



2026 World Conference on Lung Cancer

SEPTEMBER 12 – 15, 2026 | SEOUL, REPUBLIC OF KOREA

EHRsmoker: Recovering missed lung cancer screening eligibility via longitudinal smoothing of EHR smoking data

Tae Yoon (Harry) Lee, PhD

on behalf of Mengrui Zhang, PhD and Ingrid Luo, MS

USA

Disclosure

Artificial Intelligence (AI) Use Disclaimer

AI was not used for the content of these slides.

Background

- Identifying screening-eligible individuals under the 2021 U.S. Preventive Services Task Force (USPSTF) guidelines requires **reliable smoking data** on pack-years and time since quitting.
- In routine care, **electronic health record (EHR) smoking histories** are often **missing, internally inconsistent, or contradictory** because of recall bias, inconsistent documentation, and errors in derived exposure measures.

Objective

- We developed an **automated data-smoothing algorithm** (EHRsmoker) that leverages **validated large language models (LLMs)** for clinical-note extraction and **evaluated its performance** across academic and community healthcare systems.

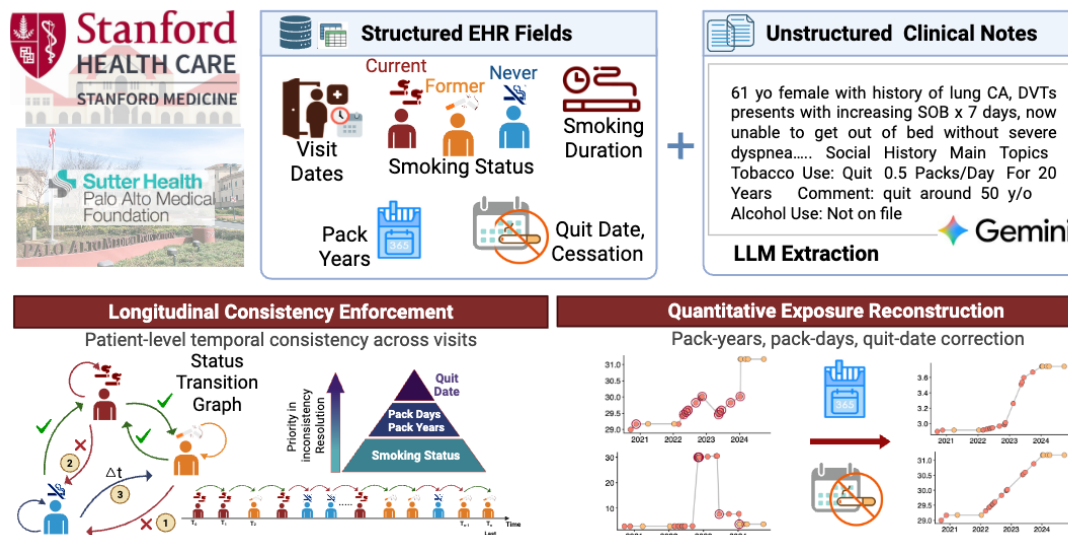
Methods: EHRsmoker algorithms & implementation

EHRsmoker algorithms

- Combines **structured EHR and LLM-extracted** smoking data.
- **Resolves inconsistencies** using clinically plausible transitions.
- **Reconstructs** pack-years and corrects missing or implausible values.

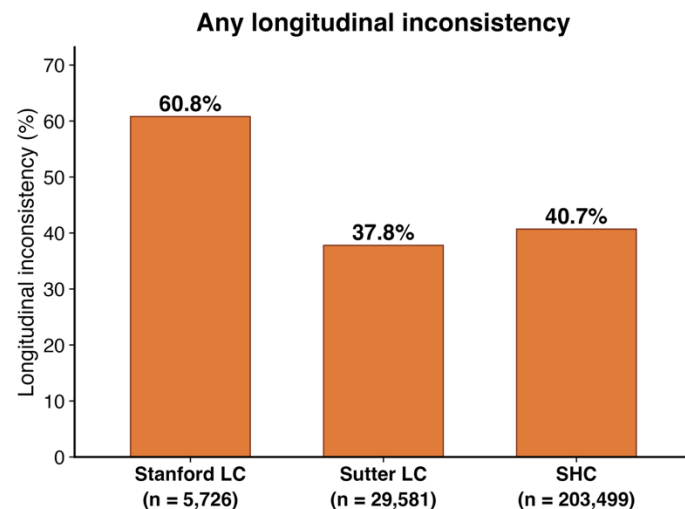
EHRsmoker implementation

- Retrospectively evaluated in Stanford lung cancer, Sutter lung cancer, and Stanford primary care cohorts.
- Assessed **improvements in USPSTF 2021 lung cancer screening sensitivity in lung cancer cases** after data smoothing.



