From the EHR to Digital Twins: Foundational Concepts for Advancing AI-enabled Cancer Research

Aaron Cohen, MD, MSCE
Senior Medical Director | Clinical Lead for AI and Digital Oncology @ Flatiron Health
Adjunct Assistant Professor @ NYU Grossman School of Medicine

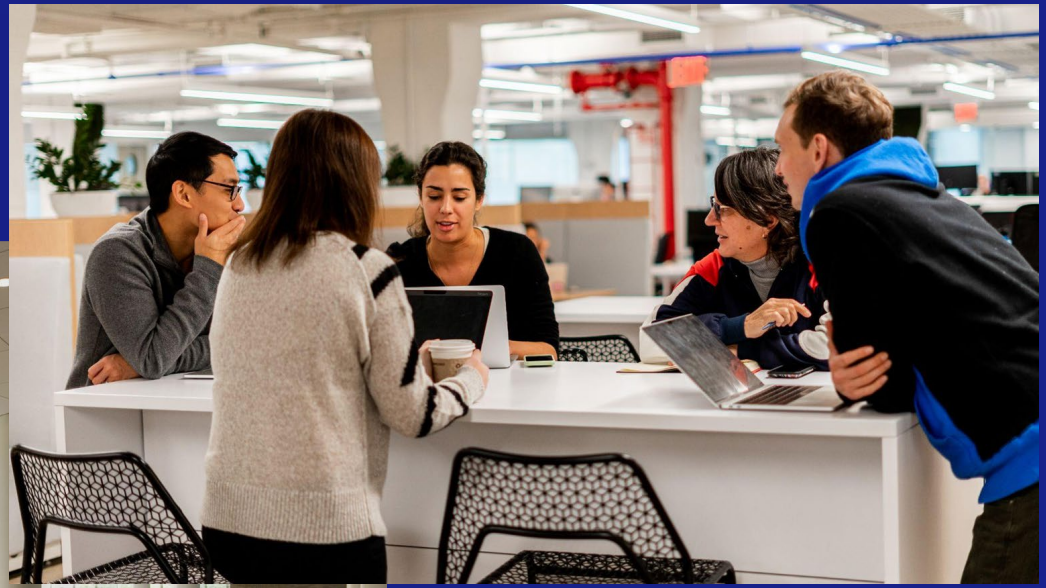
Employee Disclaimer

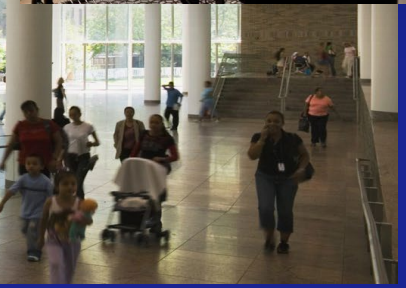
Aaron Cohen is an employee of Flatiron Health, an independent subsidiary of Roche Group. He holds stock ownership in Roche.

“Poor Daddy. You work so hard trying to cure cancer and it never works!”

Elyse ~2 weeks ago


HOSPITAL & EMERGENCY ROOM





Agenda

- Data extraction
- Model validation
- Digital twin simulation

Agenda

- **Data extraction**
- Model validation
- Digital twin simulation

Notes | Review | Mark All As Reviewed | Summarize Notes

Create Note | Telehealth Visit 1

My Note

Progress Notes • Oncology • 2/26/2026 0923

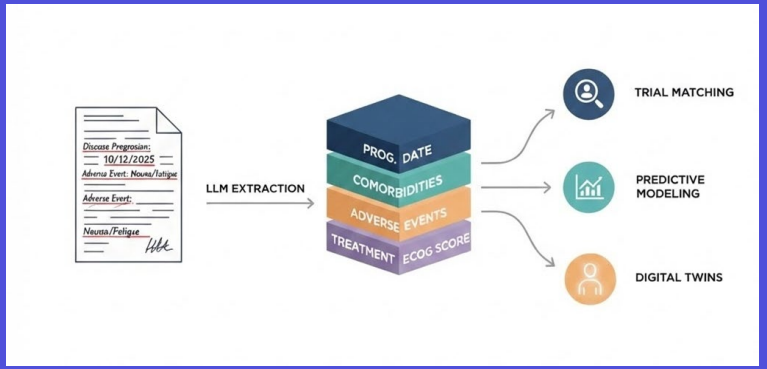
Summary:

Insert SmartText

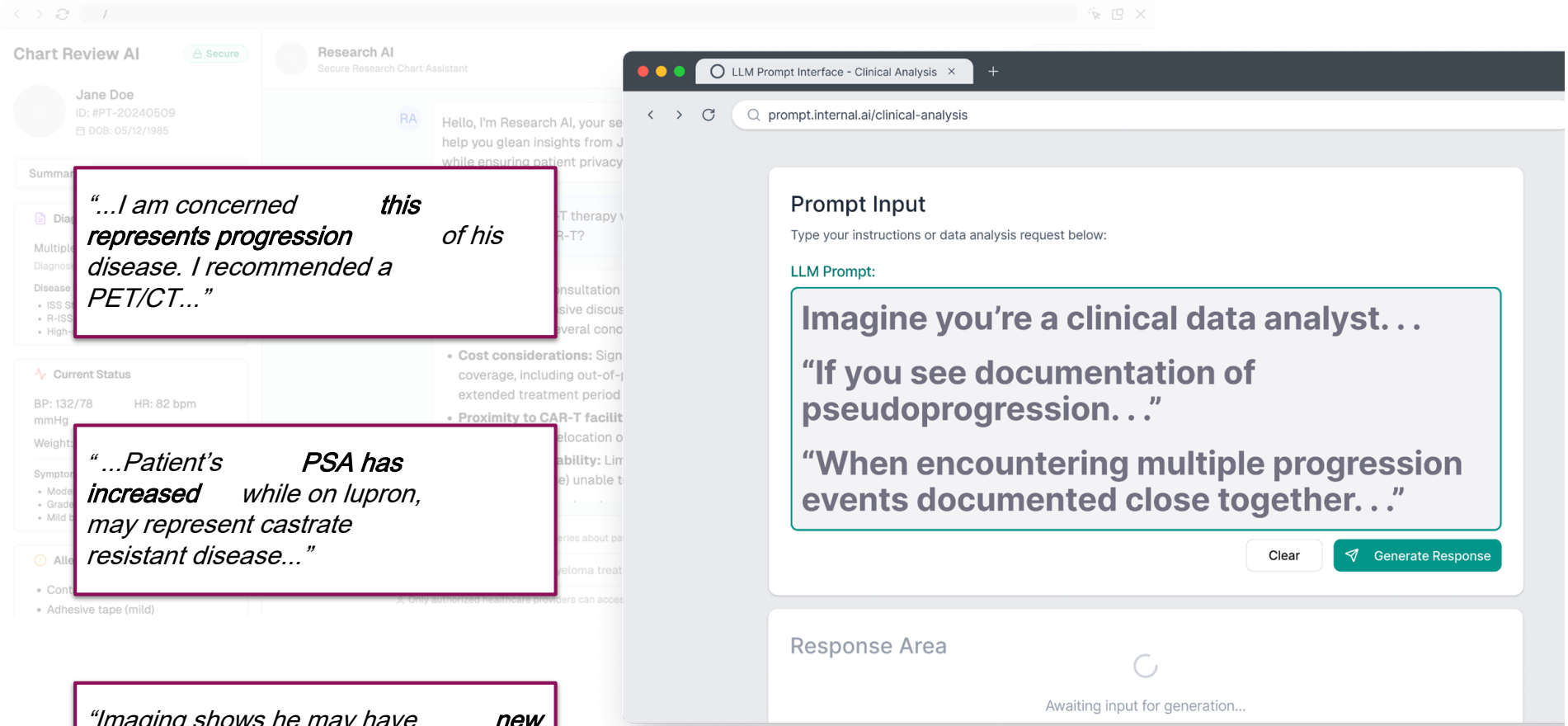
MEDICAL ONCOLOGY PATIENT VISIT
DATE: 3/9/2026
NAME: Patient A
DOB: 1/1/1970

Interval History: Patient A presents to clinic for his diagnosis of **advanced NSCLC**. He was **diagnosed 2/1/26 with metastatic disease** to the **liver and bone**. **EGFR, ALK, ROS1 neg, PDL1 <5%**. Discussed risks of immunotherapy due to history of **rheumatoid arthritis**. Started **carbo/pem/pem 2 weeks ago**, patient reports **fatigue** and **diarrhea**. Labs show **TSH elevated at 9**.

