MEDS & OHDSI

Matthew McDermott
Assistant Professor, Columbia University Department of Biomedical Informatics

mattmcdermott8@gmail.com

MEDS

https://medical-event-data-standard.github.io/

Outline

1. What is MEDS?
2. *Research Direction 1*: The ImageNet Moment for AI on Structured EHR
3. *Research Direction 2*: Learning the Latent Vocabulary of Medicine

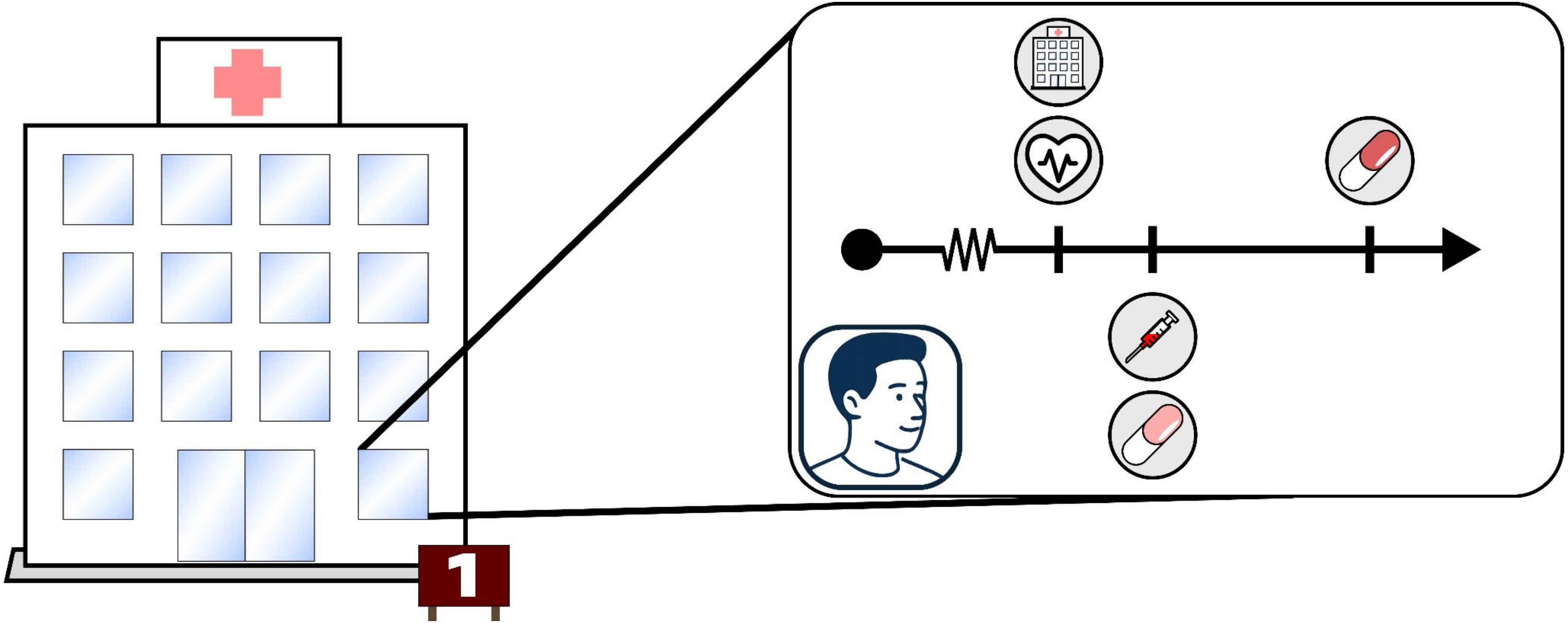
✘ *Research Direction 3*: Agentic Data Interactivity: OHDSI, MEDS, and GenAI

The Medical Event Data Standard (MEDS)

MEDS ML4H
2025 Tutorial

www.tinyurl.com/
MEDS-ML4H-pres

A Patient’s Data can be represented as a **Timeline**



Design Principles for MEDS

1. Transportability of model training *algorithms*
   The fundamental goal of MEDS is to enable *frictionless* transportability of *algorithms* first and foremost.

2. Capture the simplest view of the fundamental structure of health data
   MEDS asserts that EHR data can be most simply and fundamentally represented by capturing the longitudinal timeline of patient observations. Emphasizing simplicity in this way empowers and simplifies downstream use.

3. Empower development of an open-source ecosystem for health AI
   MEDS must enable development of an ecosystem of models, tools, & datasets.

4. Support Computational Performance for Foundation Models
   MEDS should support foundation-model scale systems.

MEDS Use Case

MEDS is designed first to enable development of models for the prediction of...

- ... some future property
- ... at an individual patient level
- ... within a single dataset

Design Principles for MEDS

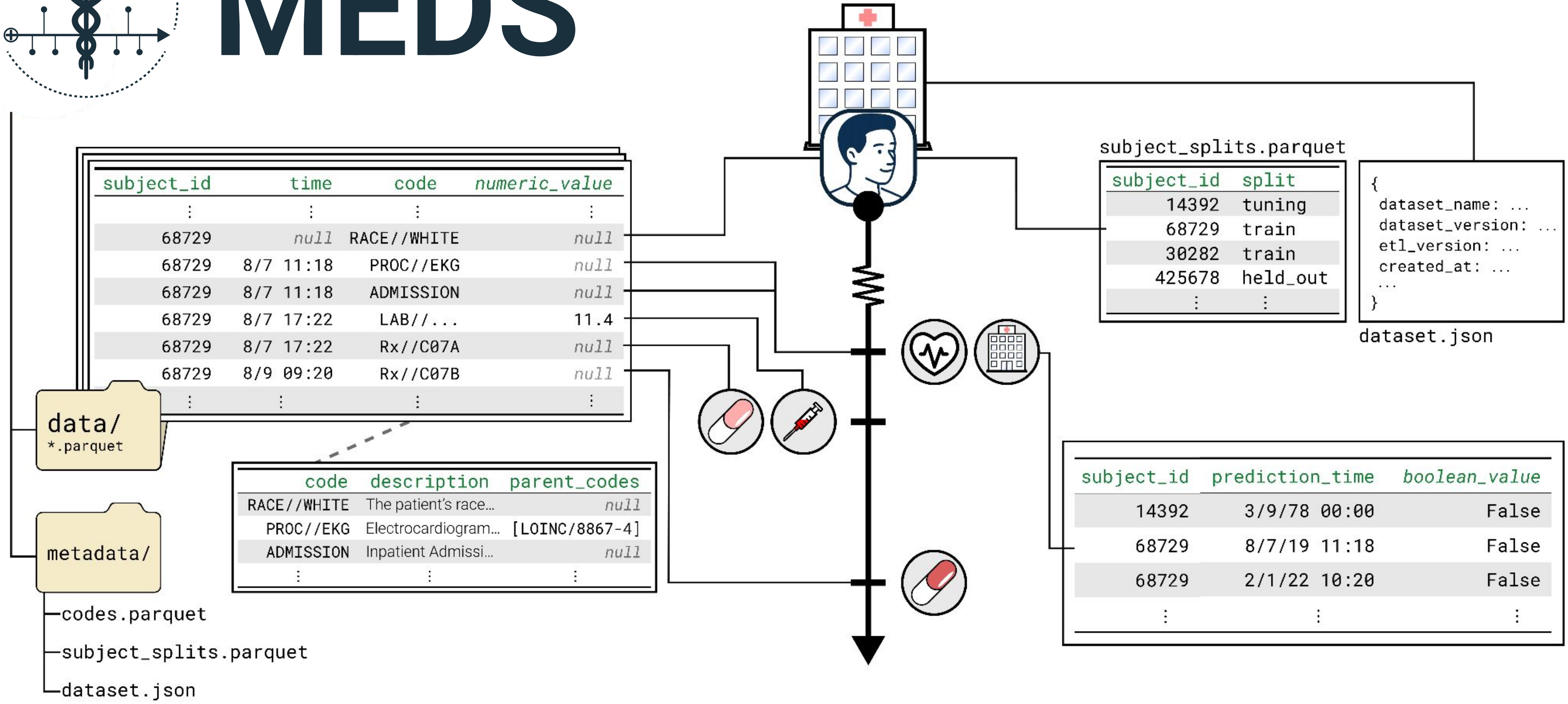
1. Transportability of model training *algorithms*
   The fundamental goal of MEDS is to enable *frictionless* transportability of *algorithms* first and foremost.

2. Capture the simplest view of the fundamental structure of health data
   MEDS asserts that EHR data can be most simply and fundamentally represented by capturing the longitudinal timeline of patient observations. Emphasizing simplicity in this way empowers and simplifies downstream use.

3. Empower development of an open-source ecosystem for health AI
   MEDS must enable development of an ecosystem of models, tools, & datasets.

4. Support Computational Performance for Foundation Models
   MEDS should support foundation-model scale systems.

MEDS



subject_id	time	code	numeric_value
⋮	⋮	⋮	⋮
68729	*null*	RACE//WHITE	*null*
68729	8/7 11:18	PROC//EKG	*null*
68729	8/7 11:18	ADMISSION	*null*
68729	8/7 17:22	LAB//...	11.4
68729	8/7 17:22	Rx//C07A	*null*
68729	8/9 09:20	Rx//C07B	*null*
⋮	⋮	⋮	⋮

code	description	parent_codes
RACE//WHITE	The patient's race...	*null*
PROC//EKG	Electrocardiogram...	[LOINC/8867-4]
ADMISSION	Inpatient Admissi...	*null*
⋮	⋮	⋮

subject_splits.parquet

subject_id	split
14392	tuning
68729	train
30282	train
425678	held_out
⋮	⋮

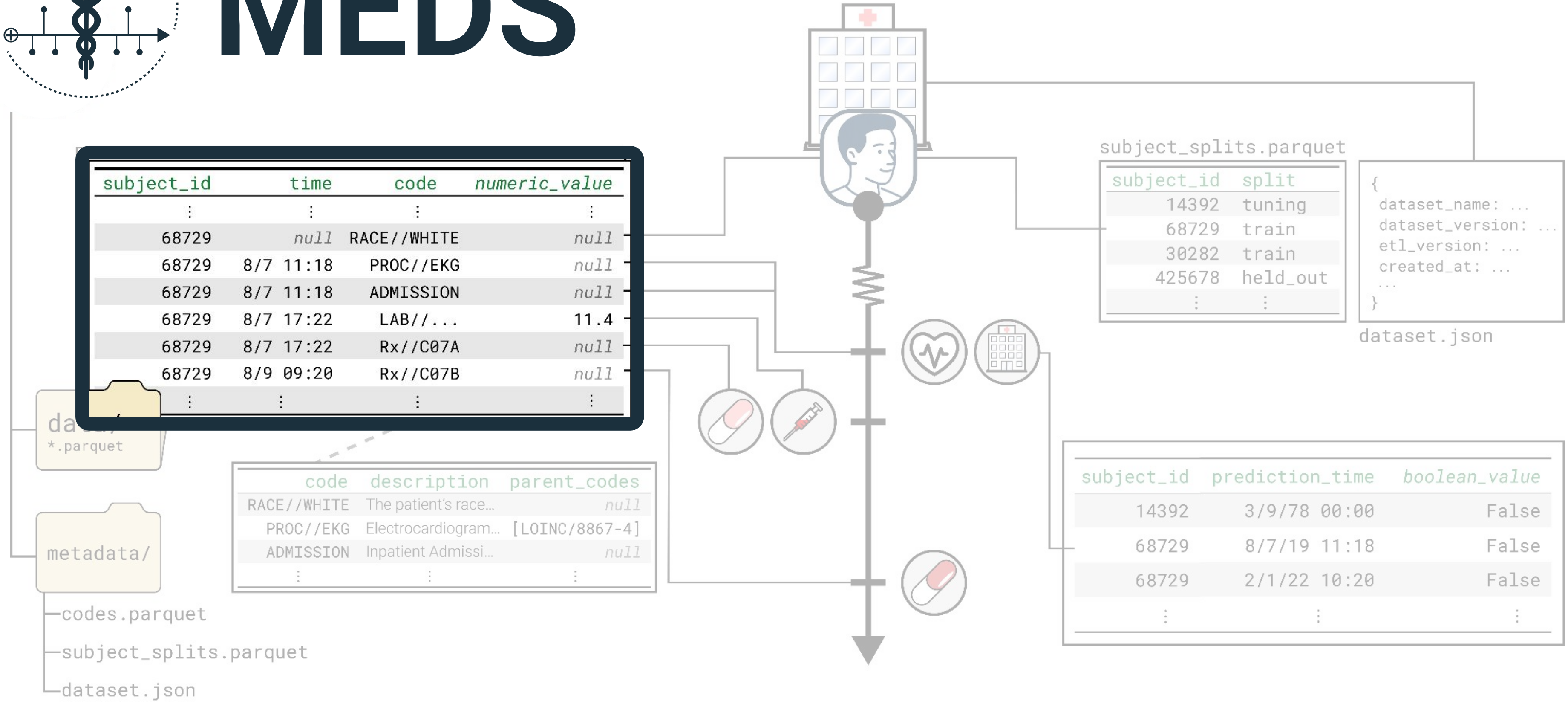
```
{
  dataset_name: ...
  dataset_version: ...
  etl_version: ...
  created_at: ...
  ...
}
```

dataset.json

subject_id	prediction_time	boolean_value
14392	3/9/78 00:00	False
68729	8/7/19 11:18	False
68729	2/1/22 10:20	False
⋮	⋮	⋮

- data/
  - *.parquet
- metadata/
  - codes.parquet
  - subject_splits.parquet
  - dataset.json

MEDS



subject_id	time	code	*numeric_value*
⋮	⋮	⋮	⋮
68729	*null*	RACE//WHITE	*null*
68729	8/7 11:18	PROC//EKG	*null*
68729	8/7 11:18	ADMISSION	*null*
68729	8/7 17:22	LAB//...	11.4
68729	8/7 17:22	Rx//C07A	*null*
68729	8/9 09:20	Rx//C07B	*null*
⋮	⋮	⋮	⋮

code	description	parent_codes
RACE//WHITE	The patient's race...	*null*
PROC//EKG	Electrocardiogram...	[LOINC/8867-4]
ADMISSION	Inpatient Admissi...	*null*
⋮	⋮	⋮

subject_splits.parquet

subject_id	split
14392	tuning
68729	train
30282	train
425678	held_out
⋮	⋮

```
{
  dataset_name: ...
  dataset_version: ...
  etl_version: ...
  created_at: ...
  ...
}
```

dataset.json

subject_id	prediction_time	*boolean_value*
14392	3/9/78 00:00	False
68729	8/7/19 11:18	False
68729	2/1/22 10:20	False
⋮	⋮	⋮

data/
*.parquet

metadata/
- codes.parquet
- subject_splits.parquet
- dataset.json

Data Shards

subject_id	time	code	numeric_value
⋮	⋮	⋮	⋮
68729	*null*	RACE//WHITE	*null*
68729	8/7 11:18	PROC//EKG	*null*
68729	8/7 11:18	ADMISSION	*null*
68729	8/7 17:22	LAB//...	11.4
68729	8/7 17:22	Rx//C07A	*null*
68729	8/9 09:20	Rx//C07B	*null*
⋮	⋮	⋮	⋮

- Each shard contains all data for a patient, in time order.

