOR

MIMIC-IV, a freely accessible electronic health record dataset

Alistair E. W. Johnson[1,2✉], Lucas Bulgarelli[1], Lu Shen[3], Alvin Gayles[3], Ayad Shammout[3], Steven Horng[3], Tom J. Pollard[1], Sicheng Hao[1], Benjamin Moody[1], Brian Gow[1], Li-wei H. Lehman[1], Leo A. Celi[1,3] & Roger G. Mark[1]

Digital data collection during routine clinical practice is now ubiquitous within hospitals. The data contains valuable information on the care of patients and their response to treatments, offering exciting opportunities for research. Typically, data are stored within archival systems that are not intended to support research. These systems are often inaccessible to researchers and structured for optimal storage, rather than interpretability and analysis. Here we present MIMIC-IV, a publicly available database sourced from the electronic health record of the Beth Israel Deaconess Medical Center. Information available includes patient measurements, orders, diagnoses, procedures, treatments, and deidentified free-text clinical notes. MIMIC-IV is intended to support a wide array of research studies and educational material, helping to reduce barriers to conducting clinical research.

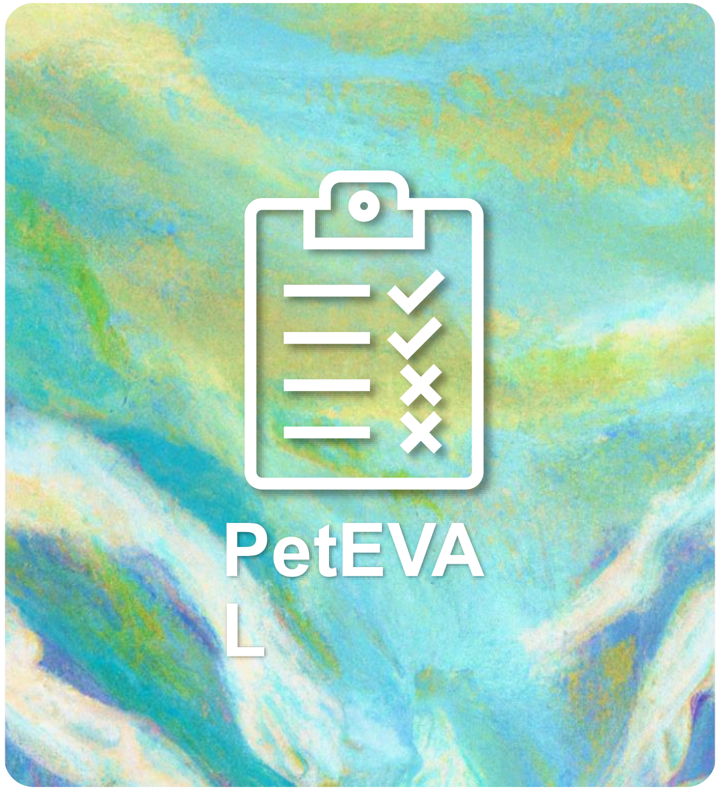
Background

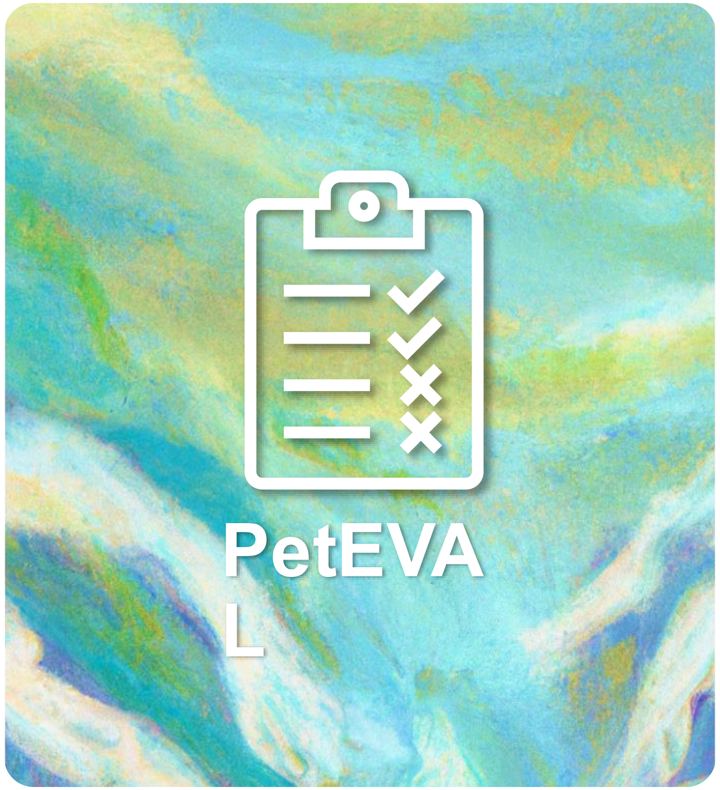
Thanks to the widespread adoption of electronic health record systems, data collected during routine clinical practice is now digitally stored in hospitals across the United States. Despite widespread storage of this data, archiving systems are often not designed to support research, making them difficult to navigate and access. In addition, routinely collected clinical data is often sporadic and noisy, reflecting the processes by which it was generated, where quality of data collection is understandably peripheral to the act of providing high quality care.

