Clauirton Siebra

AI in Action

Real-World Applications in the Brazilian Public Health System

SUS
Sistema Único de Saúde

National Health Service (NHS, 1948)

Free Health support for everyone (1988)

Government is trying to promote advances in the public health services, mainly via the use of AI

- Development rather than research (majority)
- Apply the state of the art
- Create products/services rather than a paper
- No costs

Projects

1. AI-Based Automated EHR Documentation from Doctor–Patient Conversations
2. Early prediction of clinical events
3. Predictive models to identify patients at increased risk of cardiac events
4. LLM-based virtual assistant to support health staff

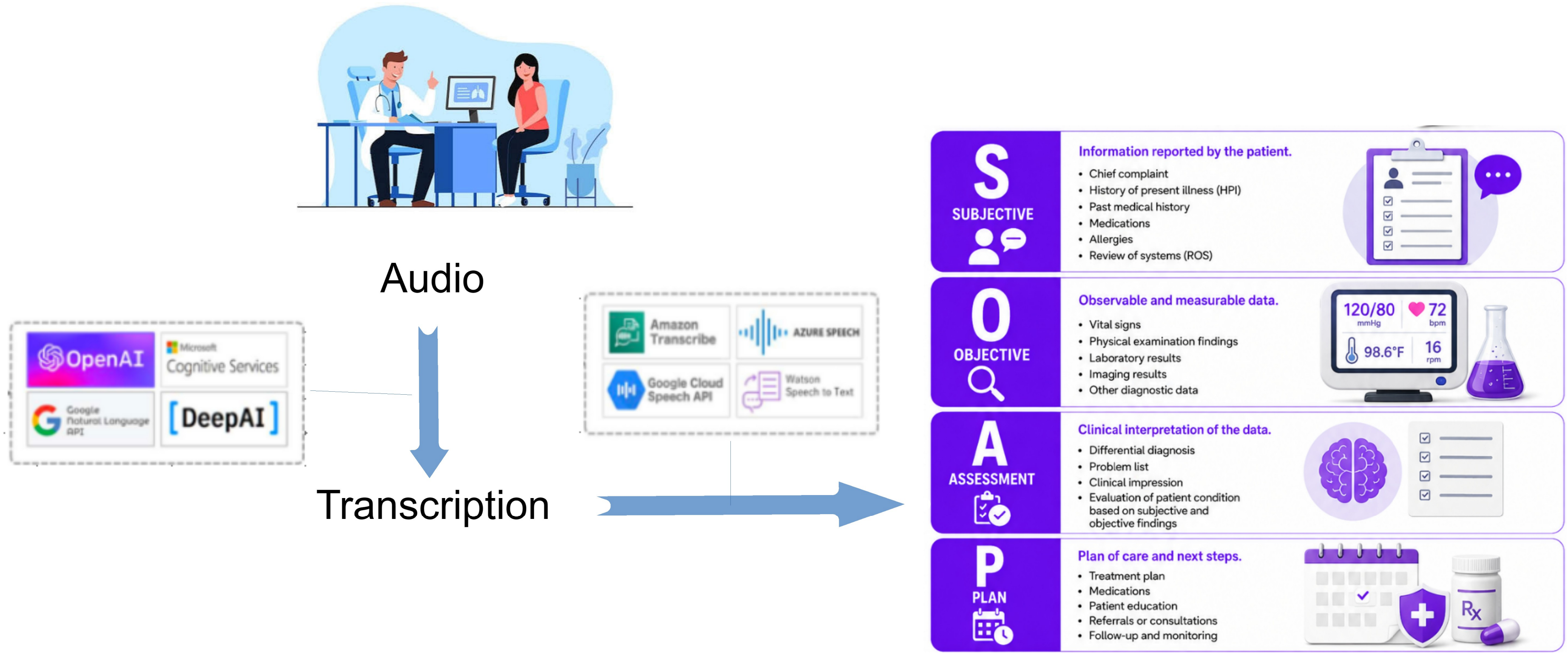
Project title

AI-Based Automated EHR Documentation from Doctor–Patient Conversations

Problem

- Studies show that doctors spend about 30% to 40% of their time on data entry (Brazil)
- Not only in Brazil
  - Pinevich Y, Clark KJ, Harrison AM, Pickering BW, Herasevich V. Interaction Time with Electronic Health Records: A Systematic Review. Appl Clin Inform. 2021 Aug;12(4):788-799. doi: 10.1055/s-0041-1733909. Epub 2021 Aug 25. PMID: 34433218; PMCID: PMC8387128.

Solution


Audio
OpenAI
Microsoft Cognitive Services
Google Natural Language API
[DeepAI]
Amazon Transcribe
AZURE SPEECH
Google Cloud Speech API
Watson Speech to Text
Transcription
S
SUBJECTIVE
Information reported by the patient.
• Chief complaint
• History of present illness (HPI)
• Past medical history
• Medications
• Allergies
• Review of systems (ROS)
O
