PP01

# The Importance of Information Extraction from Unstructured Clinical Data in Pharmacoepidemiology

Dena Jaffe[1], Elise Berliner[2], Ace Vo[3], Hasham Ul Haq[3], David Talby[3], Michael Chu[4]

[1] Oracle Health, Petah Tikva, Israel; [2] Oracle Life Sciences, Kansas City, MO, USA; [3] John Snow Labs, Lewes, DE, USA; [4] Children's Hospital of Orange County, Orange, CA, USA;

## Background

Electronic health records (EHRs) and claims are important sources of real-world data used to generate real-world evidence on the safety and effectiveness of therapies. Valuable information is contained in the unstructured clinical notes and methods such as Natural Language Processing (NLP) are needed to extract the information into a structured format for analysis. Previous work focused on developing NLP methods to extract suicidality.[1,2] However, methods to extract information on a broader array of neuropsychiatric symptoms are needed for drug safety studies and other health care use cases.

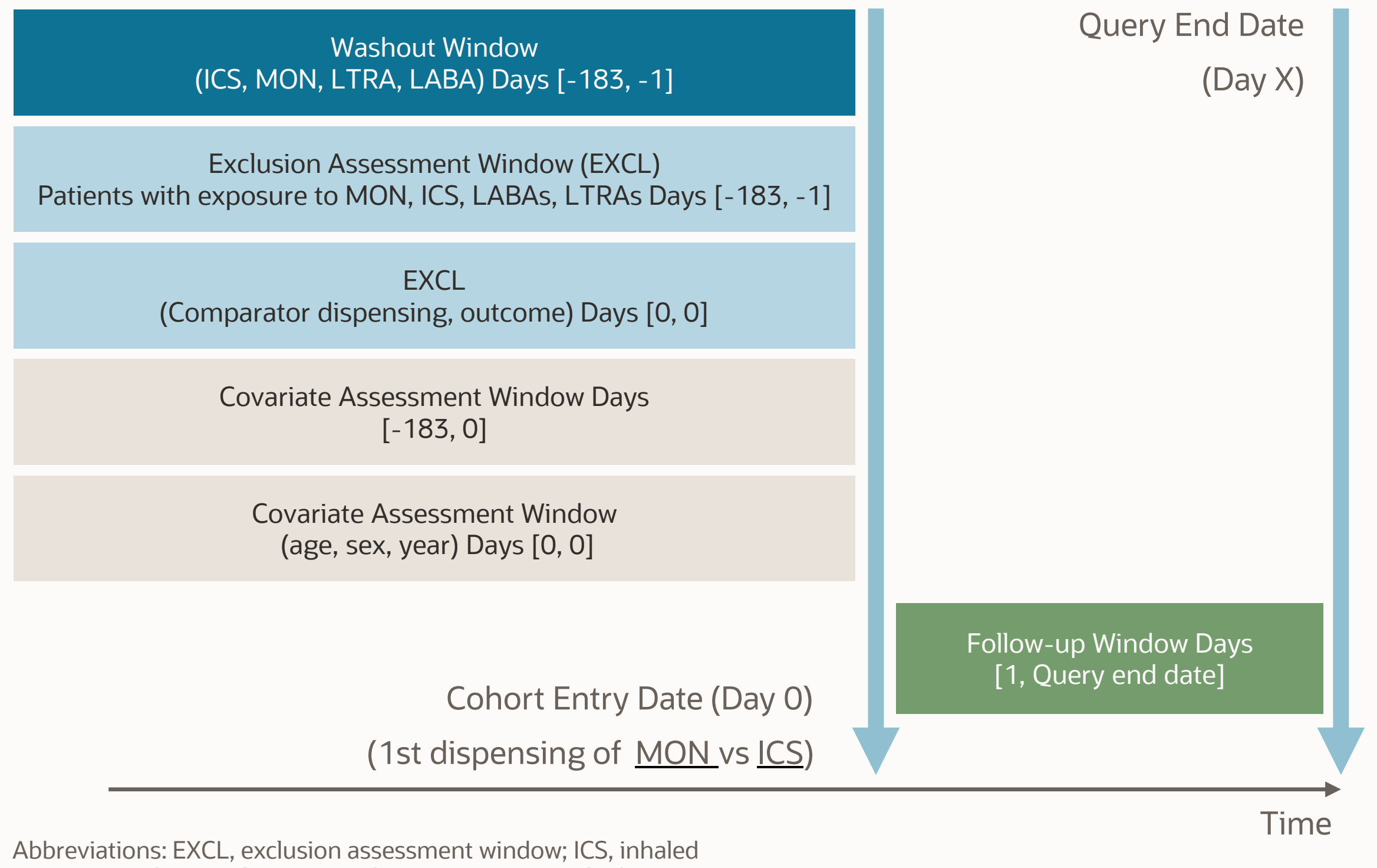
## Objective

To examine the impact on outcome identification of using unstructured EHRs in a drug safety study examining neuropsychiatric events.

## Methods

This retrospective study examined structured and unstructured data from the Oracle EHR Real-World Data (OERWD) linked to a national US claims data source during the study period 2015 to 2022.