Extracting Clinical Data from the EHR



Cohen AB, Adamson B, Larch JK, Amster G. Large Language Model Extraction of PD-L1 Biomarker Testing Details From Electronic Health Records. AI Precis Oncol. Published online 2025. doi:10.1089/aipo.2024.0043

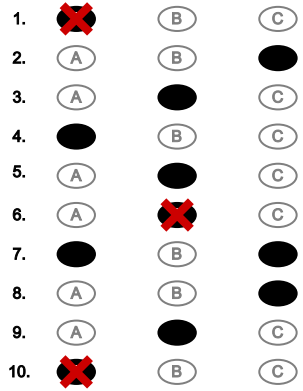


Cohen AB et al. Using large language models for scalable extraction of real -world progression events across multiple cancer types. Presented at: AACR Special Conference in Cancer Research: Artificial Intelligence and Machine Learning; July 10 - 12, 2025; Montreal, QC, Canada.

Agenda

- Data extraction
- **Model validation**
- Digital twin simulation

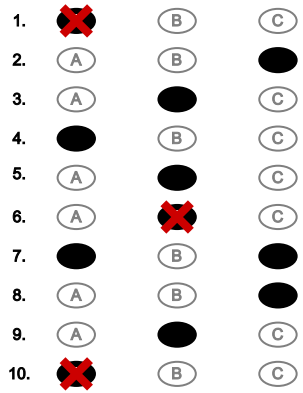
The answer key problem



7/10 Test results

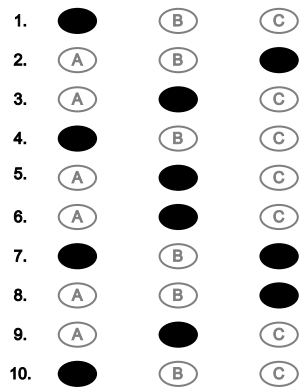
The answer key problem

Flawed answer key



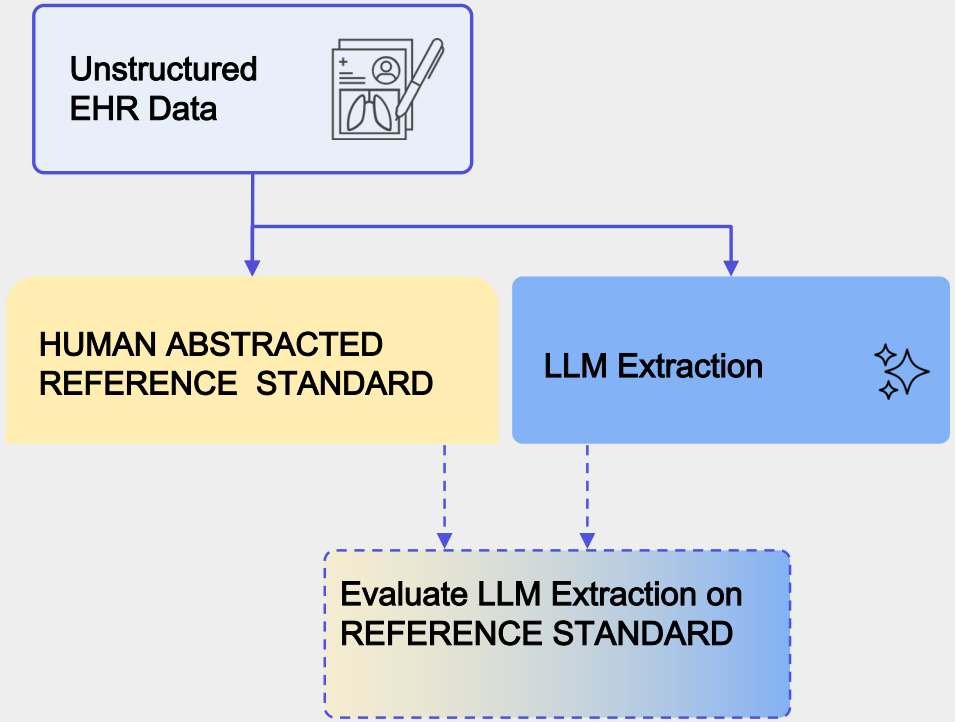
7/10 Test results

Correct answer key

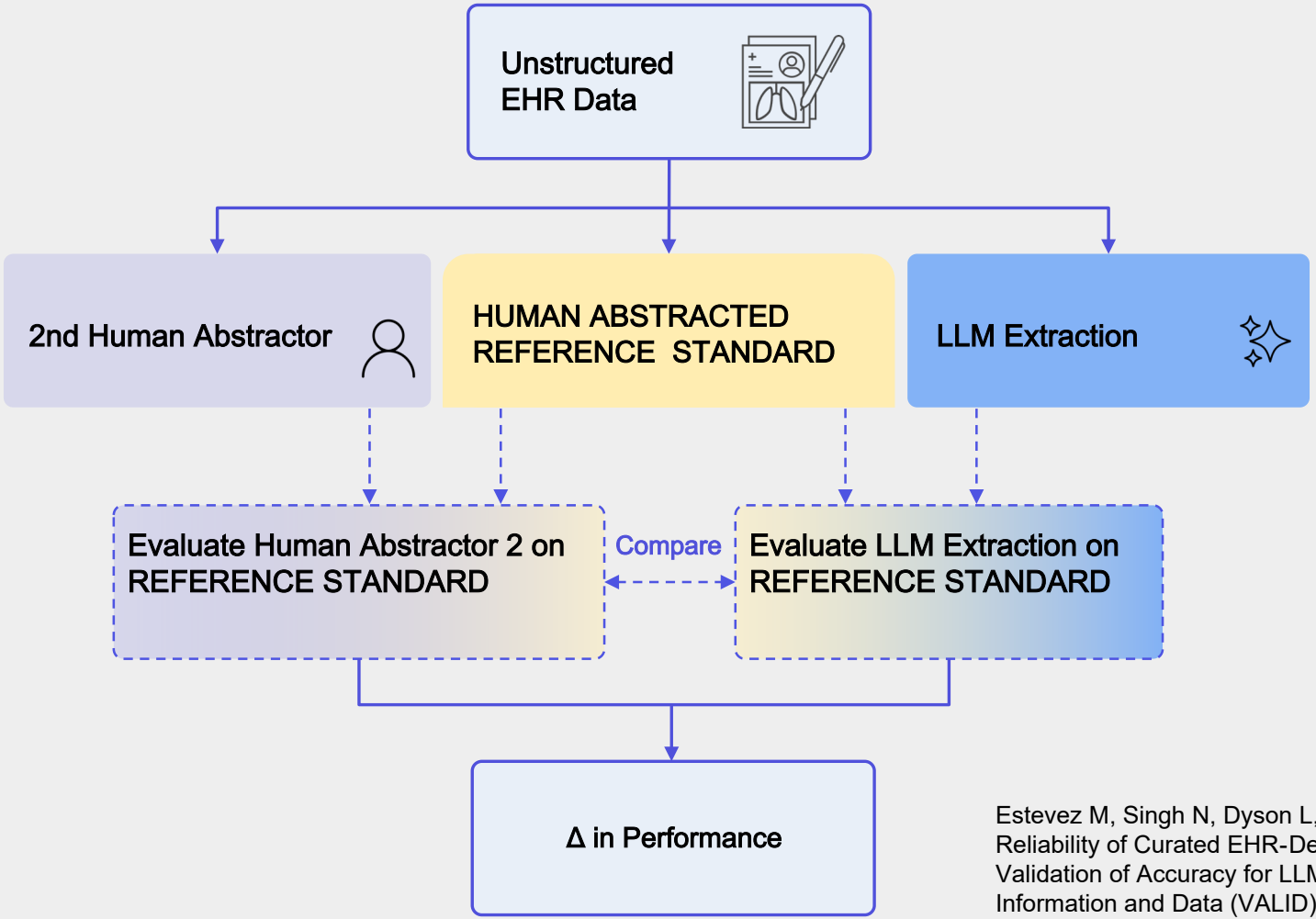


10/10 Test results

Same answers



Estevez M, Singh N, Dyson L, et al. Ensuring Reliability of Curated EHR-Derived Data: The Validation of Accuracy for LLM/ML -Extracted Information and Data (VALID) Framework. *JCO Clin Cancer Inform*. In Press.