OBJECTIVE
Observable and measurable data.
• Vital signs
• Physical examination findings
• Laboratory results
• Imaging results
• Other diagnostic data
120/80 mmHg
72 bpm
98.6°F
16 rpm
A
ASSESSMENT
Clinical interpretation of the data.
• Differential diagnosis
• Problem list
• Clinical impression
• Evaluation of patient condition based on subjective and objective findings
P
PLAN
Plan of care and next steps.
• Treatment plan
• Medications
• Patient education
• Referrals or consultations
• Follow-up and monitoring
Rx

New Problems

- Privacy and data protection rules
- 2.3 billions of clinical consultations (2023)
  - Cost around 10 billions BRL (1.56 billion CHF)

New solution

CREATE our new specialized model

LLM + prompt engineering

Fine-tunning

LLM + prompt engineering

Transcription → SOAP

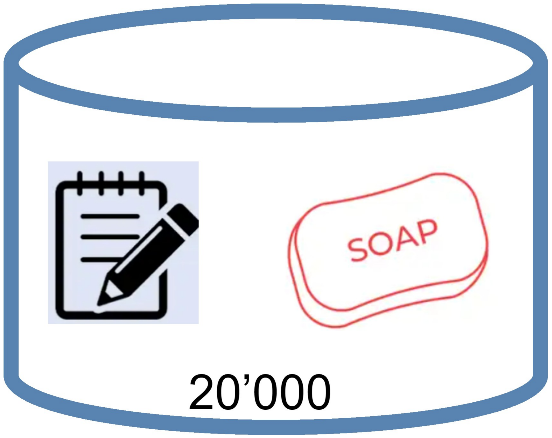
	gpt-oss-120b	llama-3.3-70b-instruct	mistral-small-4-119b-2603	deepseek-v3.2	mistral-7b-instruct-v0.3
CSE	0.73	0.73	0.75	0.69	0.75
CCI	0.84	0.56	0.75	0.79	0.88
IAC	0.92	0.87	0.88	0.87	0.96

Was every classified information correctly classified?

Was every important information classified (coverage)?

Does the model hallucinate?

Fine-tuning



Real data: app for data collection

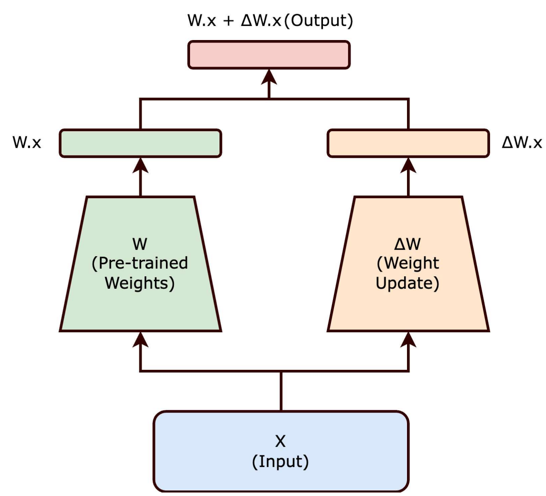
Partially synthetic data: transcription generated from real SOAP

qwen3-8b-instruct

Alibaba cloud

- Very good for Portuguese
- Already fine-tuned to accept instructions
- Small

Fine-tuning



https://medium.com/@zilliz_learn/lora-explained-low-rank-adaptation-for-fine-tuning-llms-066c9bdd0b32

