# Development of an NLP Algorithm to Identify E-Vapor Use in EHR Clinical Notes

TSRC 2025 Annual Conference

**Derek Pope, PhD**

**Altria Client Services LLC (ALCS)**

September 17, 2025

# Special Thanks to...

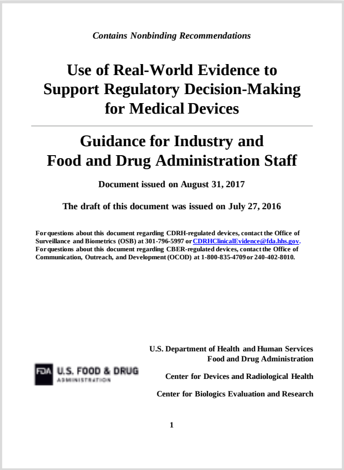
## Co-authors:

| | |
|---|---|
| **Optum Life Sciences** | Noelle N. Gronroos<br>Ami R. Buikema |
| **Optum Global Solutions** | Abhinav Nayyar<br>Shashi Khan<br>Shikha Anand<br>Tamanna Tara<br>Pankaj Kang |
| **Altria Client Services LLC (ALCS)** | Mingda Zhang |

Thanks to numerous ALCS colleagues for comments and review

# Closing the Gap Between Pre-market Authorization and Epidemiology

Premarket Review



*Contains Nonbinding Recommendations*

**Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices**

**Guidance for Industry and Food and Drug Administration Staff**

**Document issued on August 31, 2017**

**The draft of this document was issued on July 27, 2016**

For questions about this document regarding CDRH-regulated devices, contact the Office of Surveillance and Biometrics (OSB) at 301-796-5997 or CDRHClinicalEvidence@fda.hhs.gov. For questions about this document regarding CBER-regulated devices, contact the Office of Communication, Outreach, and Development (OCOD) at 1-800-835-4709 or 240-402-8010.

FDA U.S. FOOD & DRUG ADMINISTRATION

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Devices and Radiological Health
Center for Biologics Evaluation and Research

1

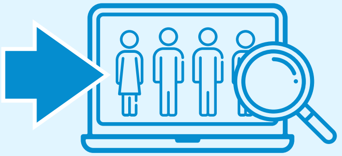
Epidemiology

The objective of our study was to develop an NLP algorithm to extract:

Electronic Nicotine Delivery Systems (ENDS; primary focus)

AND

Nicotine Pouch (secondary focus)

