# Closing Gaps in Real-World Evidence: Unlocking Clinical Value from Unstructured EHR Notes

RWD176

**Jeffrey Brown, Zuzana Dostalova, Mike Temple, Marley Boyd**
TriNetX LLC, Cambridge, MA

## Background

Oncology decision-making increasingly depends on biomarker status — yet the majority of biomarker results are not captured in structured EHR fields[1]. Molecular pathology reports, IHC scores, and genomic test results are delivered as free-text and rarely entered in structured fields, rendering biomarker-stratified RWE analyses impossible without clinical note fact extraction. NSCLC has the highest density of clinically actionable biomarkers of any solid tumor. The TriNetX Clinical Notes Program applies a validated AI-Enabled Extraction Pipeline — combining artificial intelligence with expert human annotation and clinical review — to recover this data at scale across 6 healthcare organizations. See Methods for pipeline details.

## Objectives

The TriNetX AI-Enabled Extraction Pipeline was applied to recover biomarker data from unstructured clinical notes. Human annotation and expert clinical review are embedded throughout the pipeline to ensure accuracy and clinical validity.
This study aimed to:

- Characterize ALK, EGFR, and PD-L1 biomarker testing rates and result distribution among NSCLC patients with CNP biomarker data
- Assess treatment concordance within 6 months of first biomarker test date — including targeted therapy use in biomarker-negative patients

## Methods

**NSCLC cohort identification:** Patients were identified via AI-extracted histology terms (adenocarcinoma, squamous cell carcinoma, non small cell carcinoma; NER confidence ≥0.6) from clinical notes across 6 HCOs. Note availability varies across HCOs by volume, note type, and completeness due to differences in EHR systems, documentation practices, and data-sharing agreements.

**AI-Enabled Extraction Pipeline:** Clinical notes were processed through a six-step validated pipeline combining artificial intelligence with expert human annotation: (1) keyword-based note identification developed by subject matter experts and clinical informaticists through systematic review and iterative recall analysis; (2) initial concept labeling by a pre-trained John Snow Labs NER model; (3) correction by trained annotators against a predefined clinical guideline; (4) performance evaluation against a gold standard corpus of human-annotated documents; (5) iterative model fine-tuning — steps 3–5 are repeated until performance targets are met; (6) full-corpus extraction. Human review is embedded throughout the pipeline steps, distinguishing this approach from fully automated extraction.

**Validation:** NER confidence metric (certainty of concept identification) was assessed. Target precision was ≥0.90; 73% of extracted facts met this threshold. Achieved performance: macro precision 0.96, sensitivity 0.96, F1 0.97. Full-corpus extraction is triggered only after performance targets are met.

**Biomarker status:** An ever-positive approach was applied across ALK, EGFR, and PD-L1 (positive defined as ≥1% TPS for numeric results). Where a patient had both positive and negative records, the earliest positive date was used as the index date.

**Treatment concordance:** Receipt of class-concordant targeted therapy was assessed within 6 months of first biomarker test date.

**Figure 2.** Biomarker testing rates among NSCLC patients with clinical notes (N=7,832). Bars show percentage of patients with biomarker tested (blue) and positivity rate among tested (orange). ALK: 32.9% tested, 11.1% positive. EGFR: 37.3% tested, 25.5% positive. PD-L1: 93.2% tested, 77.4% positive at ≥1% TPS (44.0% high expression ≥50% TPS).

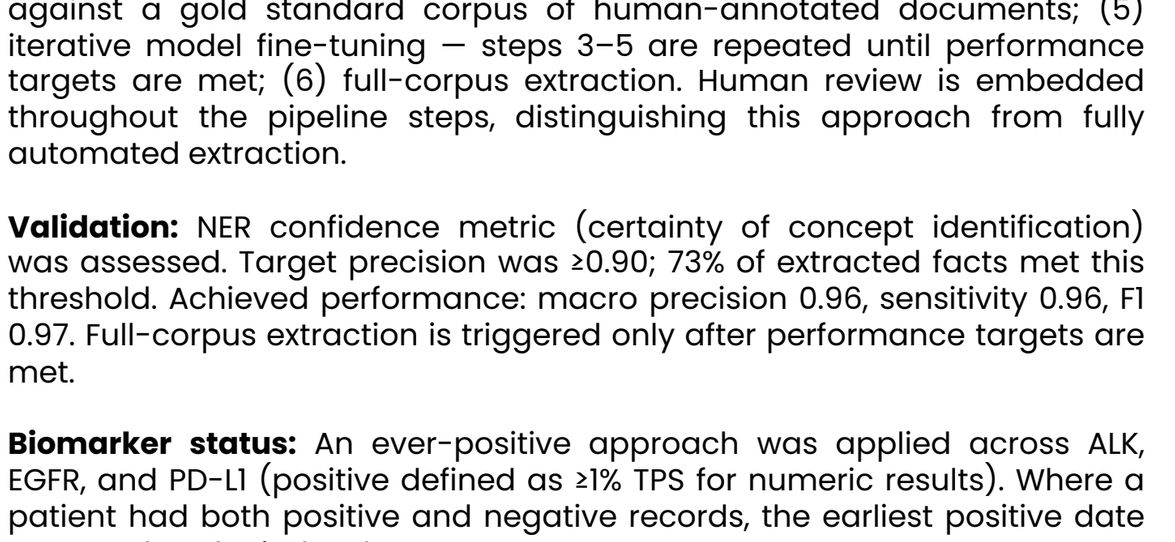
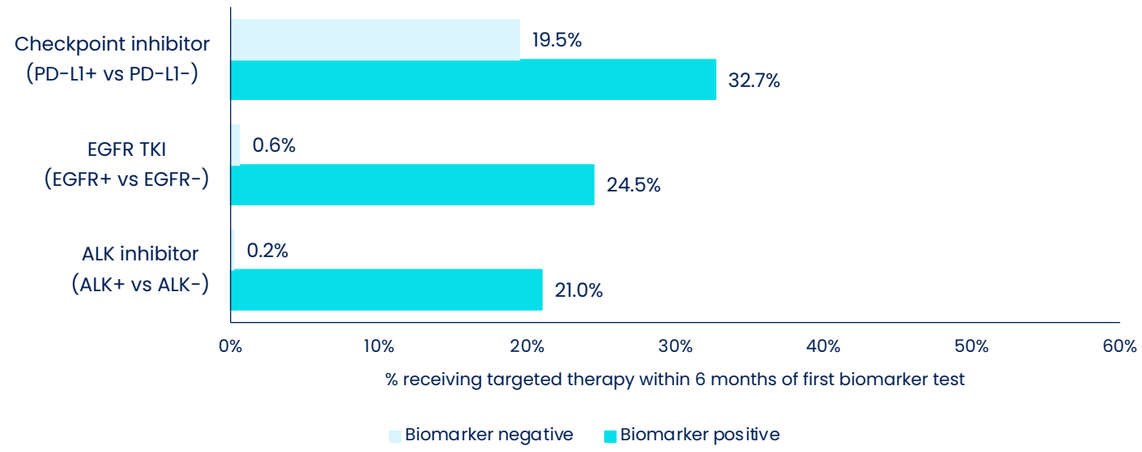