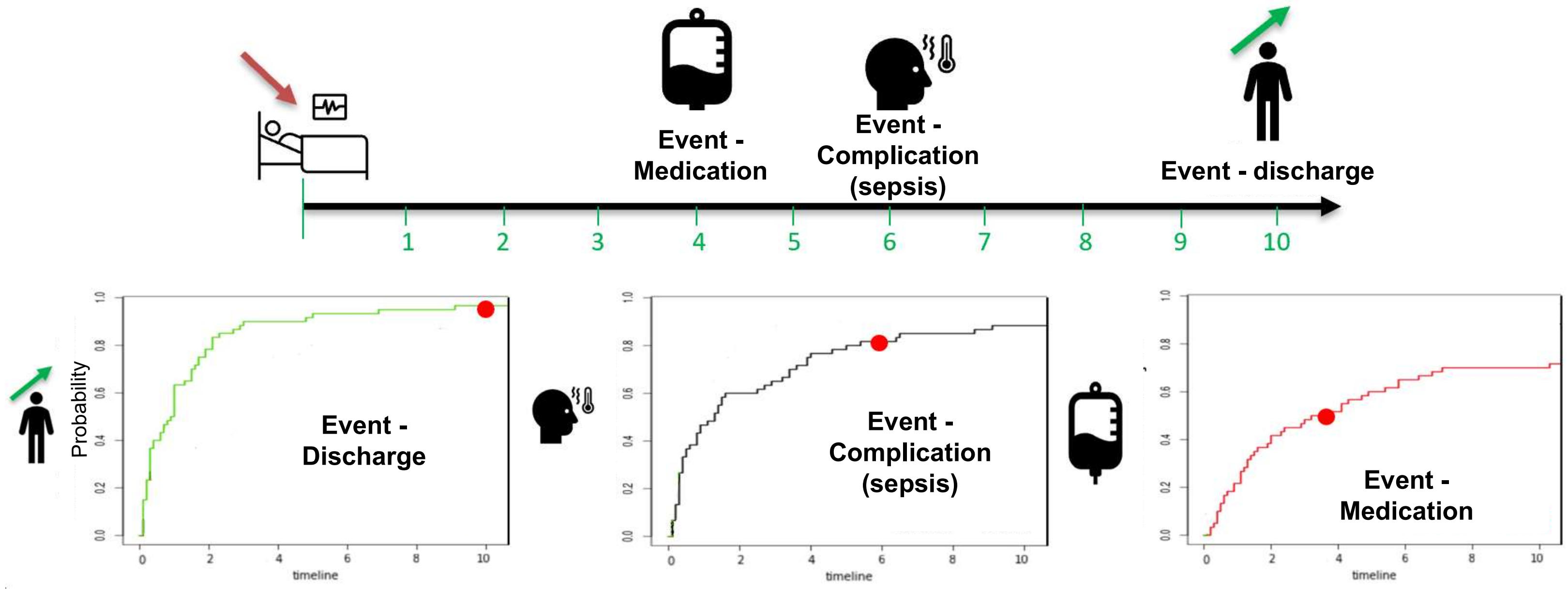
```
from peft import LoraConfig, get_peft_model
from transformers import Trainer, TrainingArguments

peft_config = LoraConfig(
    r=16,
    lora_alpha=32,
    target_modules=["q_proj", "v_proj"],
    lora_dropout=0.05,
    task_type="CAUSAL_LM"
)

model = get_peft_model(model, peft_config)

training_args = TrainingArguments(
    output_dir="./molmo-soap",
    per_device_train_batch_size=2,
    gradient_accumulation_steps=8,
    num_train_epochs=3,
    learning_rate=2e-4,
    fp16=True,
    logging_steps=50,
    save_steps=500
)

trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=train_dataset
)

trainer.train()
```

