BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Berkeley
UNIVERSITY OF CALIFORNIA
Electrical Engineering
and Computer Sciences

Computational
PRECISION HEALTH

Evaluating AI in Healthcare with End-User Feedback

Irene Y. Chen

Assistant Professor, Computational Precision Health
Electrical Engineering and Computer Science
Berkeley AI Research
UC Berkeley and UCSF

irenechen.net

@irenetrampoline

A 59-year-old overweight woman presents to the urgent care clinic with the complaint of severe abdominal pain for the past 2 hours. She also complains of a dull pain in her back with nausea and vomiting several times. Her pain has no relation with food. Her past medical history is significant for recurrent abdominal pain due to cholelithiasis. Her father died at the age of 60 with some form of abdominal cancer. Her temperature is 37°C (98.6°F), respirations are 15/min, pulse is 67/min, and blood pressure is 122/98 mm Hg. Physical exam is unremarkable. However, a CT scan of the abdomen shows a calcified mass near her gallbladder. Which of the following diagnoses should be excluded first in this patient?

A. Acute cholecystitis
B. Gallbladder cancer
C. Choledocholithiasis
D. Pancreatitis
E. Duodenal peptic ulcer

https://www.researchgate.net/figure/A-MedQA-dataset-sample-validates-the-correctness-of-our-models-answer-In-the-MedQA_fig5_394534615

A 59-year-old overweight woman presents to the urgent care clinic with the complaint of severe abdominal pain for the past 2 hours. She also complains of a dull pain in her back with nausea and vomiting several times. Her pain has no relation with food. Her past medical history is significant for recurrent abdominal pain due to cholelithiasis. Her father died at the age of 60 with some form of abdominal cancer. Her temperature is 37°C (98.6°F), respirations are 15/min, pulse is 67/min, and blood pressure is 122/98 mm Hg. Physical exam is unremarkable. However, a CT scan of the abdomen shows a calcified mass near her gallbladder. Which of the following diagnoses should be excluded first in this patient?

A. Acute cholecystitis
B. Gallbladder cancer
C. Choledocholithiasis
D. Pancreatitis
E. Duodenal peptic ulcer

https://www.researchgate.net/figure/A-MedQA-dataset-sample-validates-the-correctness-of-our-models-answer-In-the-MedQA_fig5_394534615

A 59-year-old overweight woman presents to the urgent care clinic with the complaint of severe abdominal pain for the past 2 hours. She also complains of a dull pain in her back with nausea and vomiting several times. Her pain has no relation with food. Her past medical history is significant for recurrent abdominal pain due to cholelithiasis. Her father died at the age of 60 with some form of abdominal cancer. Her temperature is 37°C (98.6°F), respirations are 15/min, pulse is 67/min, and blood pressure is 122/98 mm Hg. Physical exam is unremarkable. However, a CT scan of the abdomen shows a calcified mass near her gallbladder. Which of the following diagnoses should be excluded first in this patient?