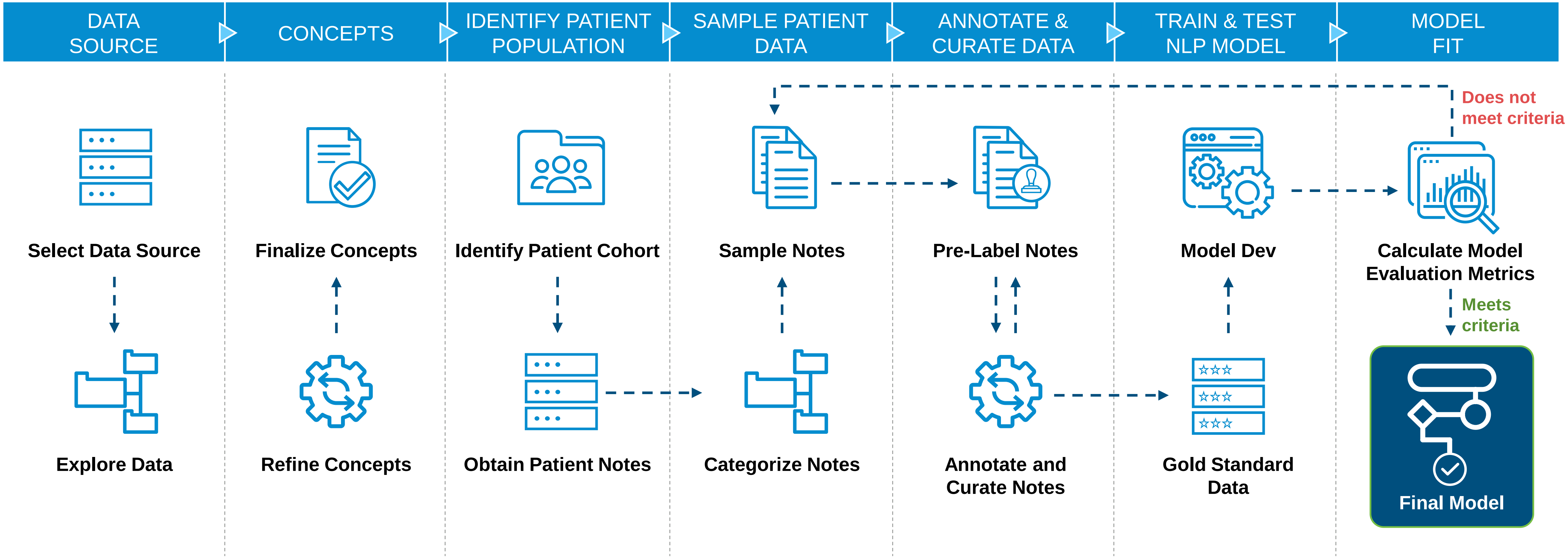


Use Data

from unstructured clinical notes from EHR to support RWE for tobacco harm reduction

Source: Guidance for Industry and Food and Drug Administration Staff August 2017. Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices.

# NLP Model Development Approach



# Large Study Populations and High EHR Note Counts

**Data Source: Optum® de-identified EHR data set (01/2007-08/2023)**

| | | PATIENTS** (n) | NOTES** (n) |
|---|---|---|---|
| **EHR Population*** | ≥ 1 ICD code for combustible tobacco use | 13,583,236* | |
| **EHR Notes Population** | ≥ 1 clinical note | 5,194,000 | 862,491,151 |
| **Study Population** | ≥ 1 note containing a term for ENDS, nicotine pouches, or combustible tobacco | 4,371,006 | 66,101,820 |

*EHR population (first row) is among the full set of patients/notes in the full EHR data asset. Each subsequent row is a subset of the prior row

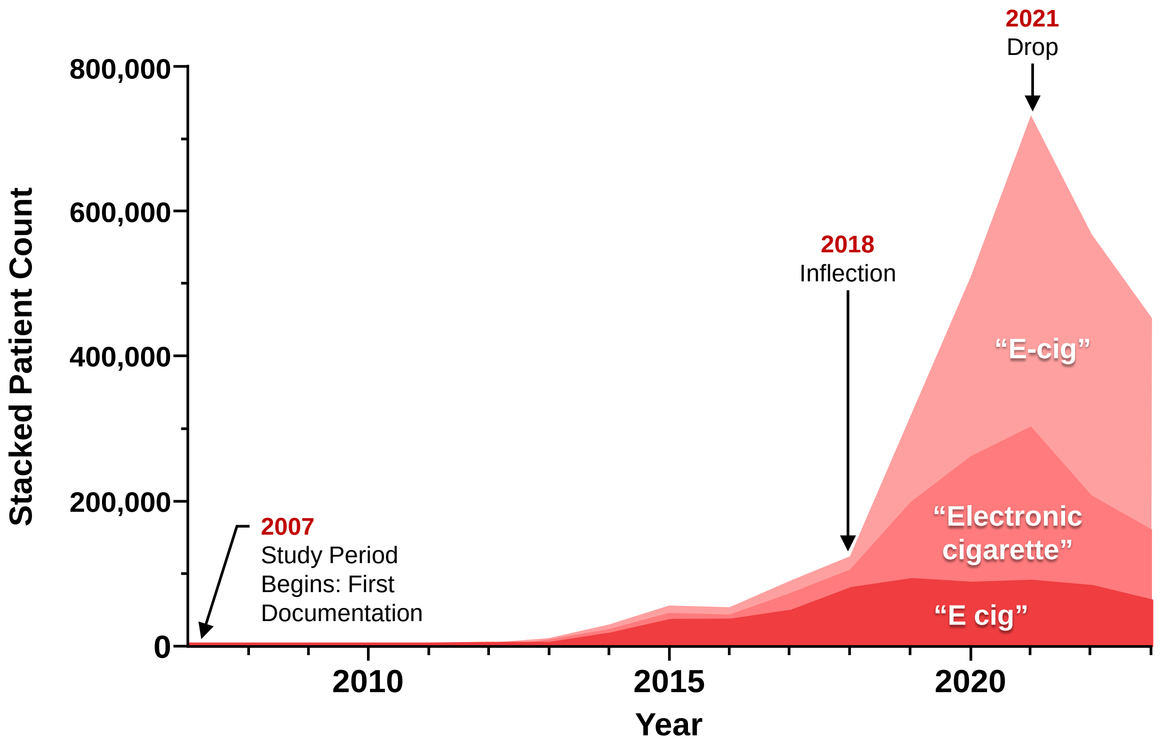
# Large Variation in Terms Used in Clinical Notes