Data Shards

subject_id	time	code	*numeric_value*
⋮	⋮	⋮	⋮
68729	*null*	RACE//WHITE	*null*
68729	8/7 11:18	PROC//EKG	*null*
68729	8/7 11:18	ADMISSION	*null*
68729	8/7 17:22	LAB//...	11.4
68729	8/7 17:22	Rx//C07A	*null*
68729	8/9 09:20	Rx//C07B	*null*
⋮	⋮	⋮	⋮

- Each shard contains all data for a patient, in time order.
- The code column is *unconstrained*.

Data Shards

subject_id	time	code	*numeric_value*
⋮	⋮	⋮	⋮
68729	*null*	RACE//WHITE	*null*
68729	8/7 11:18	PROC//EKG	*null*
68729	8/7 11:18	ADMISSION	*null*
68729	8/7 17:22	LAB//...	11.4
68729	8/7 17:22	Rx//C07A	*null*
68729	8/9 09:20	Rx//C07B	*null*
⋮	⋮	⋮	⋮

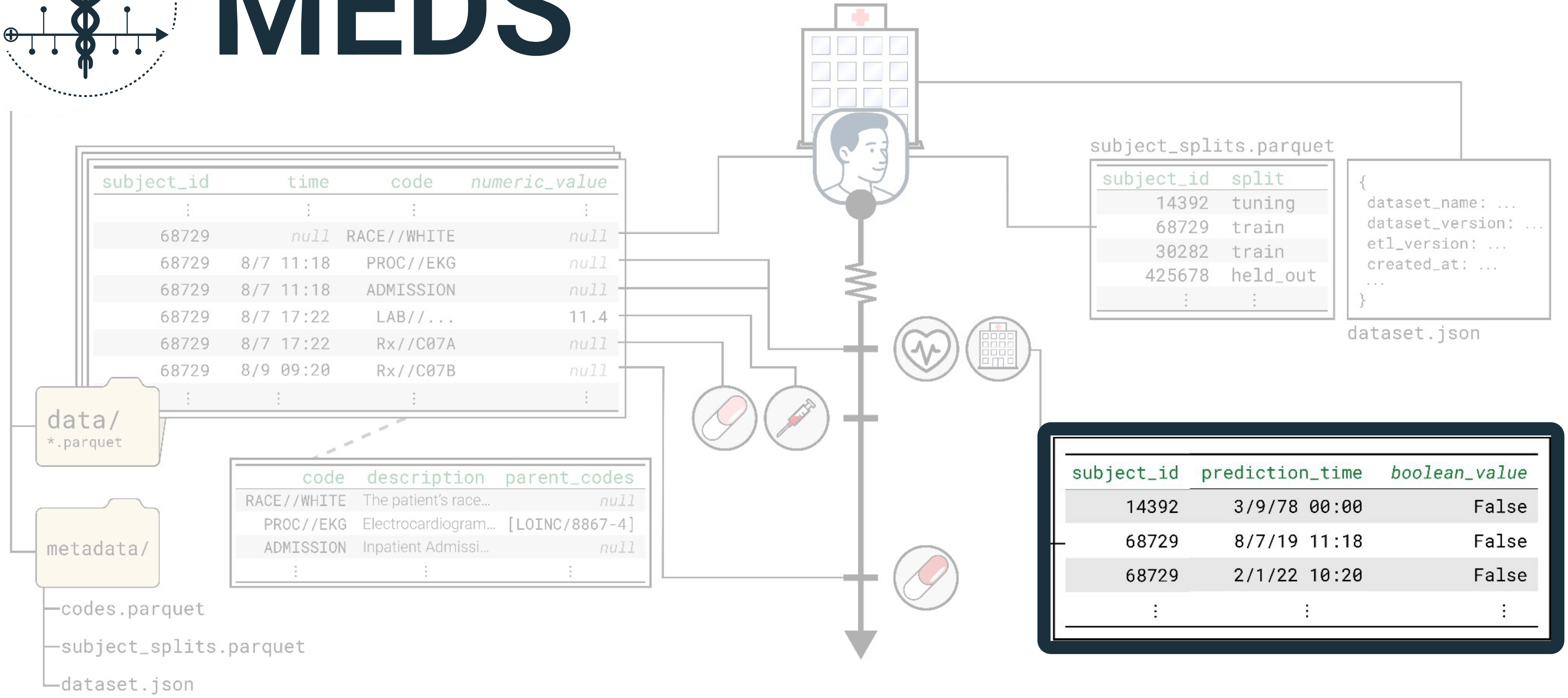
- Each shard contains all data for a patient, in time order.
- The code column is *unconstrained*.
- Additional columns can be included as needed.

Data Shards

subject_id	time	code	*numeric_value*
⋮	⋮	⋮	⋮
68729	*null*	RACE//WHITE	*null*
68729	8/7 11:18	PROC//EKG	*null*
68729	8/7 11:18	ADMISSION	*null*
68729	8/7 17:22	LAB//...	11.4
68729	8/7 17:22	Rx//C07A	*null*
68729	8/9 09:20	Rx//C07B	*null*
⋮	⋮	⋮	⋮

- Each shard contains all data for a patient, in time order.
- The code column is *unconstrained*.
- Additional columns can be included as needed.
- Intervals represented via separate start and end rows.

MEDS



subject_id	time	code	numeric_value
⋮	⋮	⋮	⋮
68729	*null*	RACE//WHITE	*null*
68729	8/7 11:18	PROC//EKG	*null*
68729	8/7 11:18	ADMISSION	*null*
68729	8/7 17:22	LAB//...	11.4
68729	8/7 17:22	Rx//C07A	*null*
68729	8/9 09:20	Rx//C07B	*null*
⋮	⋮	⋮	⋮

data/
*.parquet

code	description	parent_codes
RACE//WHITE	The patient's race…	*null*
PROC//EKG	Electrocardiogram…	[LOINC/8867-4]
ADMISSION	Inpatient Admissi…	*null*
⋮	⋮	⋮

metadata/
- codes.parquet
- subject_splits.parquet
- dataset.json

subject_splits.parquet

subject_id	split
14392	tuning
68729	train
30282	train
425678	held_out
⋮	⋮

```
{
  dataset_name: ...
  dataset_version: ...
  etl_version: ...
  created_at: ...
  ...
}
```

dataset.json

subject_id	prediction_time	boolean_value
14392	3/9/78 00:00	False
68729	8/7/19 11:18	False
68729	2/1/22 10:20	False
⋮	⋮	⋮

MEDS Label Schema

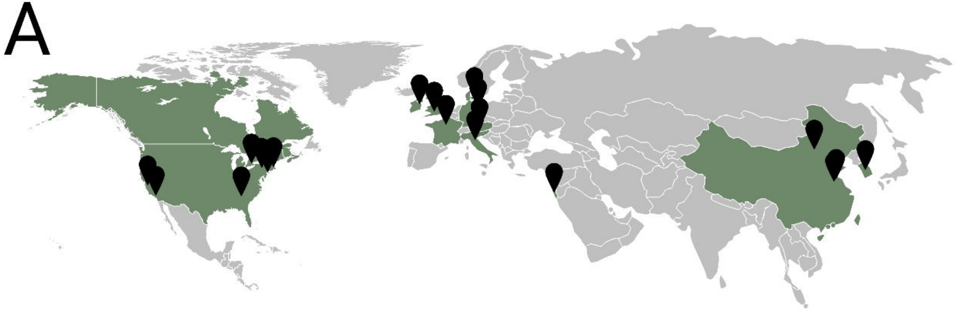
The MEDS label schema requires an index (a `subject_id`, and `prediction_time`) and permits optional labels of type including `boolean_value`, `integer_value`, `float_value`, and `categorical_value`. These labels can be predicted using any of the data of the indexed subject that occurred anytime at or before the indexed prediction time.

subject_id	prediction_time	*boolean_value*
68729	3/9/78 00:00	False
68729	5/2/10 14:22	False
68729	5/2/10 14:34	False
125829	4/9/18 18:19	True

… Why should you use it?

MEDS

A



Used in at least 21 institutions

B

27 Published Articles

17 Datasets

12 Models

C

~1.9-40000x faster than prior tools

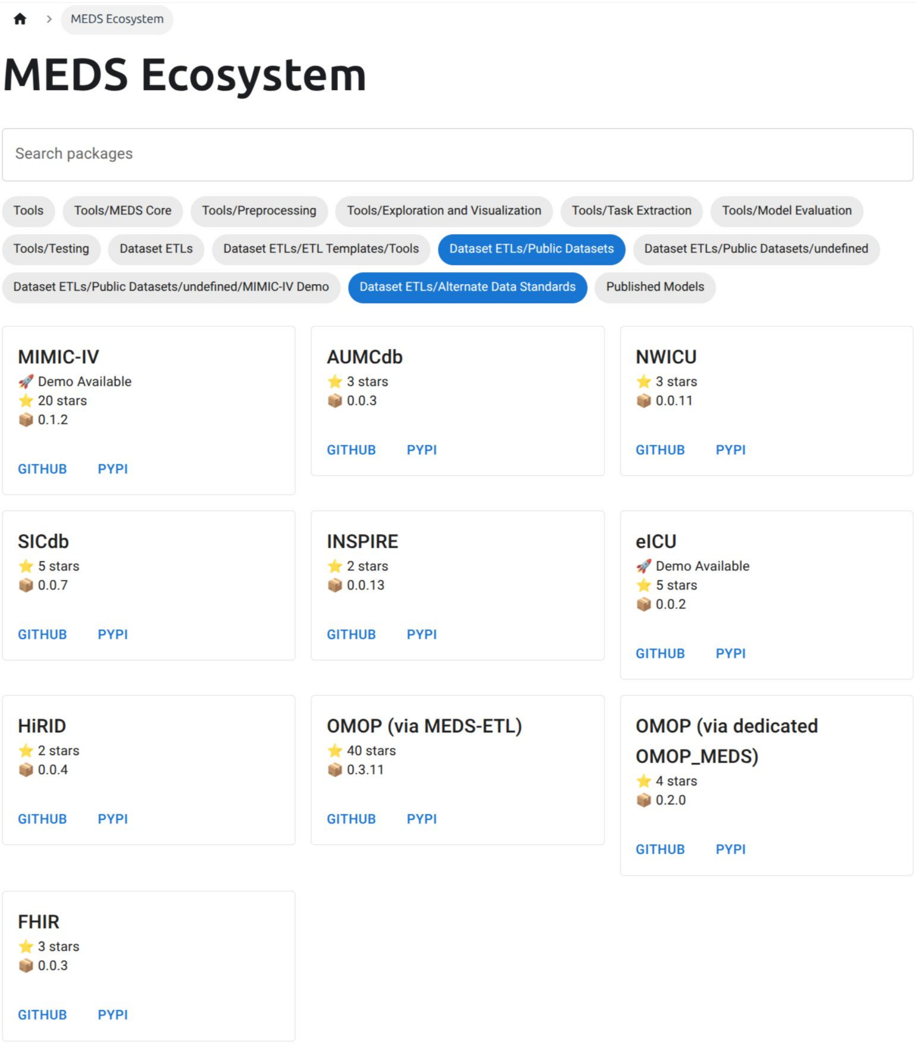
33-70% fewer lines of code

More than 14 tools available

There is a turnkey OMOP to MEDS converter.