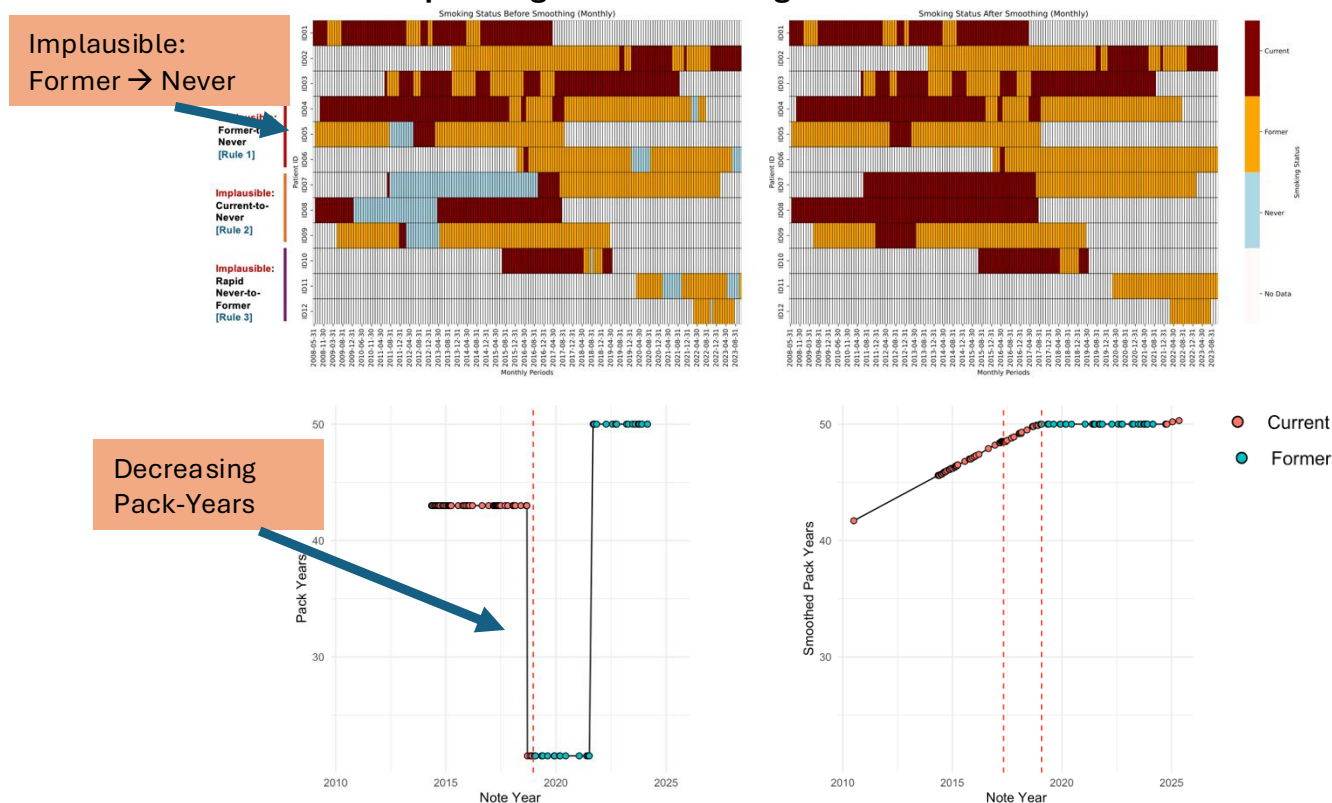
Study flow diagram

Results: longitudinal inconsistencies across different healthcare systems.



Result 1: Longitudinal smoking histories were **highly inconsistent**: **60.8%** in Stanford lung cancer, **37.8%** in Sutter lung cancer, and **40.7%** in Stanford primary care cohorts.