| Product Category | Terms Tested | | Terms Identified | |
|---|---|---|---|---|
| ≥ 1 Combustible tobacco term | 222 Total | • 128 cigarette<br>• 94 other combustible | 118 Total | • 83 cigarette<br>• 35 other combustible |
| ≥ 1 ENDS term OR<br>≥ 1 ENDS brand term | 130 Total | • 78 ENDS<br>• 52 brand | 104 Total | • 77 ENDS<br>• 27 brand |
| ≥ 1 Nicotine pouch term OR<br>≥ 1 Nicotine pouch brand term | 22 Total | • 10 pouch<br>• 12 brand | 18 Total | • 8 pouch<br>• 10 brand |

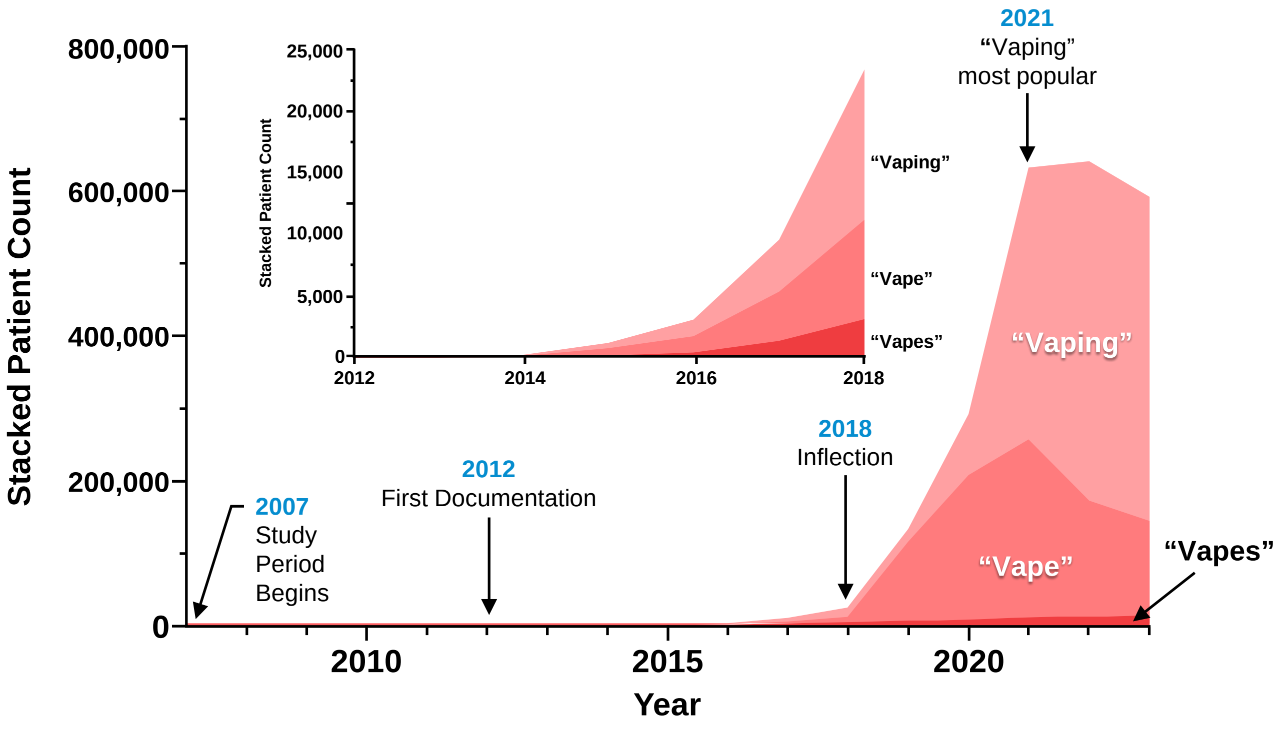
DATA SOURCE

# ENDS Terms in EHR Clinical Notes Evolved Over Time

## E-cigarette Terms




## Vape Terms




*Patients may appear in more than one year and in more than one term
**If a term's text is repeated within another term (e.g., the string of characters comprising the term "vape" is a substring within the string of characters comprising the term "vaped"), all patients/notes in the string are also counted in the total of patients/notes in the substring