A. Acute cholecystitis

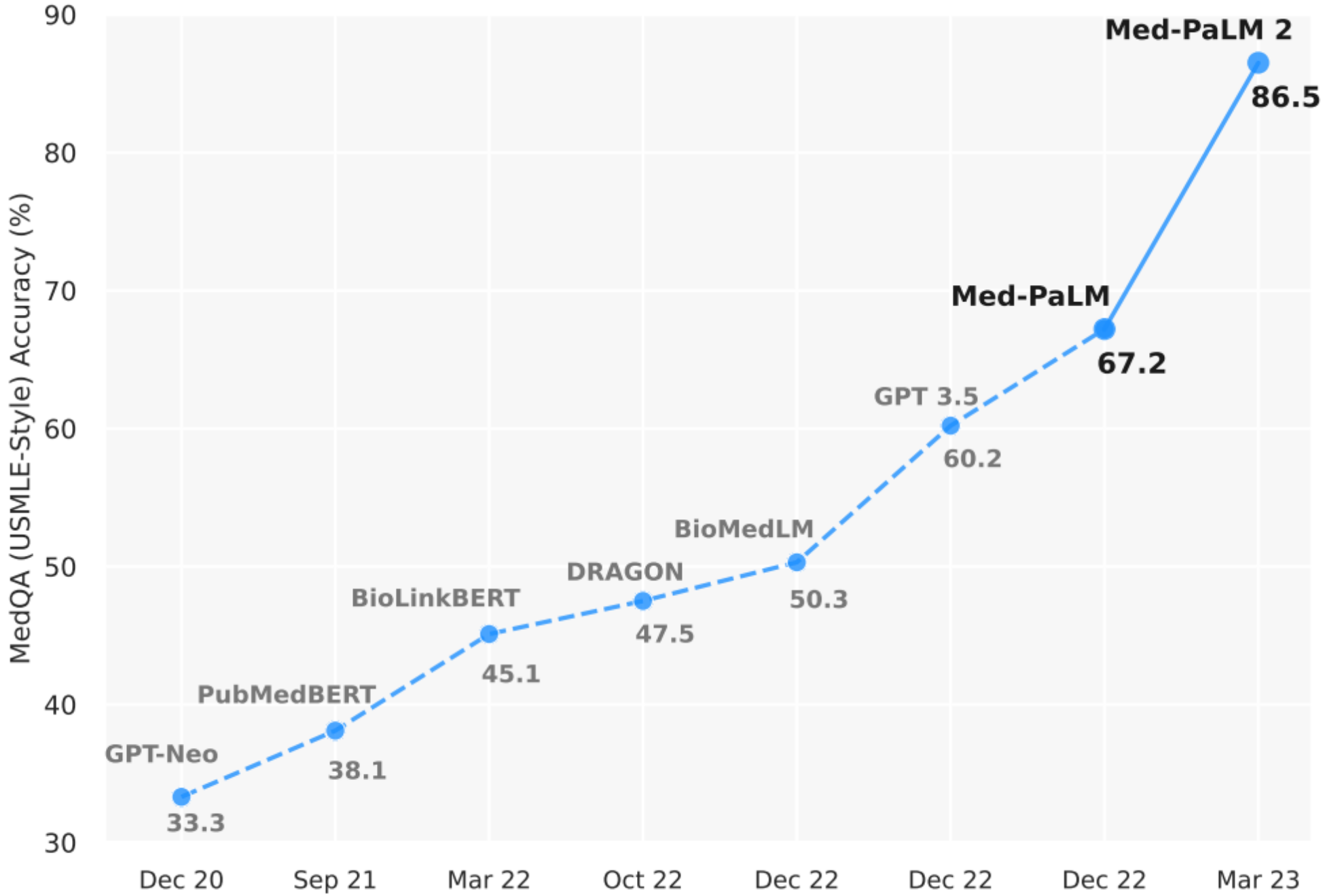
B. Gallbladder cancer

C. Choledocholithiasis

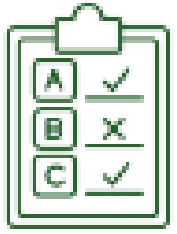
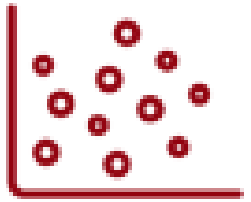
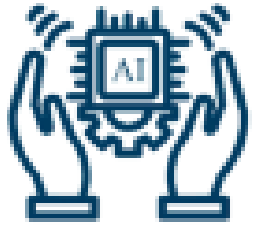
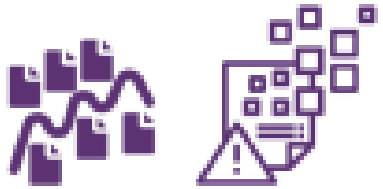
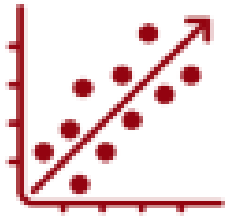
D. Pancreatitis

E. Duodenal peptic ulcer

https://www.researchgate.net/figure/A-MedQA-dataset-sample-validates-the-correctness-of-our-models-answer-In-the-MedQA_fig5_394534615



https://www.linkedin.com/pulse/quick-look-performance-googles-med-palm-2-medical-qa-jeroen-van-dalen/

	Discrepancy in Data	Discrepancy in Tasks	Discrepancy in Automated Metrics	Discrepancy in Translational Impact
Status Quo:	 LLMs are often evaluated on toy or synthetic examples	 Clinical LLM papers still primarily benchmark on multiple choice exams.	 Current automated metrics have low correlation with domain expert judgement.	 Model outputs are generally evaluated in isolation (e.g., intrinsic evaluation).
Real-world Considerations:	 Real EHR data is longitudinal and generally much noisier.	 LLMs are often being used for open-ended and interactive use cases.	 We need to develop and adopt methods that correlate more with clinical judgement.	 Models need to be evaluated in deployment context of the human-in-the-loop workflows.

Agrawal, Chen, Gulamali, Joshi. “The evaluation illusion of large language models in medicine”. *npj Digital Medicine* 2025.

What makes healthcare different?

High-stakes settings	→	Errors cause patient harm
Regulation	→	Experimentation is constrained
Data fragmentation	→	Ground truth is incomplete
Evolving knowledge	→	Benchmarks become outdated

What is a good healthcare evaluation?