Project title

Early prediction of clinical events

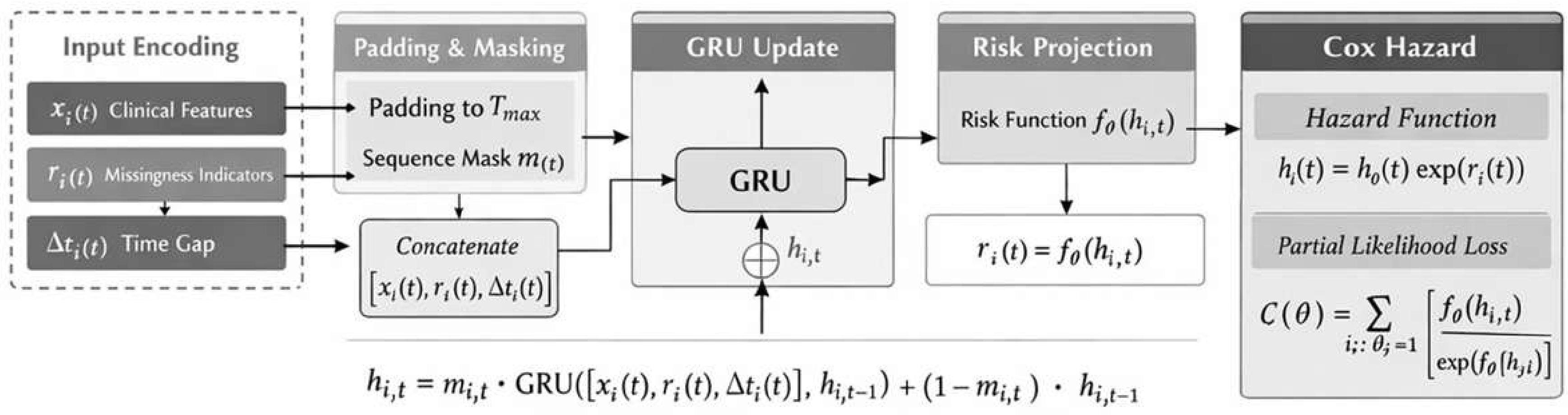
Problem

Identify as soon as possible important clinical events

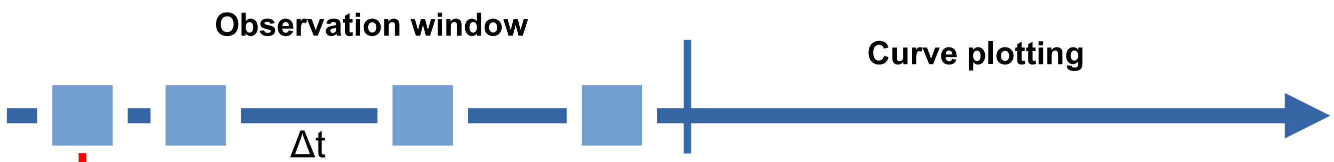


Solution – Survival analysis (Cox Hazard Function)

- However, it does not consider the evolution of attributes
- Integration with a Recurrent Neural Network
  - *GRU - Gated Recurrent Unit*



Solution - Details



X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_1	X_{12}	**attributes**
0	1	1	1	0	1	1	1	1	1	0	1	**Missingness indicator**