DATA SOURCE

# Finalized NLP Tobacco Use Concepts

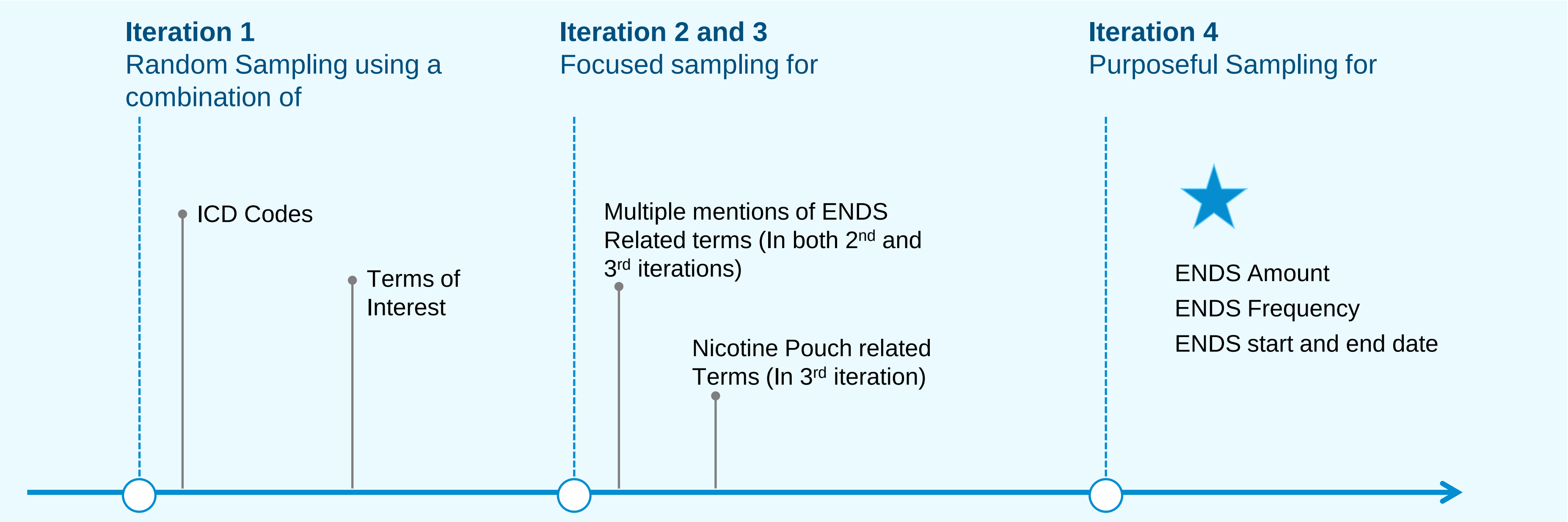
**The terms identified in the EHR Clinical notes were used to build the tobacco use concepts selected for extraction using NLP**

- ENDS usage, frequency, and duration
- Combustible tobacco usage
- Nicotine pouch usage

| Concept | Values | Brief Description |
|---|---|---|
| **ENDS** Use Status | Current use | Note indicates patient is currently using ENDS products |
| | Former use | Note indicates patient has previously used ENDS products but not using anymore |
| | Never use | Note indicates that patient has never used ENDS products. |
| | *[Exploratory]* Experimenter/trier | Note indicates that patient has not consistently used an ENDS product (either in the past or currently*)* |
| Duration | Time period *[Start/Stop dates]* | For current, former, and experimental/trier users, duration of use was captured and, where available, start and stop dates |
| Amount/Frequency | Amount *[per time period]* | For current and former users, the amount of use was captured *[If available, over what time period will be included (e.g., per day)]* |
| **Combustible** Use | Current use | Note indicates patient is currently using combustible tobacco products |
| | Former use | Note indicates patient has previously used combustible tobacco products but not using anymore |
| | Never use | Note indicates that patient has never used combustible tobacco products |
| **Nicotine Pouch** Use | Current use | Note indicates patient is currently using nicotine pouch products |
| | Former use | Note indicates patient has previously used nicotine pouch products but not using anymore |
| | Never use | Note indicates that patient has never used nicotine pouch products |