1. Related to true clinical validity
2. Reflects heterogeneous effects and subgroup performances
3. Can models demonstrates impact on workflow?
4. Measures downstream clinical consequences
5. Updates over time
6. Alerts for new and rare problems

What is a good healthcare evaluation?

1. Related to true clinical validity
2. Reflects heterogeneous effects and subgroup performances
3. Demonstrates impact on workflow
4. Measures downstream clinical consequences
5. Updates over time
6. Alerts for new and rare problems

Can it perform the task?

What happens when people use it?

What happens after deployment?


Doctor
00:04:19
Patient
00:02:26
00:07:23

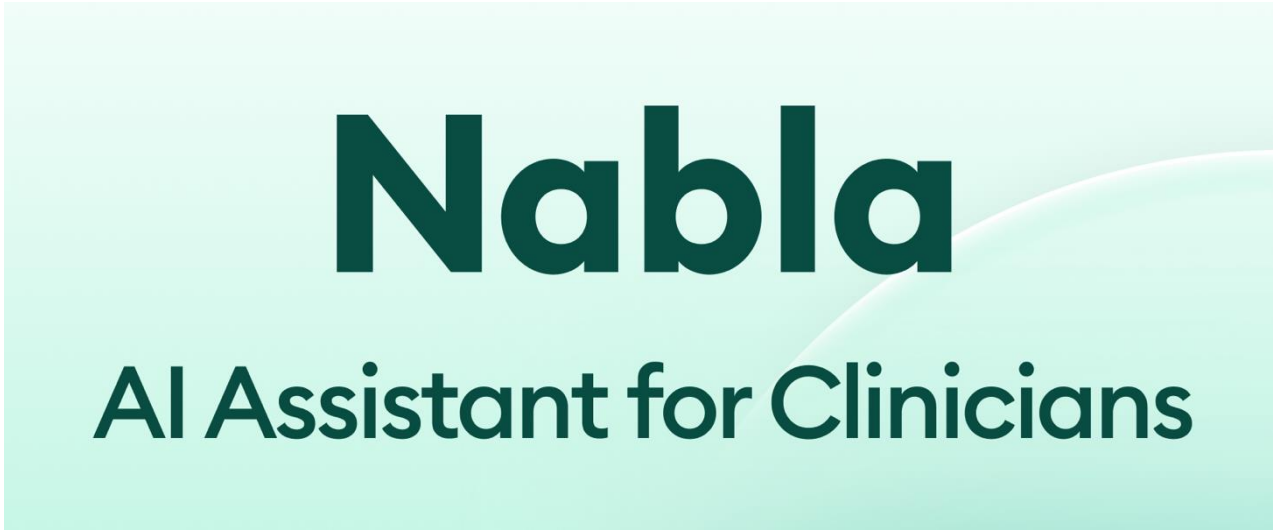
“I became ill with a fairly serious strain of viral something”

“I became ill with a fairly serious strain of viral something, but I didn't take any medication, I took Hyperactivated Antibiotics and sometimes I would think that was worse”

“I became ill with a fairly serious strain of viral something, but I didn't take any medication, I took Hyperactivated Antibiotics and sometimes I would think that was worse”

~1% hallucination rate

Koencke et al, 2024



7 million patient visits
30,000 clinicians
40 health systems

Associated Press, 2024

How well are AI scribes doing?

- Mixed results in early studies
  - Improves self-reported physician burnout (51.9% to 38.8% in 30 days)
  - RCT showed no impact on documentation time, completed appointment rate, same-day closure rate, or revenue per visit.
  - Net promoter score from 40 clinicians was 0 (from -100 to 100)
- Difficulties in evaluation
  - Concerning degree of hallucinations
  - Limited benchmark datasets and public evaluations
- Rapidly evolving landscape of model versions, hospital rollouts, patient engagement, and clinician preferences.

Olsen et al, *JAMA Network Open* 2025.
Liu et al, *NEJM AI* 2025.
Duggan et al, *JAMA Network Open* 2025

r/FamilyMedicine · 1y ago
wanna_be_doc DO

I just started using an AI scribe...

💖 Wellness 💖

I resisted for a long time to get on-board with GPT and AI, but my workplace finally integrated a dictation scribe into Epic. So I used it for the first time today.

Holy shit.

I write narrative notes and so need the more extensive notes to refresh my memory about the visits. However, this made chatting difficult and was my number one source of burnout. And it caused knockdown effects on my inbox results/messages.

Today is the first day in forever where my notes are done at 5 PM. I had time for patient messages/results during the day.

I'll never work without an AI scribe again.

uh034 · 1y ago
DO

I use DAX and it's great about 90% of the time. The other 10% I have to revise or add something that the AI missed for whatever reason.