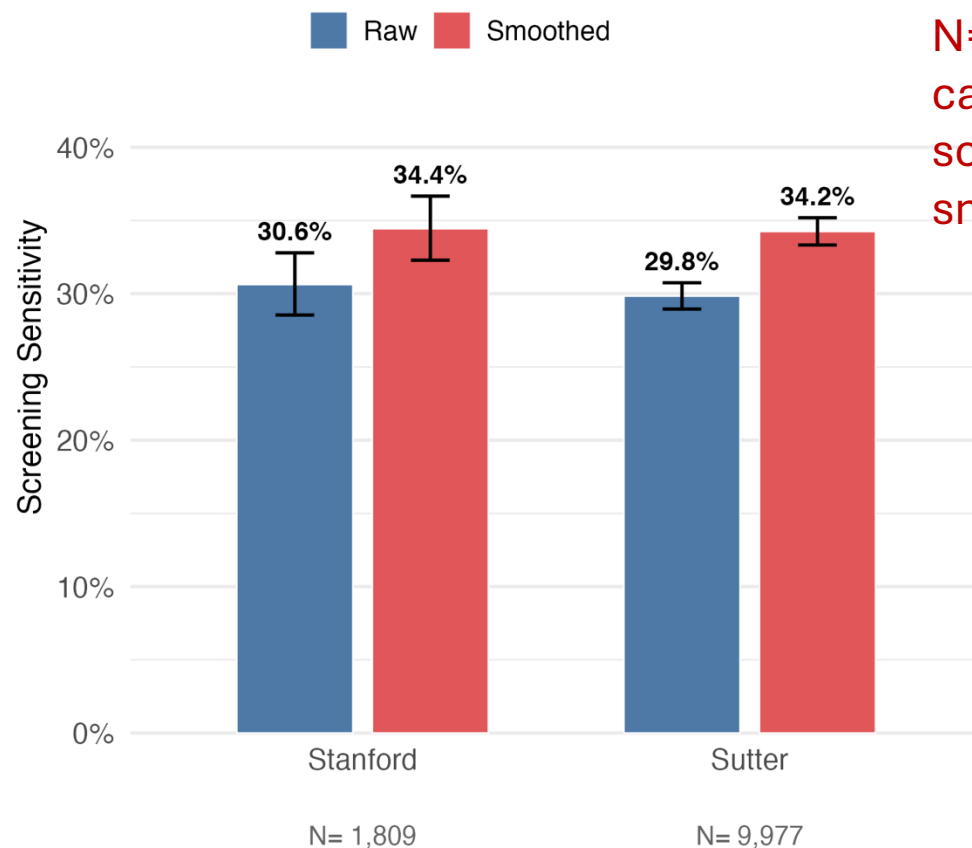
Example Longitudinal Smoking Status Data from 12 Patients



Result 2: Common errors included **conflicting quit dates**, **decreasing pack-years**, and **implausible status changes**; smoothing eliminated longitudinal inconsistencies and reduced record-level inconsistency to $\leq 0.2\%$.

Results: Screening sensitivity before vs. after smoothing

USPSTF Screening Sensitivity: Raw vs. Smoothed



N=438 additional true LC cases would have been screening-eligible if smoothed smoking data had been used

Result 4: Sensitivity increased from about 30.6% to 34.4% (Stanford) and 29.8% to 34.2% (Sutter)

Main driver of increased sensitivity was through **pack-year reconstruction** (89–97% of newly eligible patients) and **quit-date smoothing** (~10%).

Takeaway

- Our framework, implemented in the **open-source R package EHRsmoker** and recovers missed opportunities by unifying structured data with LLM-extracted narratives into complete smoking histories.
- It offers a **transportable and automated solution** to **improve screening-eligibility identification** across different healthcare systems.

Acknowledgement

Stanford University

- Summer S Han
- Ingrid Luo
- Victoria Y. Ding
- Bo Gu
- Grant M. Nieda
- Ntemena Kapula
- Annabel X. Tan
- Curtis P. Langlotz
- Ann Leung
- Joel W. Neal
- Allison W. Kurian
- Shreya Shah
- Heather A. Wakelee
- Leah M. Backhus
- Natalie S. Lui
- Aparajita Khan

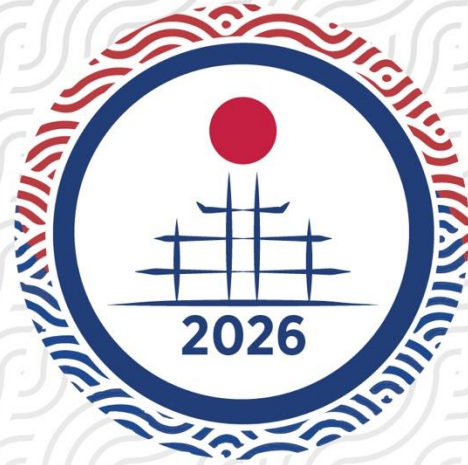
Stanford Health Care

- Mina Satoyoshi
- Archana Bhat
- A. Solomon Henry

Sutter Health

- Su-Ying Liang
- Tony Chen
- Meghan Martinez





Thank You.