Turnkey MEDS ETLs

- OMOP
- FHIR
- MIMIC-IV
- eICU
- NWICU
- AUMCdb
- SICdb
- INSPIRE
- HiRID



MEDS Ecosystem

MIMIC-IV
🚀 Demo Available
⭐ 20 stars
📦 0.1.2
GITHUB PYPI

AUMCdb
⭐ 3 stars
📦 0.0.3
GITHUB PYPI

NWICU
⭐ 3 stars
📦 0.0.11
GITHUB PYPI

SICdb
⭐ 5 stars
📦 0.0.7
GITHUB PYPI

INSPIRE
⭐ 2 stars
📦 0.0.13
GITHUB PYPI

eICU
🚀 Demo Available
⭐ 5 stars
📦 0.0.2
GITHUB PYPI

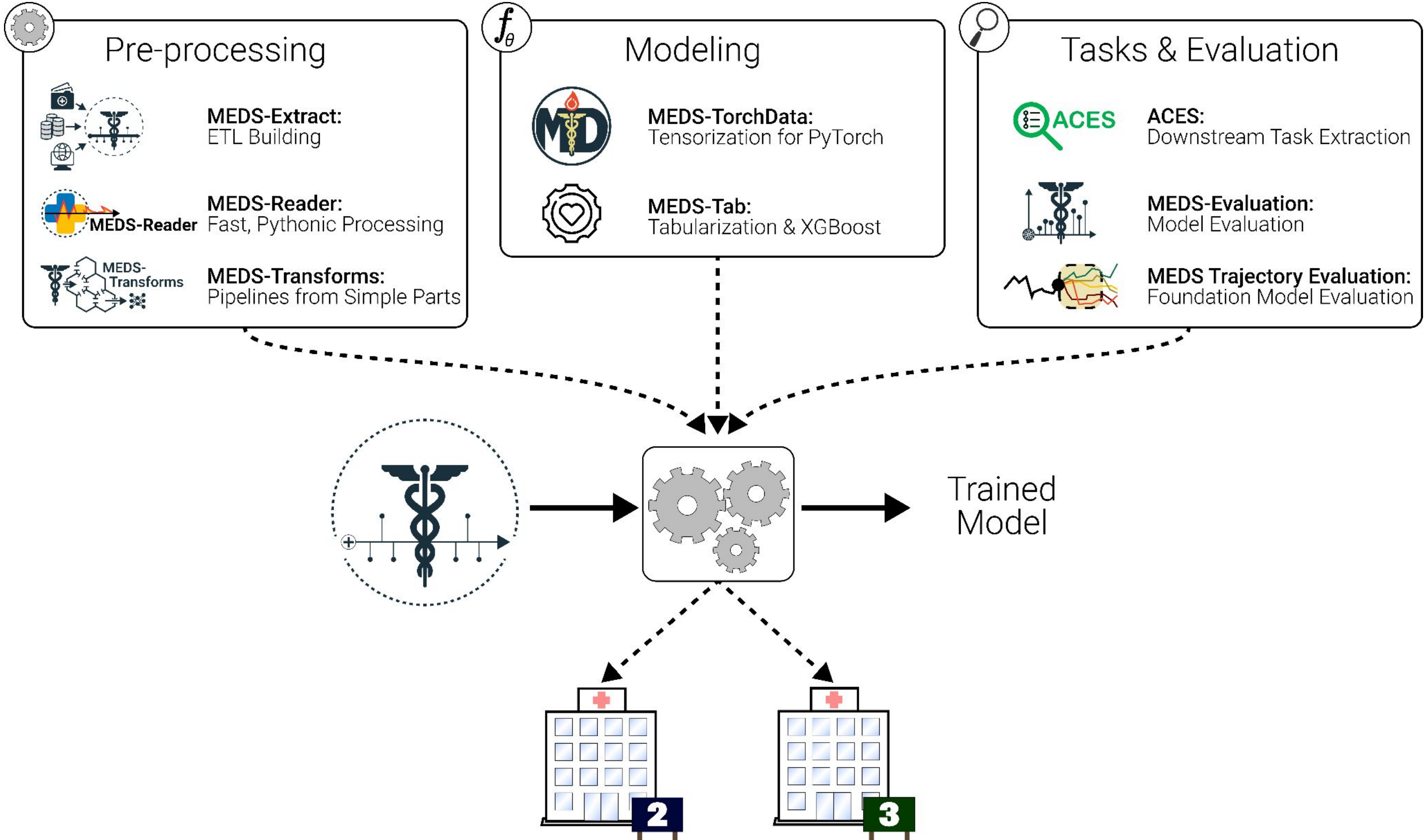
HiRID
⭐ 2 stars
📦 0.0.4
GITHUB PYPI

OMOP (via MEDS-ETL)
⭐ 40 stars
📦 0.3.11
GITHUB PYPI

OMOP (via dedicated OMOP_MEDS)
⭐ 4 stars
📦 0.2.0
GITHUB PYPI

FHIR
⭐ 3 stars
📦 0.0.3
GITHUB PYPI

MEDS



Pre-processing

MEDS-Extract: ETL Building

MEDS-Reader: Fast, Pythonic Processing

MEDS-Transforms: Pipelines from Simple Parts

Modeling

MEDS-TorchData: Tensorization for PyTorch

MEDS-Tab: Tabularization & XGBoost

Tasks & Evaluation

ACES: Downstream Task Extraction

MEDS-Evaluation: Model Evaluation

MEDS Trajectory Evaluation: Foundation Model Evaluation

MEDS Models

- MEDS-EIC-AR (AR FM)
- ETHOS (AR-FM)
- MOTOR
- CLMBR
- CEHR-BERT/-GPT/-XGPT
- Core-BEHRT
- GenHPF
- EHRMamba
- MEDS-Tab
- EveryQuery

MEDS "Everything-is-code" Autoregressive Model

A MEDS, "Everything-is-code" style Autoregressive Generative Model, capable of zero-shot inference.

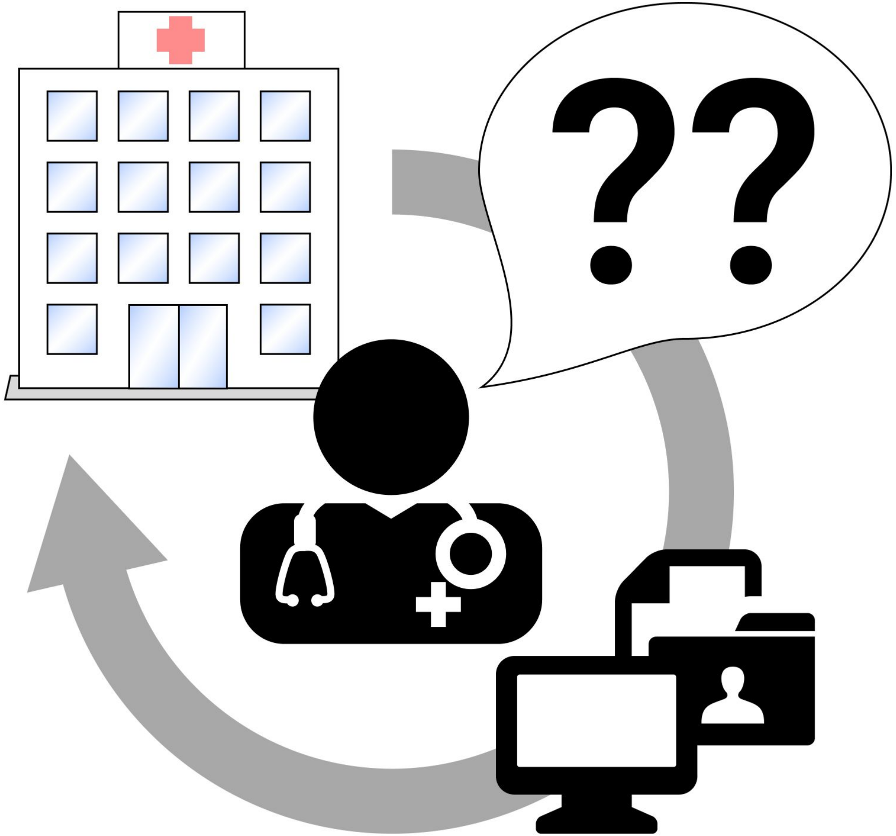
Installation

```
pip install MEDS-EIC-AR
```

Research Direction 1:
The “ImageNet” Moment for Health AI

Question:

A new colleague asks you: What model should they build for their data?



Driving Problem: Health AI has a Reproducibility Crisis

We reproduced datasets for 38 experiments corresponding to 28 published studies using MIMIC. In half of the experiments, the sample size we acquired was 25% greater or smaller than the sample size reported.

- Alistair Johnson et. al., 2017

https://proceedings.mlr.press/v68/johnson17a.html

MLH papers scored even more poorly when it came to code release... with only ~21% of the papers we analyzed releasing their code publicly

- Matthew McDermott et. al., 2021

https://doi.org/10.1126/scitranslmed.abb1655

Of the 218 included articles, 73 (34%) shared code, with 24 (33% of code sharing articles and 11% of all articles) sharing reproducible code

- Kesavan Venkatesh et. al., 2022

https://doi.org/10.1148/ryai.220081

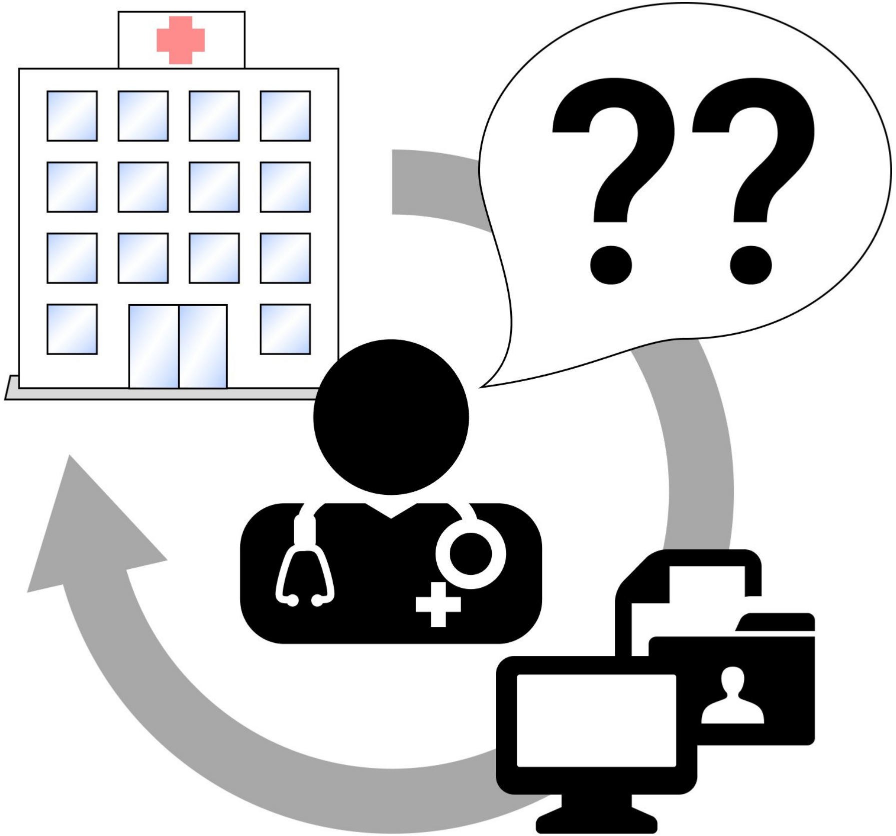
The (lack of?) Science of Machine Learning for Healthcare.

McDermott, M. (2025). *Proceedings of the 4th Machine Learning for Health Symposium*, in *Proceedings of Machine Learning Research* 259:19-29

Available from https://proceedings.mlr.press/v259/mcdermott25a.html

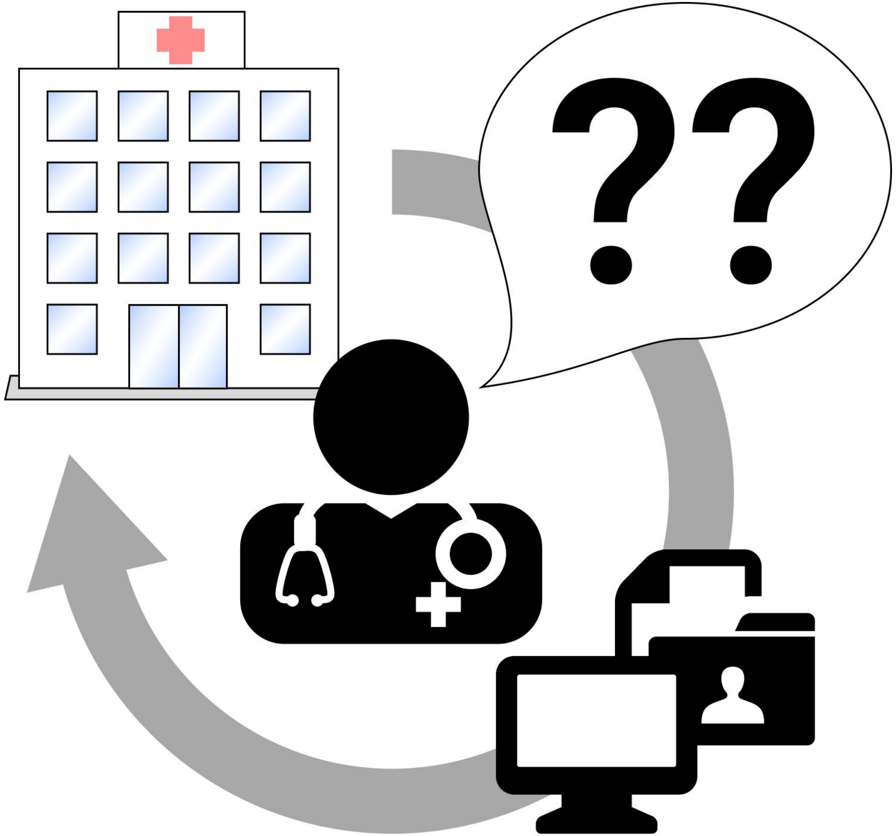
Question:

A new colleague asks you: What model should they build for their data?