74 Award Share ···

https://www.reddit.com/r/FamilyMedicine/comments/1jx1wdk/i_just_started_using_an_ai_scribe/

How can we evaluate AI scribes with clinician feedback?

Jessica Dai

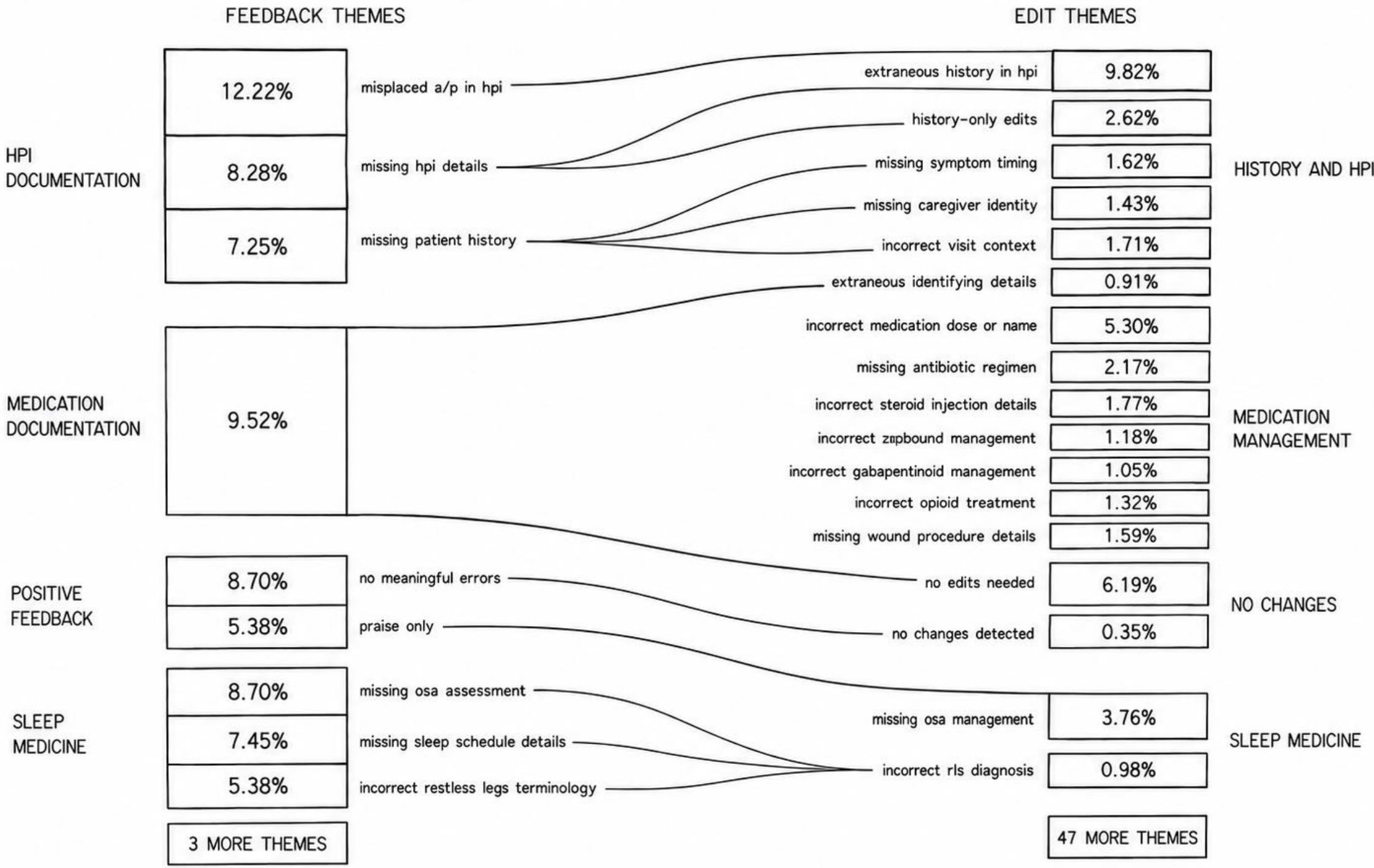
Anwen Huang

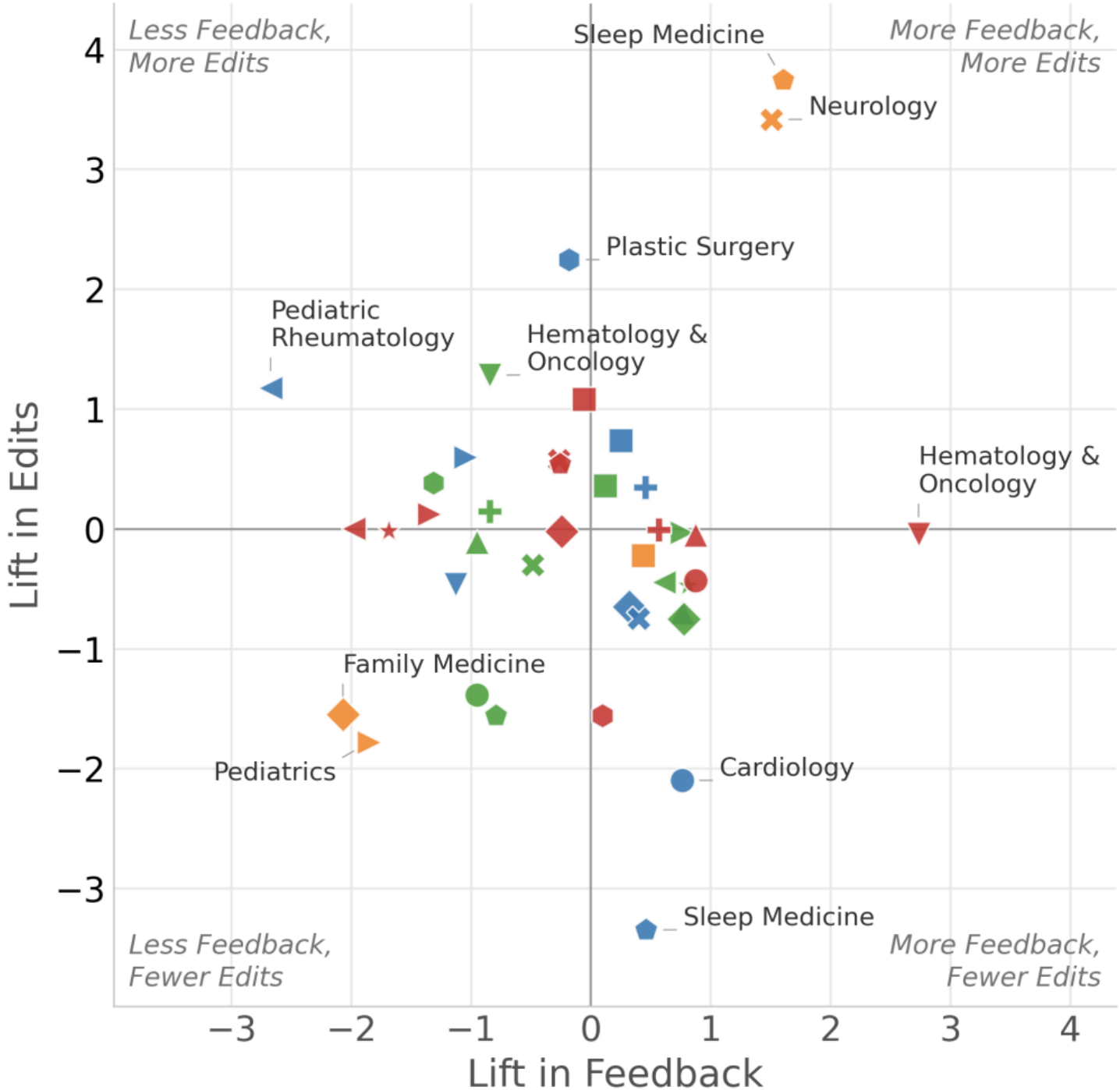
Clinician Feedback

50,151 encounters
483 feedback texts
0.9% feedback rate

Clinician Edits

37,933 encounters
23,771 with recorded edits
62.7% edit rate


FEEDBACK THEMES
EDIT THEMES
HPI DOCUMENTATION
12.22%
8.28%
7.25%
misplaced a/p in hpi
missing hpi details
missing patient history
MEDICATION DOCUMENTATION
9.52%
POSITIVE FEEDBACK
8.70%
5.38%
no meaningful errors
praise only
SLEEP MEDICINE
8.70%
7.45%
5.38%
missing osa assessment
missing sleep schedule details
incorrect restless legs terminology
3 MORE THEMES
extraneous history in hpi
history-only edits
missing symptom timing
missing caregiver identity
incorrect visit context
extraneous identifying details
incorrect medication dose or name
missing antibiotic regimen
incorrect steroid injection details
incorrect zɑpbound management
incorrect gabapentinoid management
incorrect opioid treatment
missing wound procedure details
no edits needed
no changes detected
missing osa management
incorrect rls diagnosis
47 MORE THEMES
9.82%
2.62%
1.62%
1.43%
1.71%
0.91%
5.30%
2.17%
1.77%
1.18%
1.05%
1.32%
1.59%
6.19%
0.35%
3.76%
0.98%
HISTORY AND HPI
MEDICATION MANAGEMENT
NO CHANGES
SLEEP MEDICINE



Theme

- Positive Feedback / No Changes
- Sleep Medicine
- HPI Documentation
- Medication Documentation

Specialty

- Cardiology
- Child Neurology
- Critical Care Medicine
- Family Medicine
- Hematology & Oncology
- Internal Medicine
- Neurology
- Pain Medicine
- Pediatric Rheumatology
- Pediatrics
- Plastic Surgery
- Sleep Medicine

Nuances

- Audio data is unavailable.
- Differences in rollout and clinician incentives.
- What is the theory of change here?
- Vendor can make unannounced model changes.
- What about patient feedback?