The intensive care unit (ICU) is an especially data-rich environment as patients require close monitoring. The typically acute nature of ICU patient illness and the importance of immediate intervention also make the environment of high-interest to researchers. Uniquely, there are a number of publicly available critical care datasets which have enabled research in this area. These projects largely build upon MIMIC, a waveform database with demographics digitally transcribed from paper records for over 90 patients[1]. MIMIC-II followed with a significantly increased sample size and breadth of information due to the clinical information being entirely sourced from various digital information systems[2]. More recently, MIMIC-III was published in 2015 and significantly expanded MIMIC-II, containing data for over 40,000 patients[3]. Outside of the MIMIC projects, a number of other critical care datasets have been made available to the worldwide research community. The eICU Collaborative Research Database (eICU-CRD) v2.0 comprises of 200,859 stays at ICUs and step-down units across 208 hospitals in the continental United States[4]. The AmsterdamUMCdb provides granular information for 23,106 admissions of 20,109 unique individuals admitted to a single academic medical center in the Netherlands[5]. The HiRID database contains high-resolution data for almost 34,000 admissions between 2008–2016 at Bern University Hospital in Switzerland[6,7]. HiRID contains 712 routinely collected physiological variables with one data entry every two minutes. The Pediatric Intensive Care (PIC) database is sourced from The Children's Hospital at Zhejiang University School of Medicine with 12,881 patients and 13,941 ICU stays admitted from 2010–2018[8].

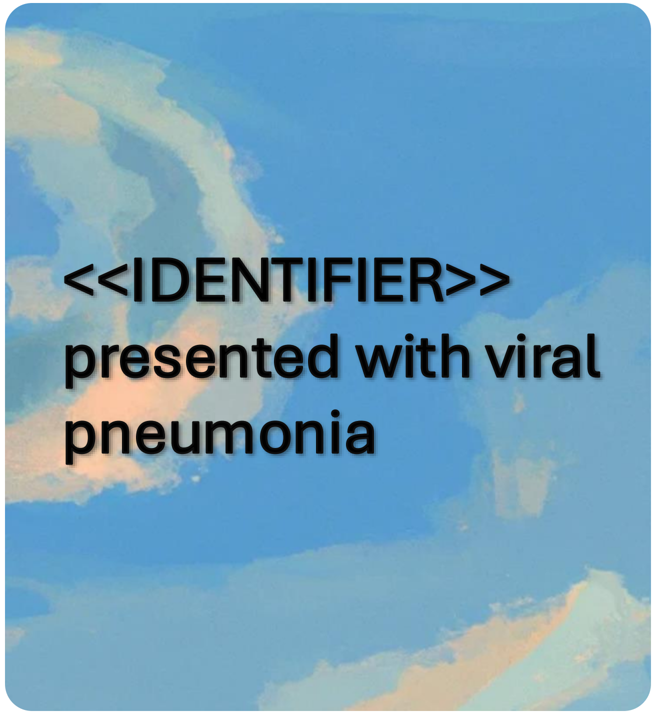
Although the increasing number of datasets publicly available for research is encouraging, a number of areas for improvement remain. Data content varies considerably across datasets, with each having a particular strength. HiRID contains high resolution physiologic variables, eICU-CRD spans hundreds of distinct hospitals, while PIC contains pediatric patients. Clinical practice evolves quickly, requiring continual updating of the resources in order for derivative research to remain relevant. Finally, most datasets comprise of only one modality of information, clinical observations, and omit other important domains such as imaging, free-text, physiologic waveforms, and genomics.