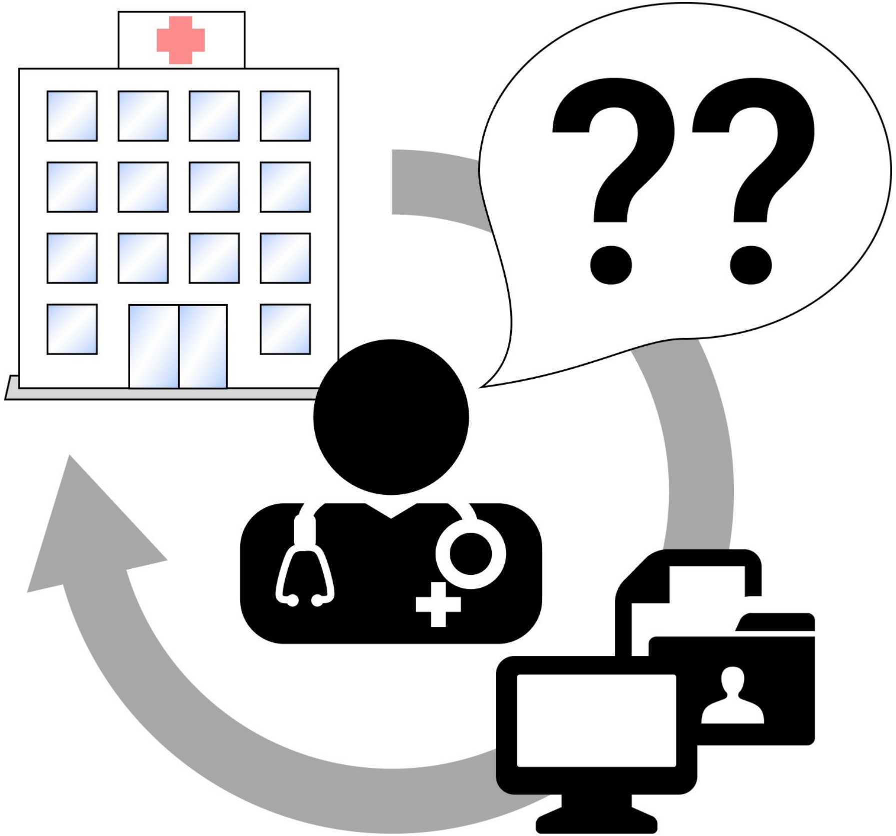


To answer this question, we need empirical evidence of what models work best, when trained on...

- Data like the target data
- Tasks like the target task

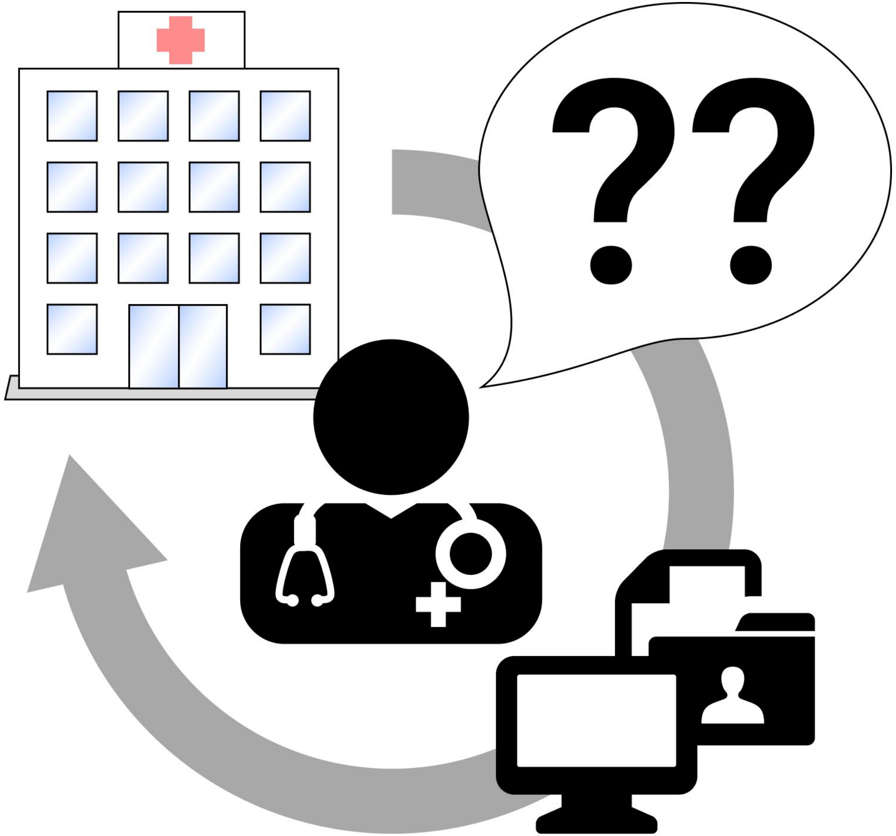


To answer this question, we need a benchmark!



To answer this question, we need a benchmark!

(that’s designed for health AI)

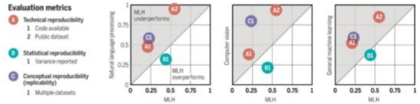


MEDS-DEV: Enabling Reproducible Science of AI for EHR Data

Aleksia Kolo, Chao Pang, Edward Choi, Ethan Steinberg, Hyewon Jeong, Jack Gallifant, Jason A. Fries, Jeffrey N. Chiang, Jungwoo Oh, Justin Xu, Kamilė Stankevičiūtė, Kiril V. Klein, Matthew McDermott, Mikkel Odgaard, Nassim Oufattole, Patrick Rockenschaub, Pawel Renc, Robin van de Water, Shalmali Joshi, Simon A. Lee, Teya S. Bergamaschi, Tom J. Pollard, Vincent Jeanselme, Young Sang Choi

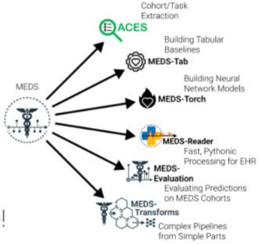
Health AI has a Reproducibility Crisis

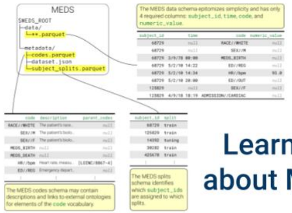
Health AI faces a systemic reproducibility crisis, limiting our ability to do effective science. The MEDS health AI ecosystem, including ACES and MEDS-DEV, changes that.



MEDS: A Health AI Data Ecosystem

- MEDS enables a robust ecosystem of health AI tools.
- This allows groups to share software, models, & results!
- Check out the link below to learn more!





- MEDS is simple and easy to use.
- Converting to MEDS is easy!

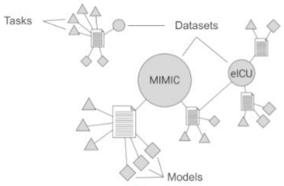
Learn More about MEDS:

MEDS-DEV
Decentralized Extensible Validation

Meaningful shared tasks in healthcare require:

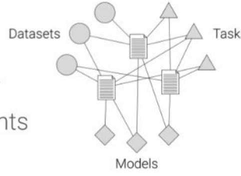
- Decentralized evaluation over private data.
- *Frictionlessly reproducible* task definitions, data preprocessing pipelines, model training recipes, and evaluation protocols.
- A community curated set of tasks spanning clinical problems, settings, and data needs.

MEDS-DEV enables meaningful benchmarking and reproducible AI in ML4H by satisfying these needs.



Benchmarks today are disconnected and irreproducible, leading to uninformative results and a lack of progress.

MEDS-DEV enables more interconnected, comparable experiments and better science



How do you use MEDS-DEV?

Need	Tool
Reproducible, transparent task definitions	ACES task configs are transparent, reproducible, & can be shared across datasets
Shared evaluation protocols	MEDS-Evaluation ensures evaluation is consistent and aspects such as fairness and equity can be included
Communal contributions on tasks & models Decentralized, extensible data and task support	MEDS-DEV configuration files and results are stored & versioned in an open-source GitHub, enabling communal development and decentralization of results

- Adding new results and models is easy!
- Preliminary results showcase adding new models, tasks, and datasets!

Model	**MIMIC-IV/Columbia** Long LOS	 ICU Mortality
Log. Reg.	0.752/0.677	0.754/0.509
LightGBM	0.783/0.757	0.798/0.661
MOTOR	0.804/0.735	**0.854**/0.727
CEHR-BERT	0.808/0.741	0.845/0.726
MEDS-Tab	**0.811/0.761**	0.830/**0.785**
GenHPF	0.779/0.662	0.790/0.633

Check out detailed MEDS-DEV Tutorials Here:

Acknowledgements

MBAM gratefully acknowledges support from a Berkowitz Postdoctoral Fellowship; KS acknowledges support from AstraZeneca; JAF acknowledges support for this effort from the Debra and Mark Leslie fund for supporting AI in Healthcare research; TP is supported by the National Institutes of Health (NIH OT2OD032701 and NIH R01EB030362); Edward Choi is supported by the IITP grant (No.RS-2019-II190075) and the NRF grant (NRF-2020H1D3A2A03100945) funded by the Korean government; Part of the PHAIR project supported by the Innovation Fund Denmark; NHS acknowledges support for this effort from the Debra and Mark Leslie fund for supporting AI in Healthcare researchPioneer Centre for AI, DNRF Grant P1; Novo Nordisk Foundation

What do we need:

1. A large set of datasets in the MEDS format, spanning diverse clinical settings.

What do we need:

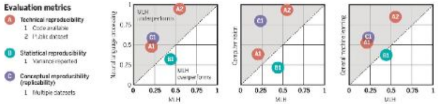
1. A large set of datasets in the MEDS format, spanning diverse clinical settings.
2. A collection of tasks that are “conceptually identical” across disparate datasets.

ACES: Reproducible Extraction of Task Cohorts for EHRs

Justin Xu, Jack Gallifant, Alistair E. W. Johnson, Matthew B. A. McDermott

Health AI has a Reproducibility Crisis

Health AI faces a **systemic reproducibility crisis**, limiting our ability to do effective science. We need to build a health AI ecosystem to change that, and ACES builds on that foundation.



ACES Leverages Event-Stream Schemas

The MEDS data schema epitomizes simplicity and has only 4 required columns: `subject_id`, `time`, `code`, and `numeric_value`.

subject_id	time	code	numeric_value
68729	null	RACE//WHITE	null
68729	null	SEX//M	null
68729	3/9/78 00:00	MEDS_BIRTH	null
68729	5/2/10 14:22	ED//REG	null
68729	5/2/10 14:34	HR//bpm	93.0
68729	5/2/10 20:00	ED//OUT	null
125829	null	SEX//F	null
125829	4/9/18 18:19	ADMISSION//CARDIAC	null

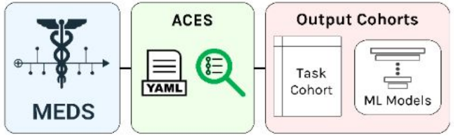
code	description	parent_codes
RACE//WHITE	The patient's race is r...	null
SEX//M	The patient's biologic...	null
SEX//F	The patient's biologic...	null
MEDS_BIRTH	null	null
MEDS_DEATH	null	null
HR//bpm	Heart rate, measured...	[LOINC/8867-4]
ED//REG	Emergency departme...	null
⋮	⋮	⋮

The MEDS codes schema *may* contain descriptions and links to external ontologies for elements of the `code` vocabulary.

- Single stream of events!
- Simple and flexible to use!
- Easy transformations!

Learn More about MEDS:

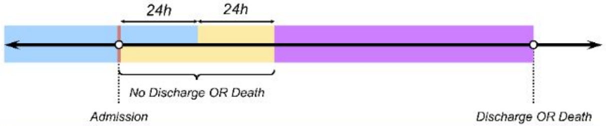
Task Extraction Made Easy!



- Transparent
- Reproducible
- Extractable on diverse datasets

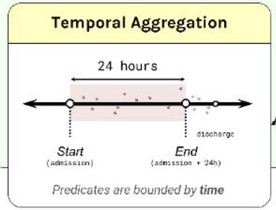
ACES Configuration Files

In-hospital Mortality: Given the first 24 hours of a patient's stay, predict whether or not they will die within this hospital admission, with a gap time of 48h.