Results

Ablation test method

Model / Hospital	M1 (static)	M2 (pure dynamic)	M3 (dynamic + masked)	M4. (dynamic + masked + Δt)	M5 (dynamic + Δt)
H1	0.700 ± 0.015	0.847 ± 0,017	0.823 ± 0.016	0.901 ± 0.014	**0.903 ± 0.014**
H2	0.697 ± 0.046	0.858 ± 0.011	0.831 ± 0.024	0.873 ± 0.020	**0.887 ± 0.017**
H3	0.700 ± 0.012	0.779 ± 0.007	0.769 ± 0.009	0.795 ± 0.007	**0.802 ± 0.007**
H4	0.628 ± 0.023	0.741 ± 0.010	0.707 ± 0.018	0.769 ± 0.022	**0.800 ± 0.014**
H5	0.706 ± 0.013	0.782 ± 0.009	0.771 ± 0.011	0.786 ± 0.011	**0.796 ± 0.012**

Models are not the same for every hospital

Project title

LLM-based virtual tutors to support the health staff

Problem

- The analysis of a patient EHR in a time-demanding activity
- Talk to the EHR is better than search for information
  - We could ask:

Patient history: [history up to today]
Question: What diagnosis is most likely?

Patient history: [history up to today]
Question: What treatment would be appropriate to consider?

Patient history: [history up to today]
Question: What important clinical information is missing?

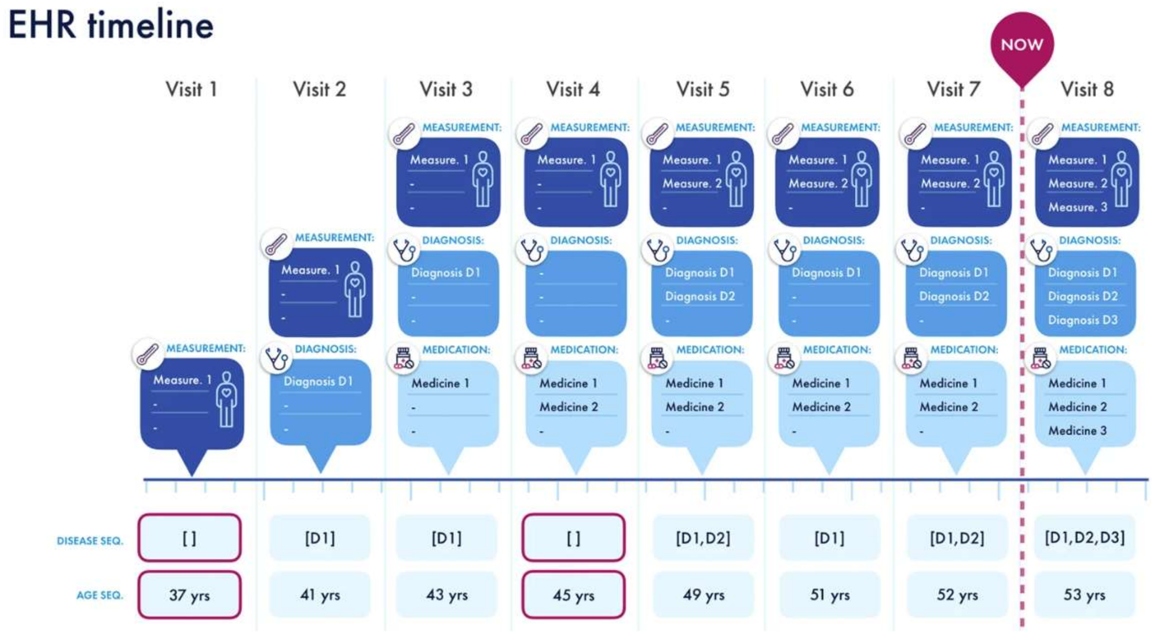
Model must understand the domain

Current approaches

BEHRT: Transformer for Electronic Health Records

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi & Gholamreza Salimi-Khorshidi