[1]Massachusetts Institute of Technology, Cambridge, MA, USA. [2]The Hospital for Sick Children, Toronto, ON, Canada. [3]Beth Israel Deaconess Medical Center, Boston, MA, USA. ✉e-mail: aewj@mit.edu

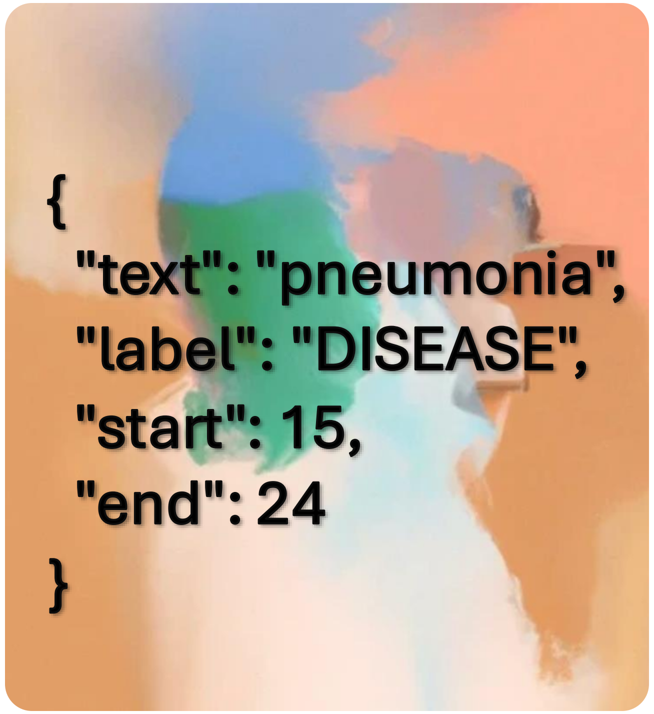




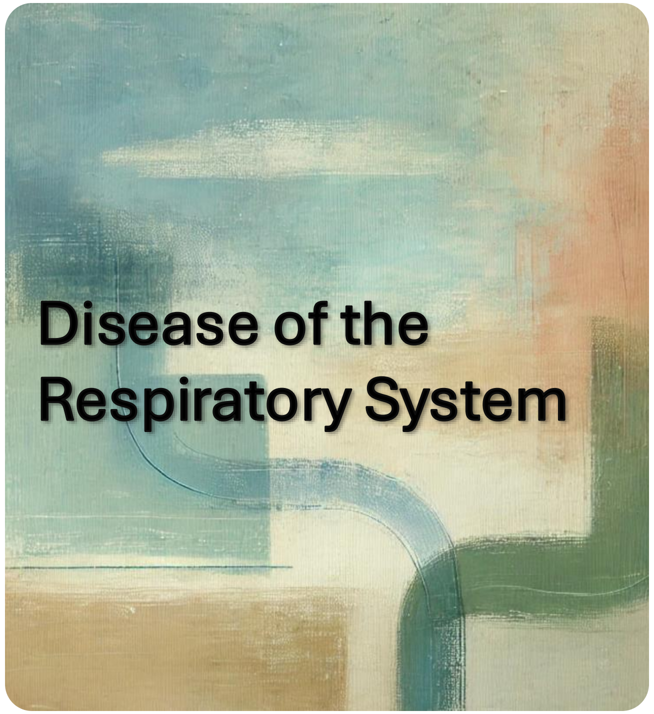
- **17,000 first opinion electronic health records**
  - 10,000 train set
  - 5,000 test set
  - 2,000 eval set
- **HuggingFace with leaderboard**
- **Full code availability**