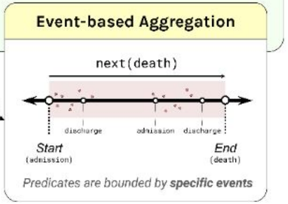


```
predicates:
  admission:
    code: ADMISSION
  discharge:
    code: DISCHARGE
  death:
    code: DEATH
  discharge_or_death:
    expr: or(discharge, death)

trigger: admission
```

```
windows:
  input:
    start: NULL
    end: trigger + 24h
    start_inclusive: True
    end_inclusive: True

  gap:
    start: trigger
    end: start + 48h
    start_inclusive: False
    end_inclusive: True

  target:
    start: gap.end
    end: start -> discharge_or_death
    start_inclusive: False
    end_inclusive: True
    label: death
```



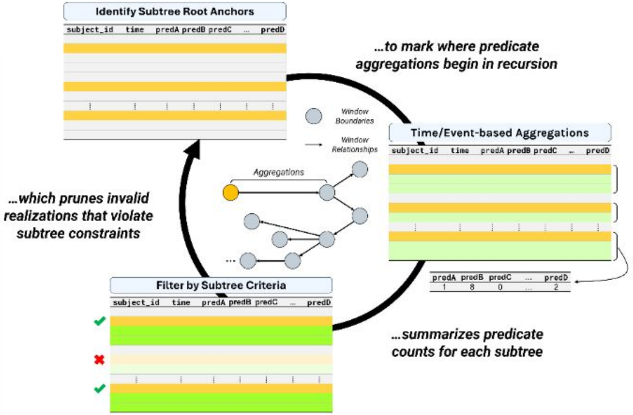
Predicates are bounded by ***time***



Predicates are bounded by ***specific events***

ACES Demo & MEDS Tutorial:

Recursive Algorithm for Extraction



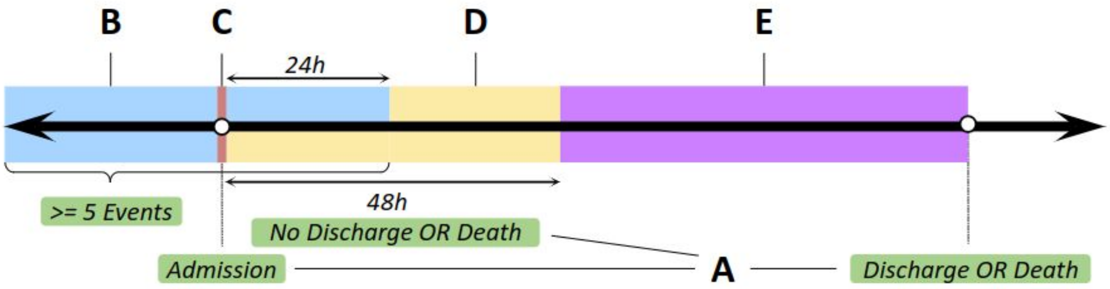
How can ACES help you?

- For a variety of pre-defined tasks, check out (or contribute to) MEDS-DEV:
- For more info on how to write your own ACES configs, check out our documentation:

Acknowledgements

MBAM gratefully acknowledges support from a Berkowitz Postdoctoral Fellowship at Harvard Medical School. JG is funded by the National Institutes of Health through NIH-USA R01CA294033. JX greatly appreciates support from supervisors David Eyre (University of Oxford) and Curtis Langlotz (Stanford University). We also acknowledge valuable contributions by Tom Pollard (Massachusetts Institute of Technology) and by the broader MEDS ecosystem of contributors and users.

In-hospital Mortality Prediction



Task Configuration

A

```
predicates:
    admission:
        code: ADMISSION
    discharge:
        code: DISCHARGE
    death:
        code: DEATH
    discharge_or_death:
        expr: or(discharge, death)
```

B

```
input:
    start: NULL
    end: trigger + 24 hours
    start_inclusive: True
    end_inclusive: True
    has:
        _ANY_EVENT: (5, None)
```

C

```
trigger: admission
```

D

```
gap:
    start: trigger
    end: start + 48 hours
    start_inclusive: False
    end_inclusive: True
    has:
        admission: (None, 0)
        discharge: (None, 0)
        death: (None, 0)
```

E

```
target:
    start: gap.end
    end: start -> discharge_or_death
    start_inclusive: False
    end_inclusive: True
    label: death
```

What do we need:

1. A large set of datasets in the MEDS format, spanning diverse clinical settings.
2. A collection of tasks that are “conceptually identical” across disparate datasets.
3. A collection of model algorithms that we can train from scratch on said datasets and tasks, then evaluate in a consistent manner.

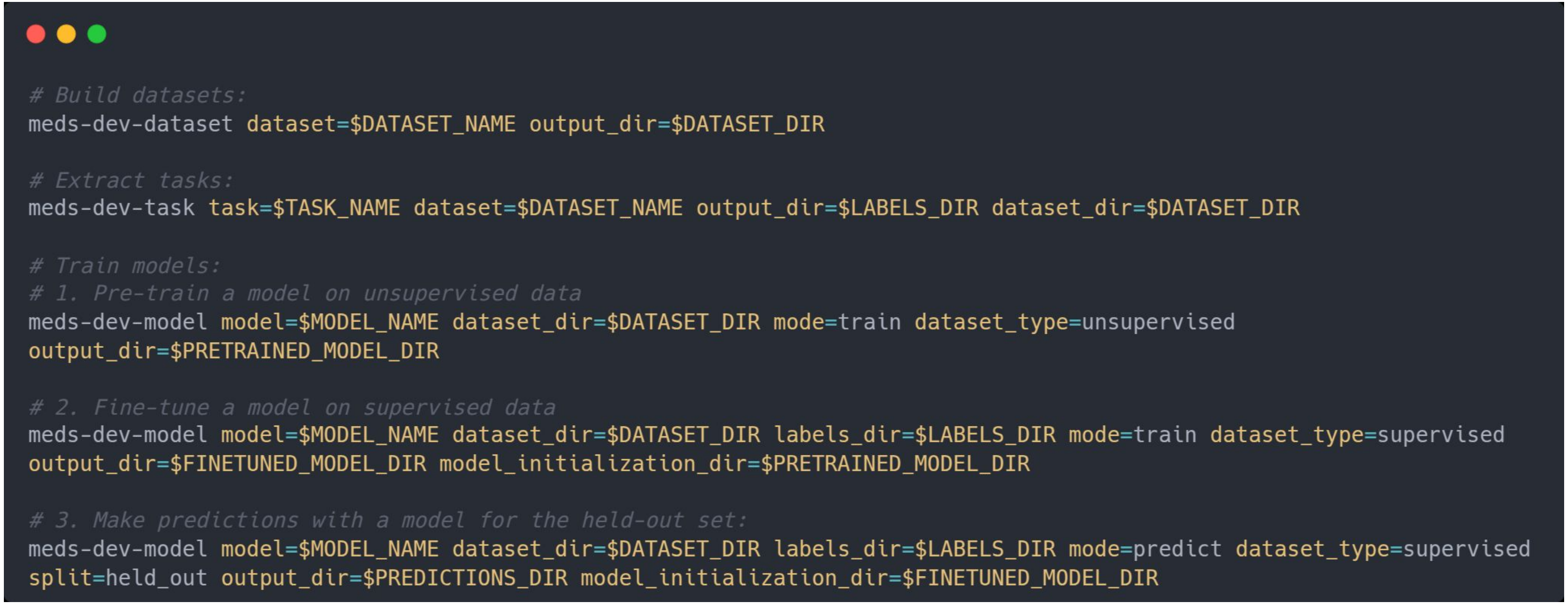
What do we need:

1. A large set of datasets in the MEDS format, spanning diverse clinical settings.
2. A collection of tasks that are “conceptually identical” across disparate datasets.
3. A collection of model algorithms that we can train from scratch on said datasets and tasks, then evaluate in a consistent manner.
4. An interface to make training these models on these tasks trivial -- *while keeping it easy to add new models, datasets, and tasks by external contributors!*

MEDS-DEV:

The MEDS Decentralized, Extensible, Validation Benchmark

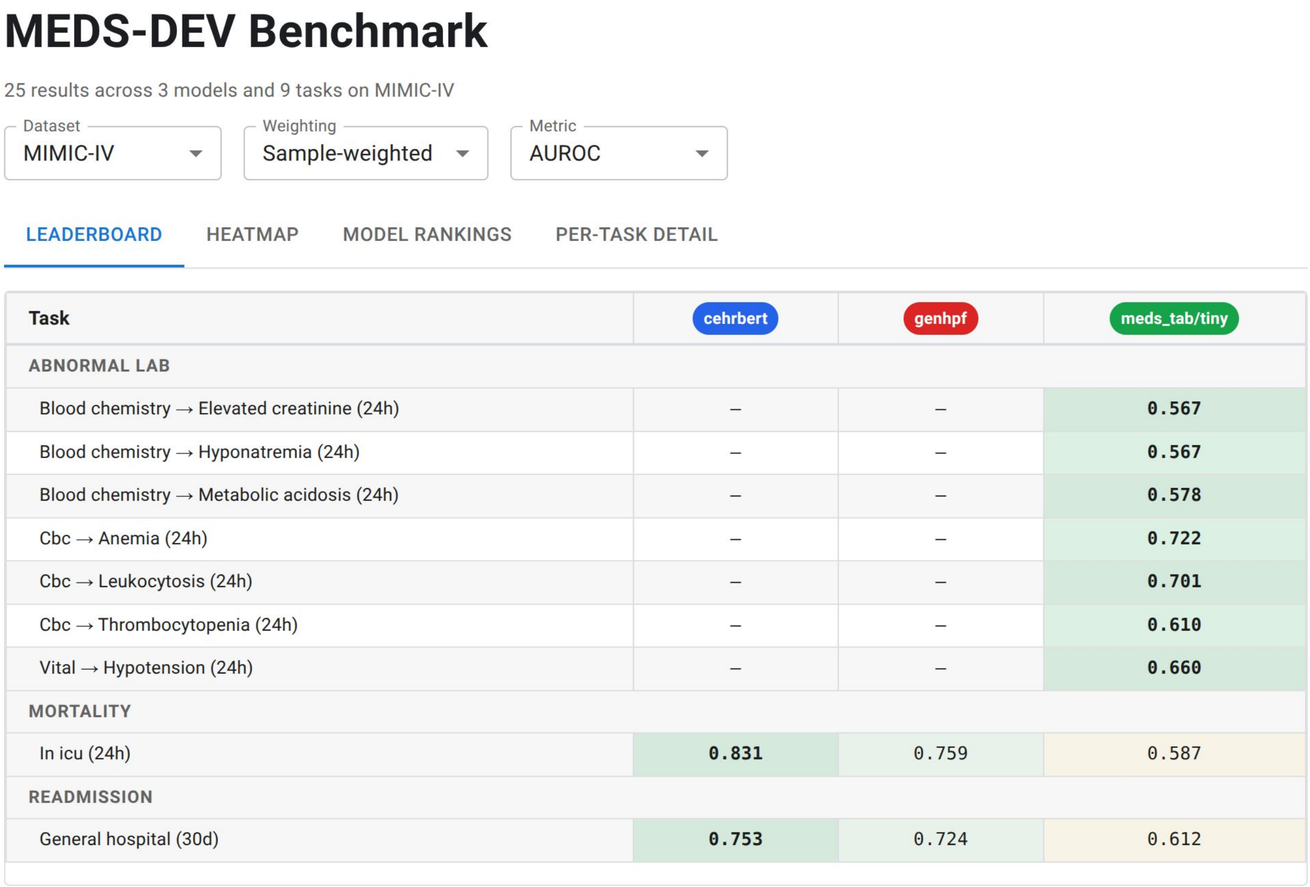
If reproducibility is made trivial, we can realize all aspects of assessing a Health AI algorithm under a simple, easy to use interface



```
# Build datasets:
meds-dev-dataset dataset=$DATASET_NAME output_dir=$DATASET_DIR

# Extract tasks:
meds-dev-task task=$TASK_NAME dataset=$DATASET_NAME output_dir=$LABELS_DIR dataset_dir=$DATASET_DIR

# Train models:
# 1. Pre-train a model on unsupervised data
meds-dev-model model=$MODEL_NAME dataset_dir=$DATASET_DIR mode=train dataset_type=unsupervised
output_dir=$PRETRAINED_MODEL_DIR

# 2. Fine-tune a model on supervised data
meds-dev-model model=$MODEL_NAME dataset_dir=$DATASET_DIR labels_dir=$LABELS_DIR mode=train dataset_type=supervised
output_dir=$FINETUNED_MODEL_DIR model_initialization_dir=$PRETRAINED_MODEL_DIR

# 3. Make predictions with a model for the held-out set:
meds-dev-model model=$MODEL_NAME dataset_dir=$DATASET_DIR labels_dir=$LABELS_DIR mode=predict dataset_type=supervised
split=held_out output_dir=$PREDICTIONS_DIR model_initialization_dir=$FINETUNED_MODEL_DIR
```

What do we need:

1. A large set of datasets in the MEDS format, spanning diverse clinical settings.
2. A collection of tasks that are “conceptually identical” across disparate datasets.
3. A collection of model algorithms that we can train from scratch on said datasets and tasks, then evaluate in a consistent manner.
4. An interface to make training these models on these tasks trivial -- *while keeping it easy to add new models, datasets, and tasks by external contributors!*
5. A pool of interested researchers to train models (both their own and those contributed by external collaborators) on their data and across diverse tasks.