MK H.
Novato, CA
0 23 5

Jul 9, 2014

I went to Dr. Kvitash to discuss going on immuntherapy again and to get prescriptions for allergy medicine I used to get from Kaiser. I have dealing with my allergies since I had asthma many years ago as a small child. I have had multiple rounds of allergy shots and try to keep up to date on new information in the field.

Dr. Kvitash said he could not discuss my allergies and options for treatments with me until I had an "exam", blood tests and three rounds of skin tests. The cost would be $2950 from his office for the office based testing if I had insurance or $1500 if I do not. Plus an additional unknown amount for the blood tests (they were unwilling to estimate) and possibly a fee for doctor review and consultation. Until I agreed to it they would not tell me what blood tests they wanted me to get. I could look at the list of $2950 charges but I could not have a copy of it ("its our policy"). I would have to come back to his office for three rounds of skin tests. If he advised immunotherapy the 22 Battery office is only open on Wednesday afternoons for shots.

When I asked what possible treatments might come out of this testing Dr. Kvitash said he could not tell me anything. He said he needed to know "the state" of my immune system. His directive was he needed all this info to have any kind of discussion.

Some people might be comfortable with this approach but I was not. I am glad that one poster mentioned this in advance so I was prepared for it. Otherwise I might have gone along with it and questioned it later.

Dr. Kvitash is a very pleasant man. His office however, looks like it was furnished from a dumpster. Old lumpy dusty maroon and plaid chairs from the 70s. The meeting left me longing for the good old days when I had Kaiser and felt like someone was watching over the doctor's shoulder and when I did not worry that the doctor might be more interested in the patient fees than helping the patient understand their health.

I welcome hearing from other patients how this worked for them.

“The cost would be $2950 from his office for the office based testing if I had insurance or $1500 if I do not”

“Until I agreed to it they would not tell me what blood tests they wanted me to get”

https://www.yelp.com/biz/vadim-kvitash-md-phd-san-francisco-2

MK H.
Novato, CA
0 23 5

Jul 9, 2014

I went to Dr. Kvitash to discuss going on immuntherapy again and to get prescriptions for allergy medicine I used to get from Kaiser. I have dealing with my allergies since I had asthma many years ago as a small child. I have had multiple rounds of allergy shots and try to keep up to date on new information in the field.

Dr. Kvitash said he could not discuss my allergies and options for treatments with me until I had an "exam", blood tests and three rounds of skin tests. The cost would be $2950 from his office for the office based testing if I had insurance or $1500 if I do not. Plus an additional unknown amount for the blood tests (they were unwilling to estimate) and possibly a fee for doctor review and consultation. Until I agreed to it they would not tell me what blood tests they wanted me to get. I could look at the list of $2950 charges but I could not have a copy of it ("its our policy"). I would have to come back to his office for three rounds of skin tests. If he advised immunotherapy the 22 Battery office is only open on Wednesday afternoons for shots.

When I asked what possible treatments might come out of this testing Dr. Kvitash said he could not tell me anything. He said he needed to know "the state" of my immune system. His directive was he needed all this info to have any kind of discussion.

Some people might be comfortable with this approach but I was not. I am glad that one poster mentioned this in advance so I was prepared for it. Otherwise I might have gone along with it and questioned it later.

Dr. Kvitash is a very pleasant man. His office however, looks like it was furnished from a dumpster. Old lumpy dusty maroon and plaid chairs from the 70s. The meeting left me longing for the good old days when I had Kaiser and felt like someone was watching over the doctor's shoulder and when I did not worry that the doctor might be more interested in the patient fees than helping the patient understand their health.

I welcome hearing from other patients how this worked for them.