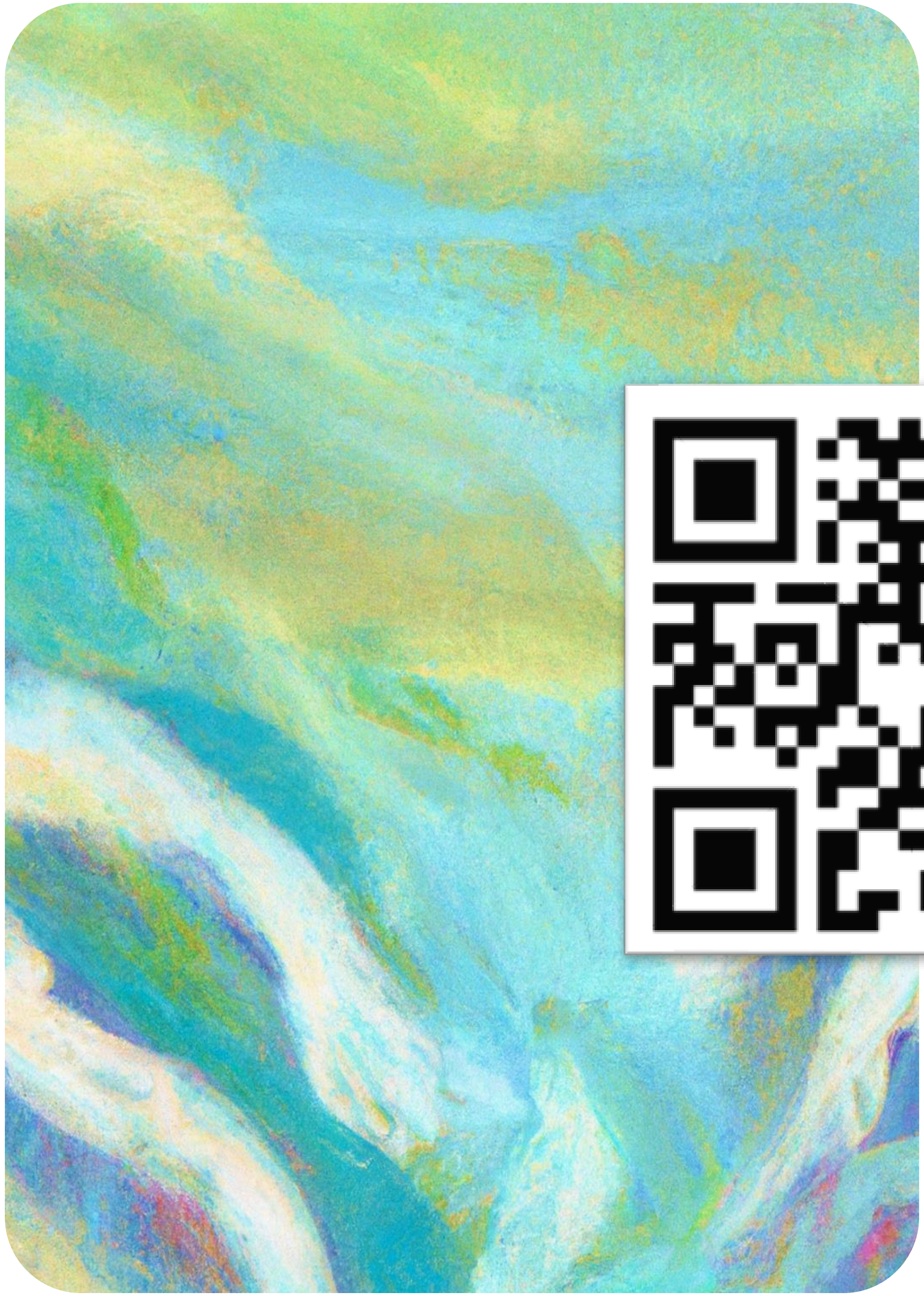
Anonymisation



Disease Extraction



Syndromic Classification



Thanks!

Sean Farrell
sean.farrell2@durham.ac.uk

BBSRC Durham University UNIVERSITY OF LIVERPOOL