MEDS-DEV Benchmark

25 results across 3 models and 9 tasks on MIMIC-IV

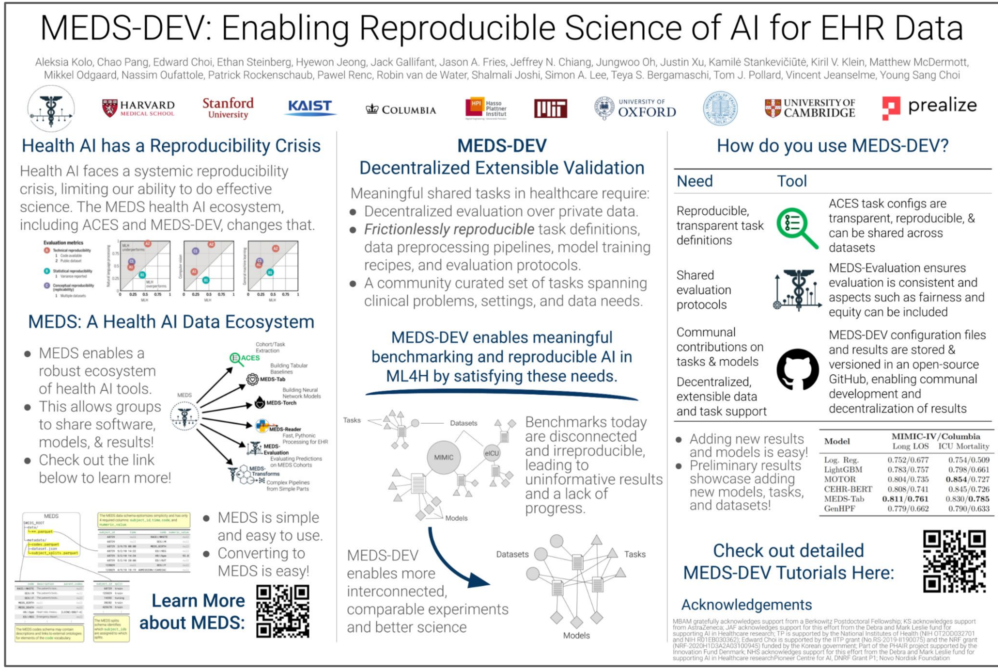
Dataset: MIMIC-IV

Weighting: Sample-weighted

Metric: AUROC

Task	cehrbert	genhpf	meds_tab/tiny
ABNORMAL LAB			
Blood chemistry → Elevated creatinine (24h)	–	–	**0.567**
Blood chemistry → Hyponatremia (24h)	–	–	**0.567**
Blood chemistry → Metabolic acidosis (24h)	–	–	**0.578**
Cbc → Anemia (24h)	–	–	**0.722**
Cbc → Leukocytosis (24h)	–	–	**0.701**
Cbc → Thrombocytopenia (24h)	–	–	**0.610**
Vital → Hypotension (24h)	–	–	**0.660**
MORTALITY			
In icu (24h)	**0.831**	0.759	0.587
READMISSION			
General hospital (30d)	**0.753**	0.724	0.612

If successful, this effort can catalyze significant advancement in our understanding of what model algorithms work for EHR data and why.



MEDS-DEV: Enabling Reproducible Science of AI for EHR Data

Aleksia Kolo, Chao Pang, Edward Choi, Ethan Steinberg, Hyewon Jeong, Jack Gallifant, Jason A. Fries, Jeffrey N. Chiang, Jungwoo Oh, Justin Xu, Kamilė Stankevičiūtė, Kiril V. Klein, Matthew McDermott, Mikkel Odgaard, Nassim Oufattole, Patrick Rockenschaub, Pawel Renc, Robin van de Water, Shalmali Joshi, Simon A. Lee, Teya S. Bergamaschi, Tom J. Pollard, Vincent Jeanselme, Young Sang Choi

Health AI has a Reproducibility Crisis

Health AI faces a systemic reproducibility crisis, limiting our ability to do effective science. The MEDS health AI ecosystem, including ACES and MEDS-DEV, changes that.

MEDS: A Health AI Data Ecosystem

- MEDS enables a robust ecosystem of health AI tools.
- This allows groups to share software, models, & results!
- Check out the link below to learn more!

- MEDS is simple and easy to use.
- Converting to MEDS is easy!

Learn More about MEDS:

MEDS-DEV
Decentralized Extensible Validation

Meaningful shared tasks in healthcare require:

- Decentralized evaluation over private data.
- *Frictionlessly reproducible* task definitions, data preprocessing pipelines, model training recipes, and evaluation protocols.
- A community curated set of tasks spanning clinical problems, settings, and data needs.

MEDS-DEV enables meaningful benchmarking and reproducible AI in ML4H by satisfying these needs.

Benchmarks today are disconnected and irreproducible, leading to uninformative results and a lack of progress.

MEDS-DEV enables more interconnected, comparable experiments and better science

How do you use MEDS-DEV?

Need	Tool	
Reproducible, transparent task definitions		ACES task configs are transparent, reproducible, & can be shared across datasets
Shared evaluation protocols		MEDS-Evaluation ensures evaluation is consistent and aspects such as fairness and equity can be included
Communal contributions on tasks & models Decentralized, extensible data and task support		MEDS-DEV configuration files and results are stored & versioned in an open-source GitHub, enabling communal development and decentralization of results

- Adding new results and models is easy!
- Preliminary results showcase adding new models, tasks, and datasets!

Model	MIMIC-IV/Columbia Long LOS	MIMIC-IV/Columbia ICU Mortality
Log. Reg.	0.752/0.677	0.754/0.509
LightGBM	0.783/0.757	0.798/0.661
MOTOR	0.804/0.735	**0.854**/0.727
CEHR-BERT	0.808/0.741	0.845/0.726
MEDS-Tab	**0.811/0.761**	0.830/**0.785**
GenHPF	0.779/0.662	0.790/0.633

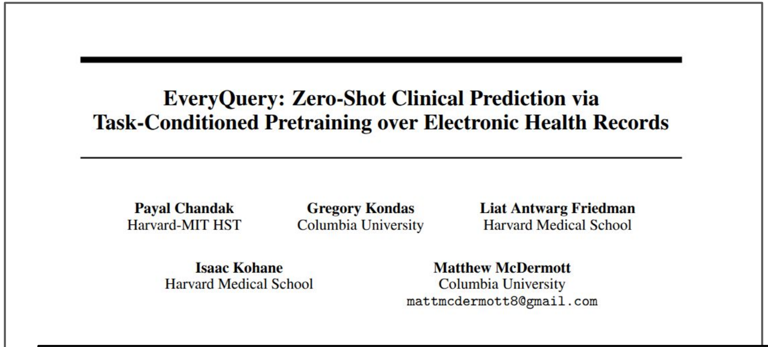
Check out detailed MEDS-DEV Tutorials Here:

Acknowledgements

MBAM gratefully acknowledges support from a Berkowitz Postdoctoral Fellowship; KS acknowledges support from AstraZeneca; JAF acknowledges support for this effort from the Debra and Mark Leslie fund for supporting AI in Healthcare research; TP is supported by the National Institutes of Health (NIH OT2OD032701 and NIH R01EB030362); Edward Choi is supported by the IITP grant (No.RS-2019-II190075) and the NRF grant (NRF-2020H1D3A2A03100945) funded by the Korean government; Part of the PHAIR project supported by the Innovation Fund Denmark; NHS acknowledges support for this effort from the Debra and Mark Leslie fund for supporting AI in Healthcare researchPioneer Centre for AI, DNRF Grant P1; Novo Nordisk Foundation

Research Direction 2:
Learning the Latent Vocabulary of Medicine

EHR foundation models are here:



EveryQuery: Zero-Shot Clinical Prediction via Task-Conditioned Pretraining over Electronic Health Records

Payal Chandak
Harvard-MIT HST

Gregory Kondas
Columbia University

Liat Antwarg Friedman
Harvard Medical School

Isaac Kohane
Harvard Medical School

Matthew McDermott
Columbia University
mattmcdermott8@gmail.com

Generative Medical Event Models Improve with Scale

Shane Waxler*[1], Paul Blazek*[1], Davis White*[1], Daniel Sneider*[1],
Kevin Chung[1], Mani Nagarathnam[1], Patrick Williams[1], Hank Voeller[1], Karen Wong[1],
Matthew Swanhorst[1], Sheng Zhang[2], Naoto Usuyama[2], Cliff Wong[2], Tristan Naumann[2],
Hoifung Poon[2], Andrew Loza[3], Daniella Meeker[3, 4], Seth Hain[1], and Rahul Shah†[1]

[1]Epic Systems
[2]Microsoft Research
[3]Yale School of Medicine
[4]Cosmos Governing Council

Abstract

Realizing personalized medicine at scale calls for methods that distill insights from longitudinal patient journeys, which can be viewed as a sequence of medical events. Foundation models pretrained on large-scale medical event data represent a promising direction for scaling real-world evidence generation and generalizing to diverse downstream tasks. Using Epic Cosmos, a dataset with medical events from de-identified longitudinal health records for 16.3 billion encounters over 300 million unique patient records from 310 health systems, we introduce the Comet models, a family of decoder-only transformer models pretrained on 118 million patients representing 115 billion discrete medical events (151 billion tokens). We present the largest scaling-law study of medical event data, establishing a methodology for pretraining and revealing power-law scaling relationships for compute, tokens, and model size. Consequently, we pretrained a series of compute-optimal models with up to 1 billion parameters. Conditioned on a patient's real-world history, Comet autoregressively predicts the next medical event to simulate patient health timelines. We studied 78 real-world tasks, including diagnosis prediction, disease prognosis, and healthcare operations. Remarkably for a foundation model with generic pretraining and simulation-based inference, Comet

npj | digital medicine

Published in partnership with Seoul National University Bundang Hospital

Article

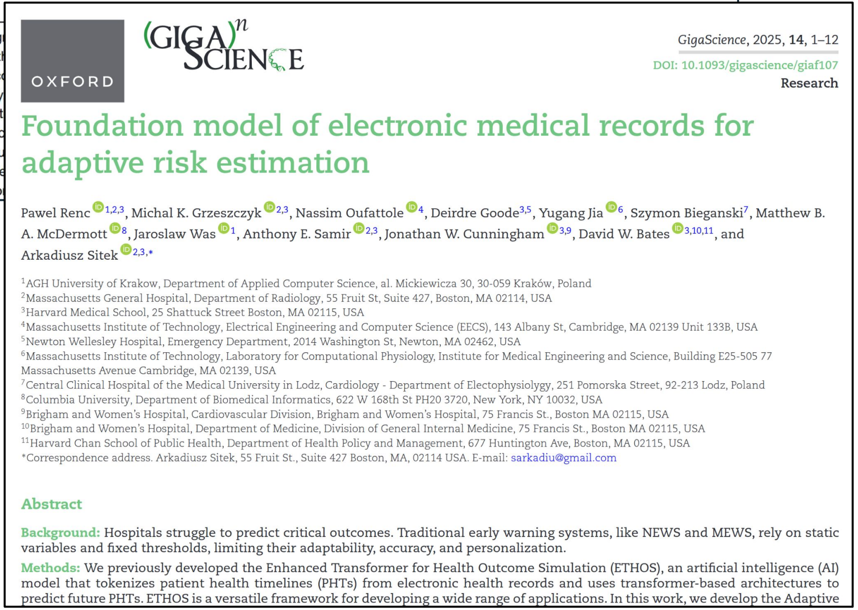
https://doi.org/10.1038/s41746-024-01235-0

Zero shot health trajectory prediction using transformer

Check for updates

Pawel Renc [1,2,3], **Yugang Jia** [4], **Anthony E. Samir** [1,2], **Jaroslaw Was**[3], **Quanzheng Li**[1,2],
David W. Bates [2,5,6] & **Arkadiusz Sitek** [1,2]

Integ
healt
Outc
analy
Healt
trajec
in fou
mode
facto



OXFORD

(GIGA)^n SCIENCE

GigaScience, 2025, **14**, 1–12
DOI: 10.1093/gigascience/giaf107

Research

Foundation model of electronic medical records for adaptive risk estimation

Pawel Renc [1,2,3], Michal K. Grzeszczyk [2,3], Nassim Oufattole [4], Deirdre Goode[3,5], Yugang Jia [6], Szymon Bieganski[7], Matthew B. A. McDermott [8], Jaroslaw Was [1], Anthony E. Samir [2,3], Jonathan W. Cunningham [3,9], David W. Bates [3,10,11], and Arkadiusz Sitek [2,3,*]