Challenges:

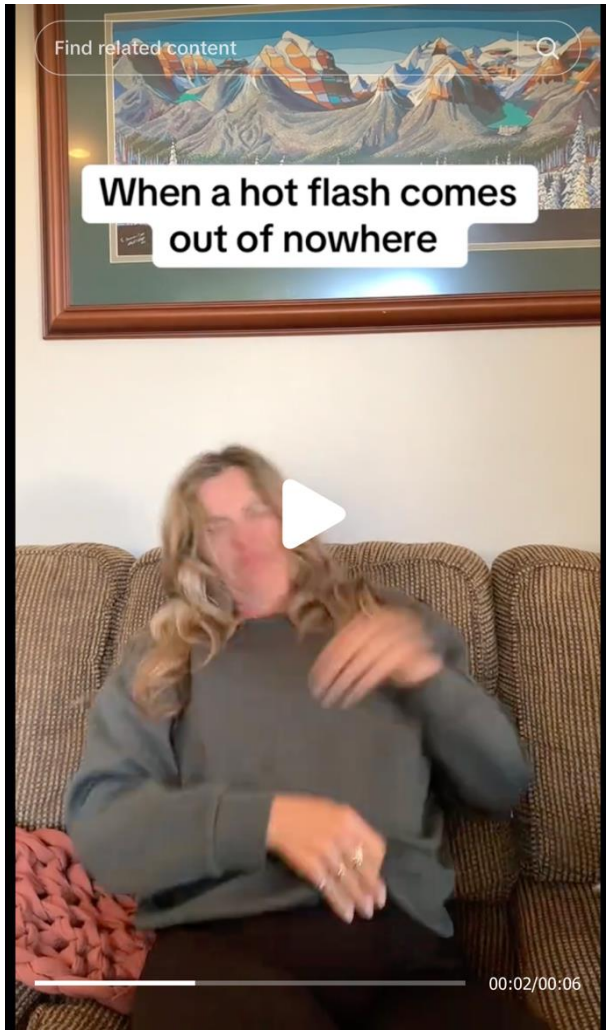
- Who is this person?
- What was the actual health encounter and the medical advice given?
- What was the long-term health impact?

https://www.yelp.com/biz/vadim-kvitash-md-phd-san-francisco-2

VAERS Vaccine Adverse Event Reporting System
www.vaers.hhs.gov

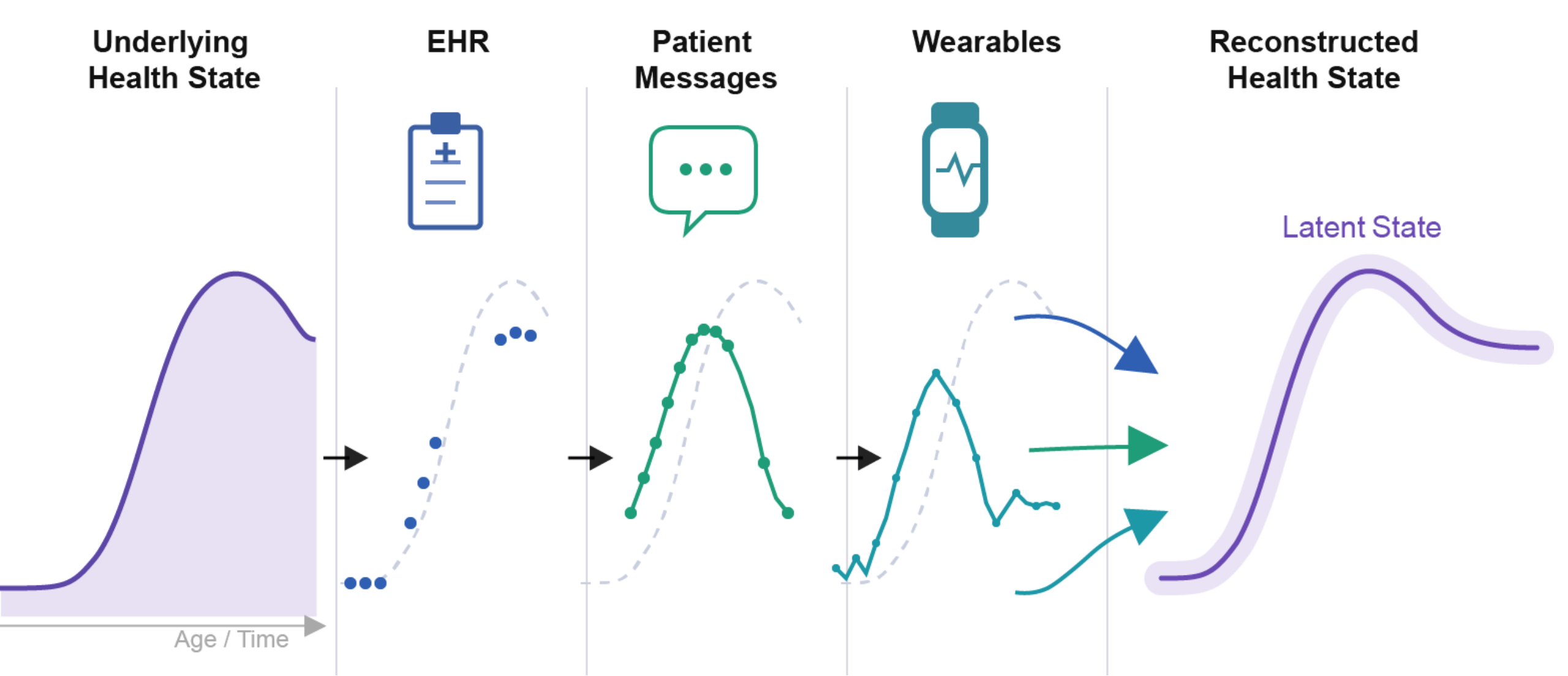
Millions of publicly available vaccine adverse-event reports since 1990