**Figure 3.** Treatment concordance rates within 6 months of first biomarker test date by biomarker status. Bars show percentage receiving concordant targeted therapy. ALK+: 21.0% vs ALK−: 0.2%. EGFR+: 24.5% vs EGFR−: 0.6%. PD-L1+ (≥1% TPS): 32.7% (high ≥50%: 40.5%; low 1–49%: 32.4%) vs PD-L1−: 19.5%.




*PD-L1-negative checkpoint inhibitor use (19.5%) is consistent with NCCN-approved combination regimens that do not require PD-L1 positivity.

## Results

The dataset comprised of 7,741,982 patients with clinical notes. Among 28,256 NSCLC patients identified via AI-extracted histology, 7,832 (27.7%) had biomarker data available. Testing rates among biomarker-eligible patients were: ALK 32.9%, EGFR 37.3%, PD-L1 93.2%. Positivity rates among tested patients were: ALK 11.1%, EGFR 25.5%, PD-L1 77.4% — of whom 44.0% showed high expression (≥50% TPS).

**Treatment concordance within 6 months of first biomarker test:**

- ALK-positive: 21.0% received an ALK inhibitor vs. 0.2% of ALK-negative patients
- EGFR-positive: 24.5% received an EGFR tyrosine kinase inhibitor vs. 0.6% of EGFR-negative patients
- PD-L1-positive: 32.7% received a checkpoint inhibitor (high expressors ≥50% TPS: 40.5%; low expressors 1–49%: 32.4%) vs. 19.5% of PD-L1-negative patients — consistent with NCCN-approved combination regimens that do not require PD-L1 positivity

**Table 1.** Biomarker testing rates and treatment concordance in NSCLC patients with CNP biomarker data (N=7,832). Concordance defined as receipt of class-concordant targeted therapy within 6 months of first biomarker test date.

| Biomarker | Tested, n (%) | Positive, n (%) | Negative, n (%) | Concordant treatment, % (n treated) [biomarker positive] | Targeted therapy, % (n treated) [biomarker negative] |
|---|---|---|---|---|---|
| PD-L1 (≥1% TPS) | 7,299 (93.2%) | 5,650 (77.4%) | 1,649 (22.6%) | 1,845 (32.7%) | 322 (19.5%) |
| EGFR | 2,918 (37.3%) | 743 (25.5%) | 2,175 (74.5%) | 182 (24.5%) | 13 (0.6%) |
| ALK | 2,577 (32.9%) | 286 (11.1%) | 2,291 (88.9%) | 60 (21.0%) | 5 (0.2%) |

Note: Targeted therapy use in biomarker-negative patients is not treatment concordance. PD-L1-negative: 19.5% (n=322) consistent with NCCN-approved combination regimens not requiring PD-L1 positivity. ALK-negative: 0.2% (n=5) and EGFR-negative: 0.6% (n=13) may represent off-label prescribing or data capture artefacts.

## Conclusion

TriNetX AI-Enabled Extraction Pipeline, combining artificial intelligence with expert human annotation, uncovers clinically meaningful data absent from structured EHR data. ALK and EGFR testing remains incomplete despite guideline mandates, while targeted therapy uptake in biomarker-positive patients remains suboptimal — fewer than 30% receive a class-concordant therapy within 6 months. PD-L1-negative checkpoint inhibitor use (19.5%) is consistent with NCCN-approved combination regimens, demonstrating CNP data can distinguish clinically justified from potentially discordant treatment patterns. These findings establish that unstructured clinical notes contain research-grade data essential for oncology at scale. Ongoing development will extend this approach to breast cancer (ER, PR, HER2, Ki-67, BRCA1/2) and colon cancer (MSI, KRAS/NRAS) biomarkers.

Download the Poster as a PDF

**References**
[1]2024 IASLC global survey: https://www.iaslc.org/science-research/scientific-projects/iaslc-global-survey-biomarker-testing-lung-cancer-2024-update