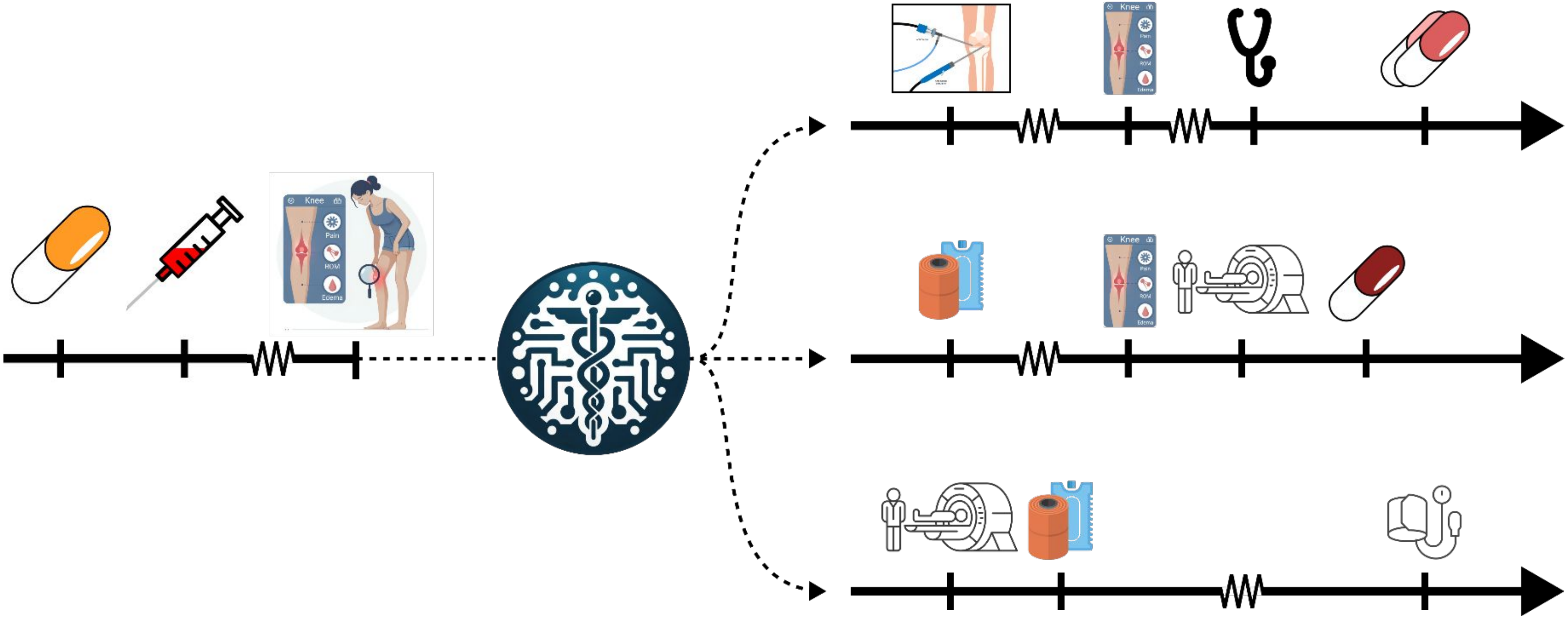
[1]AGH University of Krakow, Department of Applied Computer Science, al. Mickiewicza 30, 30-059 Kraków, Poland
[2]Massachusetts General Hospital, Department of Radiology, 55 Fruit St, Suite 427, Boston, MA 02114, USA
[3]Harvard Medical School, 25 Shattuck Street Boston, MA 02115, USA
[4]Massachusetts Institute of Technology, Electrical Engineering and Computer Science (EECS), 143 Albany St, Cambridge, MA 02139 Unit 133B, USA
[5]Newton Wellesley Hospital, Emergency Department, 2014 Washington St, Newton, MA 02462, USA
[6]Massachusetts Institute of Technology, Laboratory for Computational Physiology, Institute for Medical Engineering and Science, Building E25-505 77 Massachusetts Avenue Cambridge, MA 02139, USA
[7]Central Clinical Hospital of the Medical University in Lodz, Cardiology - Department of Electophysiolygy, 251 Pomorska Street, 92-213 Lodz, Poland
[8]Columbia University, Department of Biomedical Informatics, 622 W 168th St PH20 3720, New York, NY 10032, USA
[9]Brigham and Women's Hospital, Cardiovascular Division, Brigham and Women's Hospital, 75 Francis St., Boston MA 02115, USA
[10]Brigham and Women's Hospital, Department of Medicine, Division of General Internal Medicine, 75 Francis St., Boston MA 02115, USA
[11]Harvard Chan School of Public Health, Department of Health Policy and Management, 677 Huntington Ave, Boston, MA 02115, USA
*Correspondence address. Arkadiusz Sitek, 55 Fruit St., Suite 427 Boston, MA, 02114 USA. E-mail: sarkadiu@gmail.com

Abstract

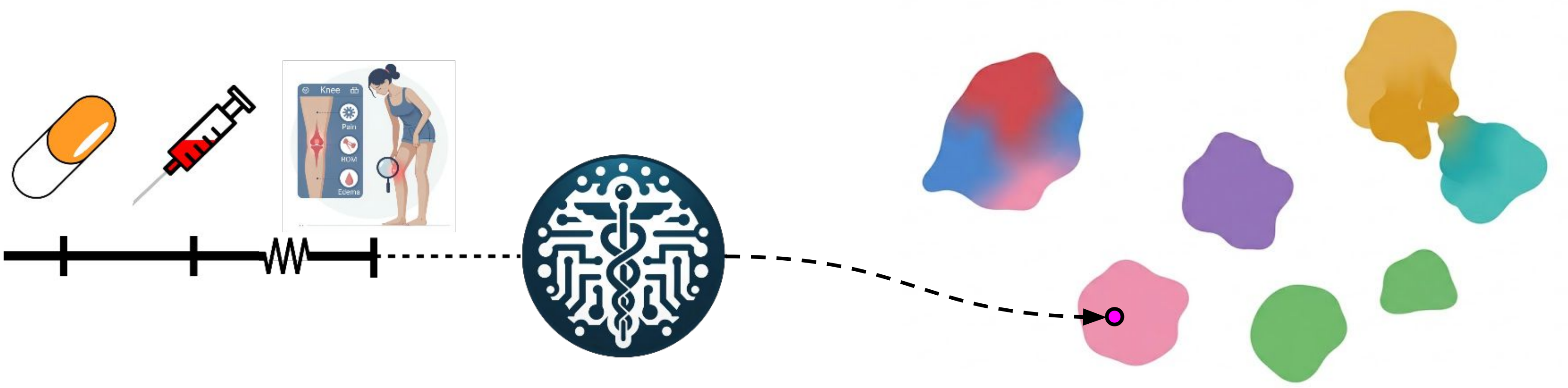
Background: Hospitals struggle to predict critical outcomes. Traditional early warning systems, like NEWS and MEWS, rely on static variables and fixed thresholds, limiting their adaptability, accuracy, and personalization.

Methods: We previously developed the Enhanced Transformer for Health Outcome Simulation (ETHOS), an artificial intelligence (AI) model that tokenizes patient health timelines (PHTs) from electronic health records and uses transformer-based architectures to predict future PHTs. ETHOS is a versatile framework for developing a wide range of applications. In this work, we develop the Adaptive

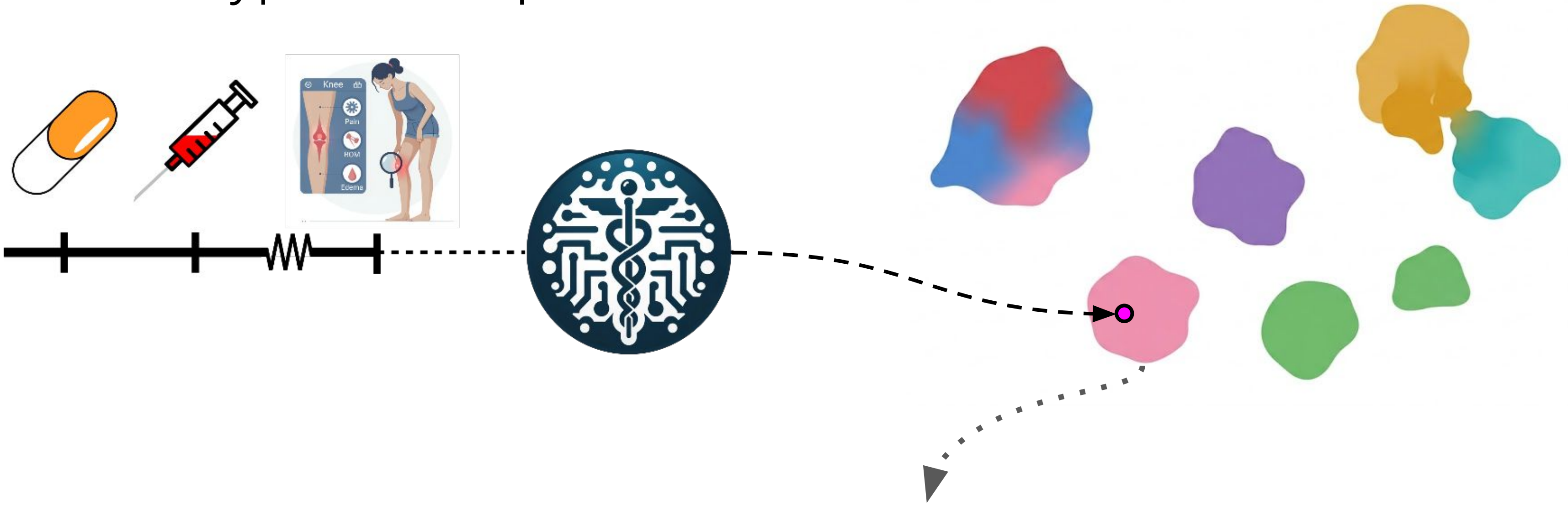
These models (usually) take a patient's input, and simulate "possible futures"



What if, instead, they mapped to a “latent phenotype”?



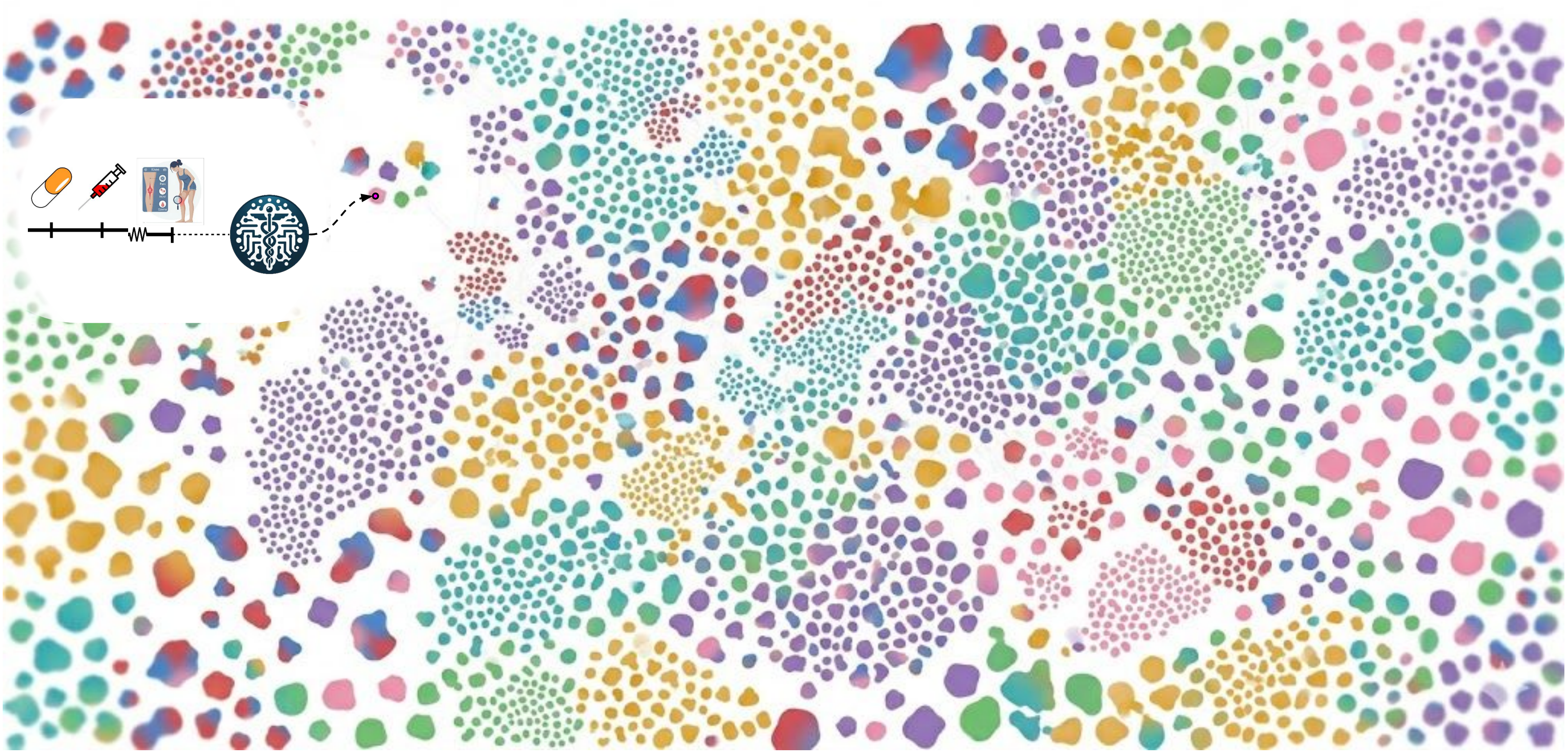
Phenotype driven predictions:



Historically, ● patients:

- Have high risk of autoimmune disease
- Have a history of pediatric infections
- Are likely to respond to drug X

… And I want a *lot* of latent phenotypes



The simple things that I want:

- Not 5-100 phenotypes, but *10 - 50 thousand*.

The simple things that I want:

- Not 5-100 phenotypes, but *10 - 50 thousand*.
- A subset of phenotypes to meaningful separate out patient populations at increased risk for clinically meaningful diseases, adverse events, or drug suitability.