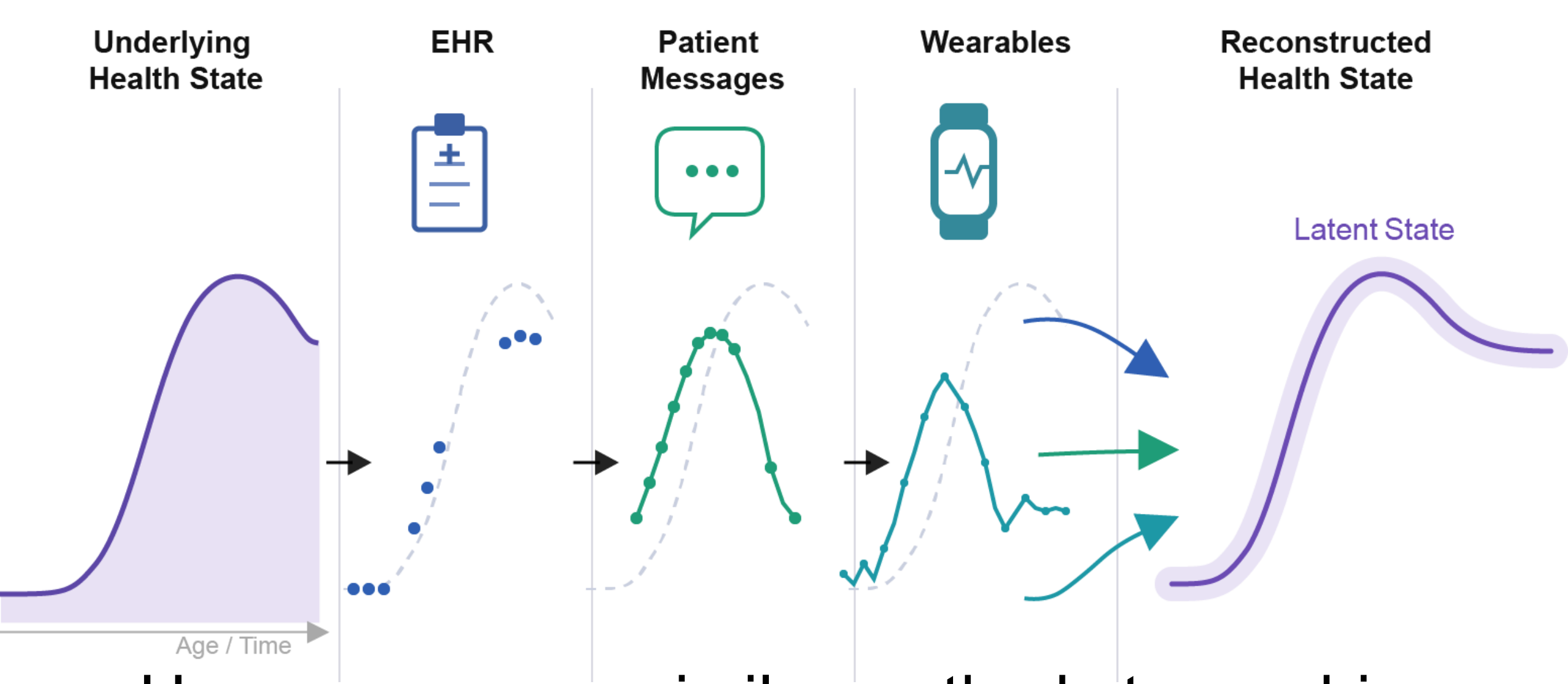
48% of nurses said that these automated reports don't match their assessment.



36% of U.S. adults get health information from social media at least sometimes.

https://vaers.hhs.gov/
https://www.nationalnursesunited.org/press/national-nurses-united-survey-finds-ai-technology-undermines-patient-safety
https://www.tiktok.com/@kristinamullen_/video/7564873597085355


Underlying Health State
EHR
Patient Messages
Wearables
Reconstructed Health State
Latent State
Age / Time



How can we use similar methods to combine disparate sources of public feedback?

iychen@berkeley.edu