Scientific Reports **10**, Article number: 7155 (2020) | Cite this article

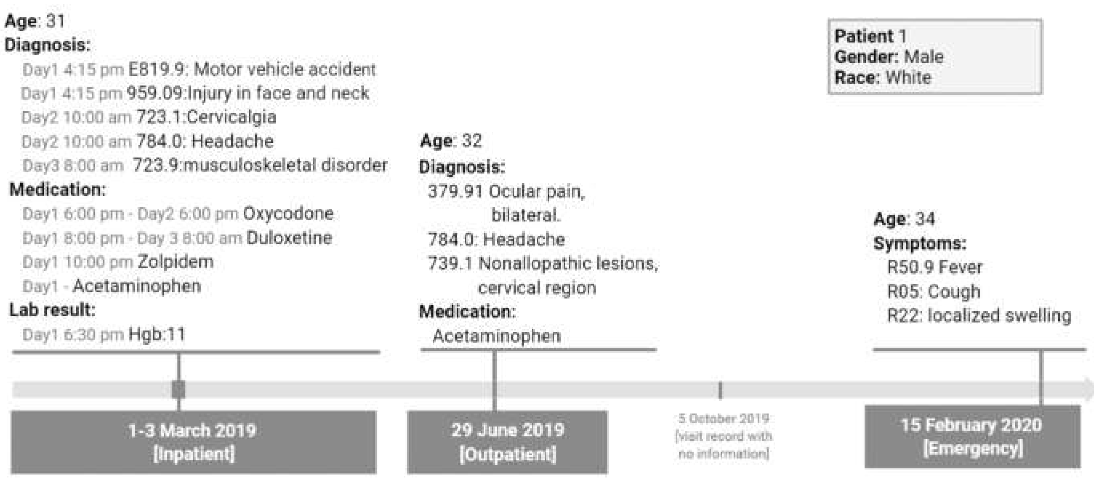


Current approaches

Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao & Degui Zhi

npj Digital Medicine **4**, Article number: 86 (2021) | Cite this article



Different approach

Raw EHR → **Patient timeline** → **Clinical narrative** → **Model fine-tuning** → **Prediction/ recommendation**

This pipeline gives the model a temporal and semantic representation of the patient's history.

2021-03-14: HbA1c 7.8%
2021-04-02: Metformin prescribed
2021-09-20: HbA1c 8.4%
2022-01-15: Metformin dose increased

"The patient has a history of type 2 diabetes diagnosed in 2021. Initial treatment with metformin resulted in limited glycemic control. HbA1c subsequently increased, leading to a dose escalation..."

Different approach

Raw EHR → **Patient timeline** → **Clinical narrative** → **Model fine-tuning** → **Prediction/ recommendation**

Normalize the EHR into clinical events/episodes (SQL functions)

```
Patient
|
├── 2021-01-10
|   ├── Diagnosis: Type 2 diabetes
|   ├── HbA1c: 8.2%
|   └── Medication: Metformin 500 mg
|
├── 2021-04-15
|   ├── HbA1c: 7.5%
|   └── Medication: Metformin 1000 mg
|
├── 2022-02-03
|   ├── HbA1c: 9.1%
|   └── Symptom: fatigue
|
└── 2022-02-20
    └── Medication: ...
```

Use of Template

Patient background:
...
Active conditions:
...
Clinical course:
...
Medications and changes:
...
Laboratory trends:
...
Investigations:
...
Procedures/hospitalizations:
...
Recent developments:
...
Current clinical state:
...

Every statement should be traceable to the EHR

Example

"The patient developed worsening renal function during 2023."

Why?

Creatinine:

2023-01-10 → 1.1

2023-06-14 → 1.5

2023-11-20 → 1.8

What are we doing now?

Approach	Input	Task
Baseline	Structured EHR	Diagnosis prediction
Baseline LLM	Raw clinical notes	Diagnosis prediction
Narrative model	Generated patient history	Diagnosis prediction
Hybrid	Structured EHR + narrative	Diagnosis prediction

Data Descriptor | Open access | Published: 03 January 2023

MIMIC-IV, a freely accessible electronic health record dataset

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

Scientific Data **10**, Article number: 1 (2023) | Cite this article

364,627 unique patients

Thank you!