Estevez M, Singh N, Dyson L, et al. Ensuring Reliability of Curated EHR-Derived Data: The Validation of Accuracy for LLM/ML -Extracted Information and Data (VALID) Framework. *JCO Clin Cancer Inform*. In Press.

Agenda

- Data extraction
- Model validation
- **Digital twin simulation**

Predicting Survival on Standard of Care Therapy

1

Predicting Survival on Standard of Care Therapy

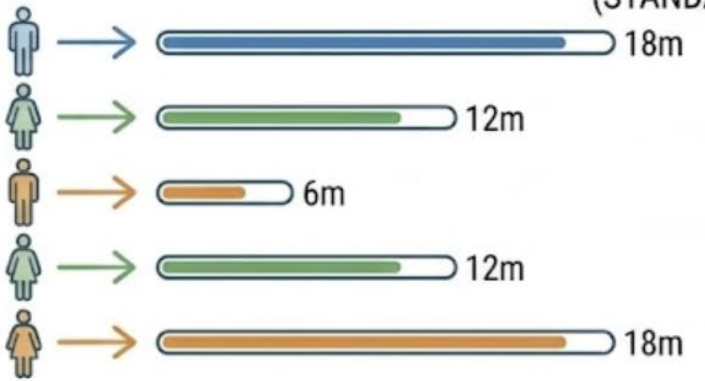
1

REAL-WORLD DATA AI MODEL
(Trained on Thousands of Patients)

LEARNS FEATURE WEIGHTS
(e.g., biomarker importance)

DIGITAL TWIN COHORT

PREDICTED SURVIVAL
(STANDARD CARE)



2

Predicting Survival on Standard of Care Therapy

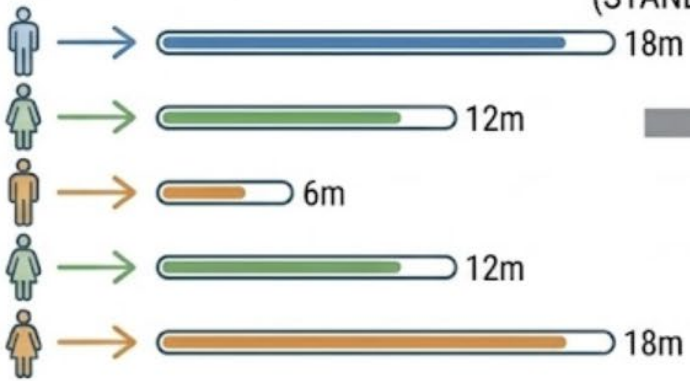
1

REAL-WORLD DATA AI MODEL
(Trained on Thousands of Patients)

LEARNS FEATURE WEIGHTS
(e.g., biomarker importance)

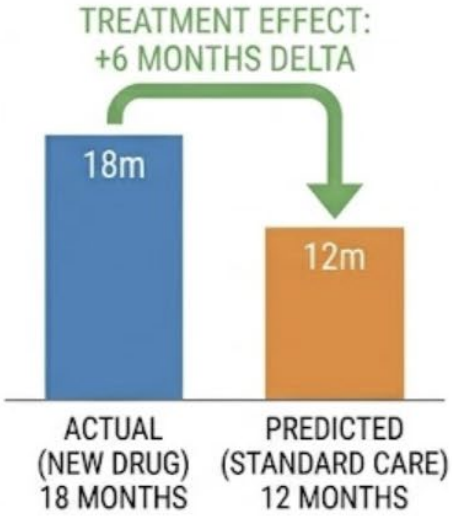
DIGITAL TWIN COHORT

PREDICTED SURVIVAL
(STANDARD CARE)



2

THE COMPARISON: ISOLATED TREATMENT EFFECT



3

Image created using Google Gemini version 3.1 Pro

CLOSING THOUGHTS

The goal: to help our patients live longer and better lives