CONCEPTS

# Clinical Notes Randomly Sampled, Annotated, & Curated (Hypothetical Data Illustrations)




Has Relation

Combustible tobacco — Combustible tobacco status

**Do you smoke tobacco?: Patient has been a former user of cigarettes**

**Pack years of tobacco: 12 Cigarettes per day, advised to stop smoking cigarettes but patient has no plans of quitting smoking**

Has Relation

ENDS — ENDS status

**Do you or have you ever used e-cigarettes or Vape?: Current user of e-cig**

ENDS — ENDS frequency

**(Notes: Vapes 10-20 puffs per day………..stopped vaping few days back)**

SAMPLE PATIENT DATA → ANNOTATE & CURRATE DATA

# NLP Model Architecture

## Model Architecture utilized:

Char-CNN → BiLSTM → CRF*

**Final models were selected by comparing fit statistics across two additional embeddings used to improve performance:**

**1 HealthCare Embeddings**
pre-trained weights from a language representation model for the healthcare domain which is trained on:
PubMed + ICD10 + UMLS + MIMIC III corpora

**2 Clinical Embeddings**
pre-trained weights from a language representation model for clinical domains, which are trained on:
PubMed corpora

*John Snow Labs (JSL) was used for model development, which utilizes the following: Char-CNN=Character-level convolutional neural network – a deep learning model whose strength is dealing with terms that were not part of the training data; BiLSTM=Bidirectional long short-term memory – uses neural networks and is good at distinguishing context (e.g., past, present, and future); CRF=Conditional random field – a probabilistic graphical model that excels prediction based on a sequence.

# Model Fit & Acceptance Criteria

## Precision (≥ 80%)

IT IS THE RATIO OF
**true positive predictions to the total number of positive predictions made by the model**

IT ANSWERS THE QUESTION
**“Of all the instances that the model predicted as positive, how many were actually positive?”**

## Recall (≥ 70%)

IT MEASURES THE
**ability of the model to identify all the relevant instances of a particular class**

IT ANSWERS,
**“Of all the actual instances for a particular class, how many did the model correctly identify?”**

## F-1 Score (≥ 75%)

IT IS A MEASURE THAT
**combines precision and recall into a single metric**

THIS IS PARTICULARLY USEFUL
**when we need to find an optimal balance between precision and recall, especially in situations where there is an uneven class distribution**

## Support

THE NUMBER OF ACTUAL INSTANCES
**of the positive class in the dataset**

THIS PROVIDES NECESSARY CONTEXT
**for these metrics, indicating the number of actual instances of each class, thus helping to understand the reliability and significance of the performance measures**

TRAIN & TEST NLP MODEL

# NLP Model Met/Exceeded Acceptance Criteria

| Concept | True Positives | False Positives | False Negatives | Precision (≥ 80%) | Recall (≥ 70%) | F1 score (≥ 75%) | Support |
|---|---|---|---|---|---|---|---|
| | | | | Acceptance Criteria | | | |
| **ENDS** | 588 | 59 | 165 | 0.90 | 0.78 | 0.84 | 753 |
| ENDS Use Status | 473 | 46 | 85 | 0.91 | 0.84 | 0.87 | 558 |
| ENDS Duration | 114 | 4 | 9 | 0.96 | 0.92 | 0.94 | 123 |
| ENDS Amount | 113 | 7 | 23 | 0.94 | 0.83 | 0.88 | 136 |
| ENDS Frequency | 117 | 29 | 46 | 0.80 | 0.71 | 0.75 | 163 |
| **Combustible Tobacco** | 424 | 80 | 63 | 0.84 | 0.87 | 0.85 | 487 |
| Combustible Tobacco Use Status | 452 | 72 | 97 | 0.86 | 0.82 | 0.84 | 549 |
| **Nicotine Pouch** | 27 | 3 | 5 | 0.90 | 0.84 | 0.87 | 32 |
| Nicotine Pouch Status | 24 | 2 | 6 | 0.92 | 0.80 | 0.85 | 30 |

MODEL FIT

# Key Findings & Summary

### Large EHR sample
enables robust RWE studies on ENDS transition and long-term health outcomes

### NLP algorithm
successful in extracting ENDS use from unstructured clinical notes

### NLP is a valid and effective tool to:
- Identify patients’ exposure to ENDS
- Scale to additional SFTPs in future research

# Thank You!