The simple things that I want:

- Not 5-100 phenotypes, but *10 - 50 thousand*.
- A subset of phenotypes to meaningful separate out patient populations at increased risk for clinically meaningful diseases, adverse events, or drug suitability.
- ... And for a subset of those phenotypes to have the property that membership can be identified with moderate to high fidelity via lower-capacity models

The simple things that I want:

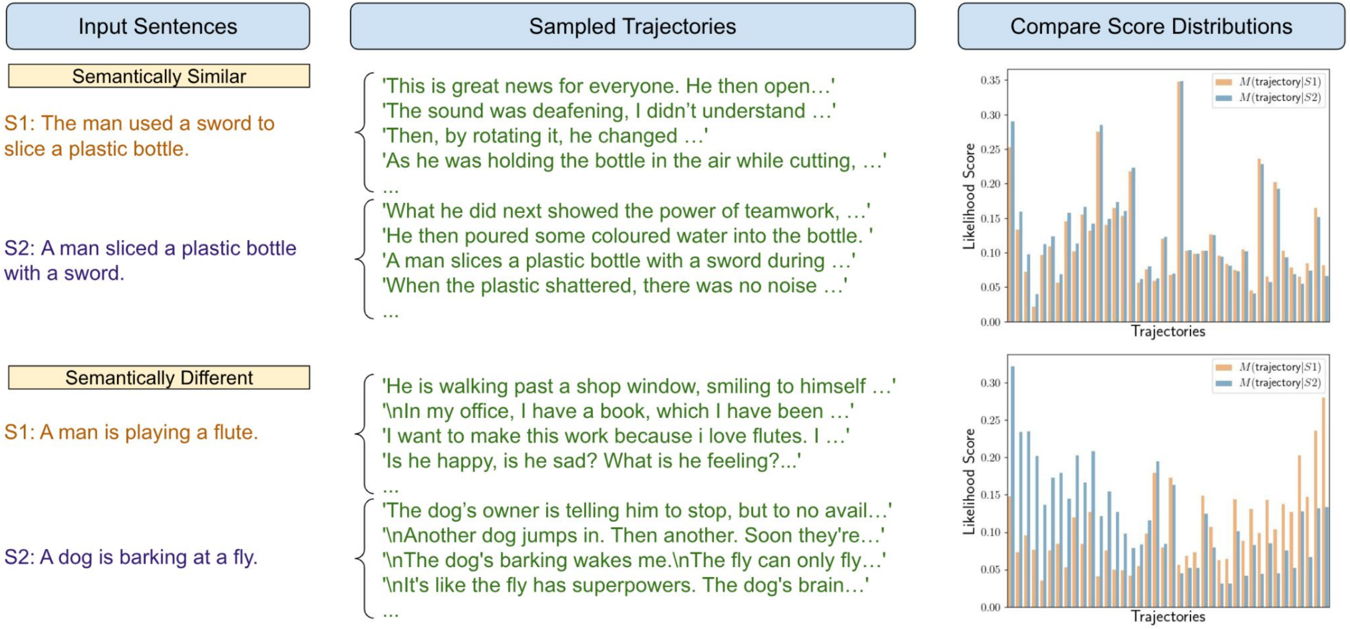
- Not 5-100 phenotypes, but *10 - 50 thousand*.
- A subset of phenotypes to meaningful separate out patient populations at increased risk for clinically meaningful diseases, adverse events, or drug suitability.
- ... And for a subset of those phenotypes to have the property that membership can be identified with moderate to high fidelity via lower-capacity models
- ... and for a subset of those lower-capacity models to be generalizable across sites.

Why should this be possible?

Why wouldn’t this have been done already?

If EHR foundation models can predict clinically meaningful outcomes in a zero-shot manner, then this information exists somewhere in the model.
All we need to do is extract it.

New Clustering Methods



Input Sentences

Semantically Similar

S1: The man used a sword to slice a plastic bottle.

S2: A man sliced a plastic bottle with a sword.

Semantically Different

S1: A man is playing a flute.

S2: A dog is barking at a fly.

Sampled Trajectories

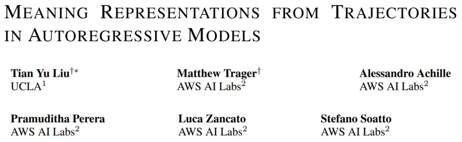
'This is great news for everyone. He then open…'
'The sound was deafening, I didn't understand …'
'Then, by rotating it, he changed …'
'As he was holding the bottle in the air while cutting, …'
…

'What he did next showed the power of teamwork, …'
'He then poured some coloured water into the bottle. '
'A man slices a plastic bottle with a sword during …'
'When the plastic shattered, there was no noise …'
…

'He is walking past a shop window, smiling to himself …'
'\nIn my office, I have a book, which I have been …'
'I want to make this work because i love flutes. I …'
'Is he happy, is he sad? What is he feeling?...'
…

'The dog's owner is telling him to stop, but to no avail…'
'\nAnother dog jumps in. Then another. Soon they're…'
'\nThe dog's barking wakes me.\nThe fly can only fly…'
'\nIt's like the fly has superpowers. The dog's brain…'
…

Compare Score Distributions



MEANING REPRESENTATIONS FROM TRAJECTORIES IN AUTOREGRESSIVE MODELS

Tian Yu Liu[†*]
UCLA[1]

Matthew Trager[†]
AWS AI Labs[2]

Alessandro Achille
AWS AI Labs[2]

Pramuditha Perera
AWS AI Labs[2]

Luca Zancato
AWS AI Labs[2]

Stefano Soatto
AWS AI Labs[2]

Emergent Clusters with non-autoregressive Foundation Models

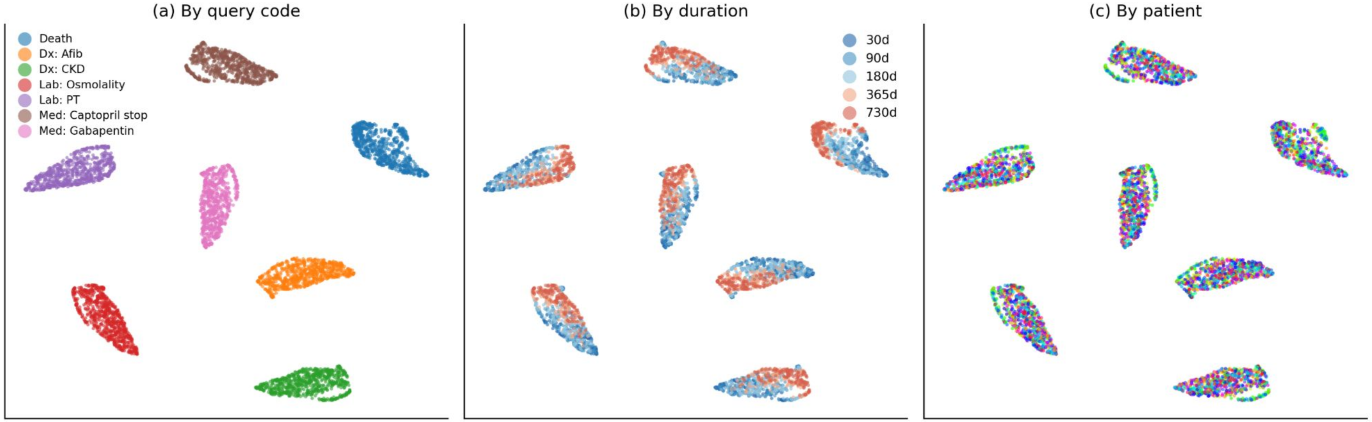
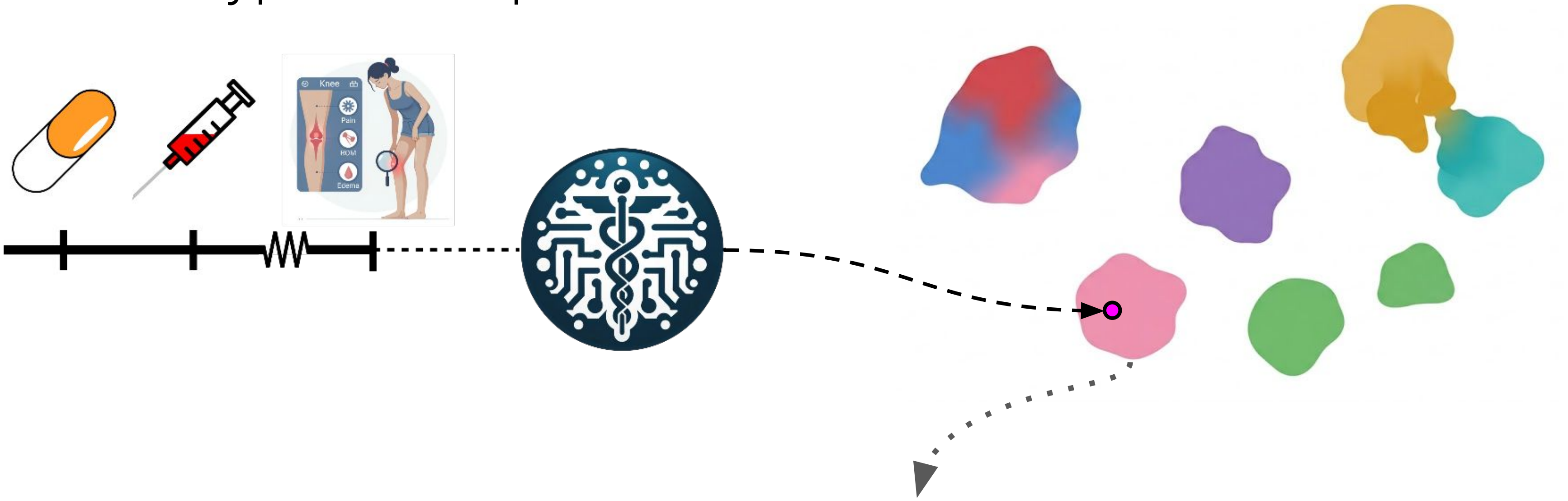


Figure 4: **EveryQuery representations are organized by task.** UMAP projection of embeddings, colored by query code (left), duration horizon (middle), and patient (right). Code identity produces distinct clusters; horizon is sequentially ordered in time within each cluster; patient identity produces less visible clustering but retains a deep structure, as discussed in Section 5.3.

What we need:

1. Refinements for how to make meaningful clusters over EHR data.
2. A better way to interrogate and understand latent clusters.
3. Stable, efficient cluster determination across large EHR datasets.
4. Assessment of cluster property consistency across diverse datasets.

Phenotype driven predictions



Historically, ● patients:

- Have high risk of autoimmune disease
- Have a history of pediatric infections
- Are likely to respond to drug X

Acknowledgements

Edward Choi
Ethan Steinberg
Jason A. Fries
Jungwoo Oh
Matthew McDermott

Michael Wornow
Nigam H. Shah
Patrick Rockenschaub
Robin P. van de Water
Tom J. Pollard

Aleksia Kolo
Chao Pang
Edward Choi
Ethan Steinberg
Hyewon Jeong
Jack Gallifant
Jason A. Fries
Jeffrey N. Chiang

Jungwoo Oh
Justin Xu
Kamilė Stankevičiūtė
Kiril V. Klein
Matthew McDermott
Mikkel Odgaard
Nassim Oufattole
Patrick Rockenschaub
Pawel Renc

Robin van de Water
Shalmali Joshi
Simon A. Lee
Teya S. Bergamaschi
Tom J. Pollard
Vincent Jeanselme
Young Sang Choi

HARVARD MEDICAL SCHOOL
Stanford University
KAIST AI Kim Jaechul Graduate School
COLUMBIA
HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam
MIT
Cornell University
University of Oxford
UCLA
University of Cambridge
prealize

MEDS Publications

McDermott, M. B. A., Steinberg, E., Fries, J. A., van de Water, R. P., Pang, C., Rockenschaub, P., Renc, P., Oh, J., Stankevičiūtė, K., Xu, J., Pollard, T. J., Oufattole, N., Wornow, M., Bergamaschi, T. S., Jeong, H., Lee, S. A., Jeanselme, V., Klein, K. V., Odgaard, M., … Shah, N. H. (2026). MEDS — An emerging data standard and ecosystem for health AI research. NEJM AI (in press).

MEDS Working Group: Bert Arnrich, Edward Choi, Jason A. Fries, Matthew B. A. McDermott, Jungwoo Oh, Tom J. Pollard, Nigam Shah, Ethan Steinberg, Michael Wornow, Robin van de Water. 2024. “Medical Event Data Standard (MEDS): Facilitating Machine Learning for Health.” In ICLR 2024 Workshop TS4H. https://openreview.net/forum?id=IsHy2ebjIG&referrer=%5BAuthor%20Console%5D(%2Fgroup%3Fid%3DICLR.cc%2F2024%2FWorkshop%2FTS4H%2FAuthors%23your-submissions).

Oufattole, Nassim, Teya Bergamaschi, Aleksia Kolo, Hyewon Jeong, Hanna Gaggin, Collin M. Stultz, and Matthew B. A. McDermott. 2024. “MEDS-Tab: Automated Tabularization and Baseline Methods for MEDS Datasets.” *arXiv [Cs.LG]*. arXiv. http://arxiv.org/abs/2411.00200.

Steinberg, E., Michael Wornow, Suhana Bedi, J. Fries, Matthew B. A. McDermott, and Nigam H. Shah. 2024. “Meds_reader: A Fast and Efficient EHR Processing Library.” In *Machine Learning for Health Symposium (Findings Track)*. Vol. abs/2409.09095. https://doi.org/10.48550/arXiv.2409.09095.

Xu, Justin, Jack Gallifant, Alistair E. W. Johnson, and Matthew B. A. McDermott. 2025. “ACES: Automatic Cohort Extraction System for Event-Stream Datasets.” In *Proceedings of the International Conference on Learning Representations*.