**Figure 1. Study Design**


Washout Window (ICS, MON, LTRA, LABA) Days [-183, -1]

Exclusion Assessment Window (EXCL) Patients with exposure to MON, ICS, LABAs, LTRAs Days [-183, -1]

EXCL (Comparator dispensing, outcome) Days [0, 0]

Covariate Assessment Window Days [-183, 0]

Covariate Assessment Window (age, sex, year) Days [0, 0]

Follow-up Window Days [1, Query end date]

Query End Date (Day X)

Cohort Entry Date (Day 0) (1st dispensing of MON vs ICS)

Time

Abbreviations: EXCL, exclusion assessment window; ICS, inhaled corticosteroid; LABA, long-acting beta agonist; LTRA, leukotriene receptor antagonist; MON, montelukast

The cohort was defined as patients with a diagnosis code of asthma in either claims or electronic health record structured data and a new prescription of montelukast or inhaled corticosteroids. Prior treatment episodes were considered using a 30-day gap period (Fig 1). The covariate assessment window was the 183 days up to and including the date of prescription. For this analysis, outcomes were measured from 1 day after the prescription until query end date.

Outcomes were neuropsychiatric events based on the FDA boxed warning for montelukast (Fig 2).[3] Outcomes from structured data were ascertained using diagnosis codes, hospitalization and emergency room codes, and dispensed treatments. Outcomes from unstructured data were ascertained through named entity recognition models from John Snow Labs. Guidelines for annotation were developed based on the clinical concepts in the boxed warning for montelukast and refined with the input of clinicians and trained annotations.[4] The models were trained in four rounds, with increasing amounts of training data and enrichment of notes with mentions of rare events.

Three analytic approaches were used to examine the value of incrementally contributing sources of data for covariates and outcomes:

Analysis 1: Claims only

Analysis 2: Claims and structured EHR data

Analysis 3: Claims and structured and unstructured EHR data.

Propensity scores using logistic regression models were used to match montelukast initiators to the ICS referent group using a 1:1 ratio and nearest neighbor matching algorithm for each analysis group. For this study, results from the total matched cohort are presented. Statistical analyses were performed using R version 4.1.

For more information on the methods, see the protocol for the study: https://www.sentinelinitiative.org/sites/default/files/documents/MOSAIC-NLP_Protocol_v1.3.pdf

**Figure 2. Outcomes of Interest: Neuropsychiatric Events**

**FDA's Boxed Warning**

- Agitation, including aggressive behavior or hostility
- Attention problems
- Bad or vivid dreams
- Depression
- Disorientation or confusion
- Feeling anxious
- Hallucinations
- Irritability
- Memory problems
- Obsessive-compulsive symptoms
- Restlessness
- Sleepwalking
- Stuttering
- Suicidal thoughts and actions
- Tremor or shakiness
- Trouble sleeping
- Uncontrolled muscle movements

**Claims/EHR Structured Data**

Hospitalization AND/OR Emergency Department utilization OR Diagnosis AND/OR Treatment of

- Depression
- Self harm
- Psychotic disorder
- Mood disorder
- Anxiety disorder
- OCD
- Manic or bipolar disorder
- Personality disorder
- Hyperactivity or aggressive behavior or harm

Diagnosis OR Treatment of Sleep Disorder

- Insomnia
- Hypersomnia
- Circadian rhythm disorder
- Parasomnia
- Movement disorder
- Other undefined sleep disorder

**EHR Unstructured Data**

- Aggressive behavior or hostility
- Agitation
- Attention problems
- Bad or vivid dreams
- Depression
- Disorientation or confusion
- Dream abnormalities
- Feeling anxious
- Hallucinations
- Irritability
- Memory problems
- Obsessive-compulsive symptoms
- Restlessness
- Sleepwalking
- Stuttering
- Suicidal thoughts and actions
- Tremor or shakiness
- Trouble sleeping
- Uncontrolled muscle movements

## Results

A total of 109,076 patients with asthma who initiated montelukast or inhaled corticosteroids from 112 health systems were examined. Demographic characteristics of the propensity-matched patients are presented in Table 1.

**Table 1. Demographic Characteristics of the Overall Matched Cohort According to Data Source**

| | Claims | Claims + Structured EHR | Claims + Structured EHR + Unstructured EHR |
|---|---|---|---|
| **Number of patients** | 76,016 | 71,620 | 71,244 |
| **Age at treatment initiation, years, mean (SD)** | 29.7 (20.9) | 29.7 (20.7) | 29.8 (20.8) |
| **Female, %** | 61.2% | 61.0% | 61.0% |
| **Married or living with a partner, %** | n/a | 19.3% | 19.4% |
| **Race, %** | | | |
| Asian, or American Indian or Alaska Native, or Native Hawaiian or Other Pacific Islander | n/a | 3.0% | 3.0% |
| Black or African American | n/a | 20.2% | 20.1% |
| White | n/a | 57.7% | 58.0% |
| Multiple races/other | n/a | 12.8% | 12.7% |
| Missing | n/a | 6.3% | 6.2% |
| **Ethnicity, %** | | | |
| Hispanic or Latino | n/a | 24.3% | 24.2% |
| Non-Hispanic or Latino | n/a | 68.1% | 68.2% |
| Multiple ethnicities/missing | n/a | 7.6% | 7.6% |
| **Insurance status, %** | | | |
| Commercial | 30.5% | 30.1% | 30.3% |
| Medicaid/Medicare | 68.1% | 68.5% | 68.4% |
| Other/missing | 1.3% | 1.4% | 1.4% |

Abbreviations: n/a, not available; SD, standard deviation

**Table 2. Neuropsychiatric Events in the Overall Matched Cohort According to Data Source**

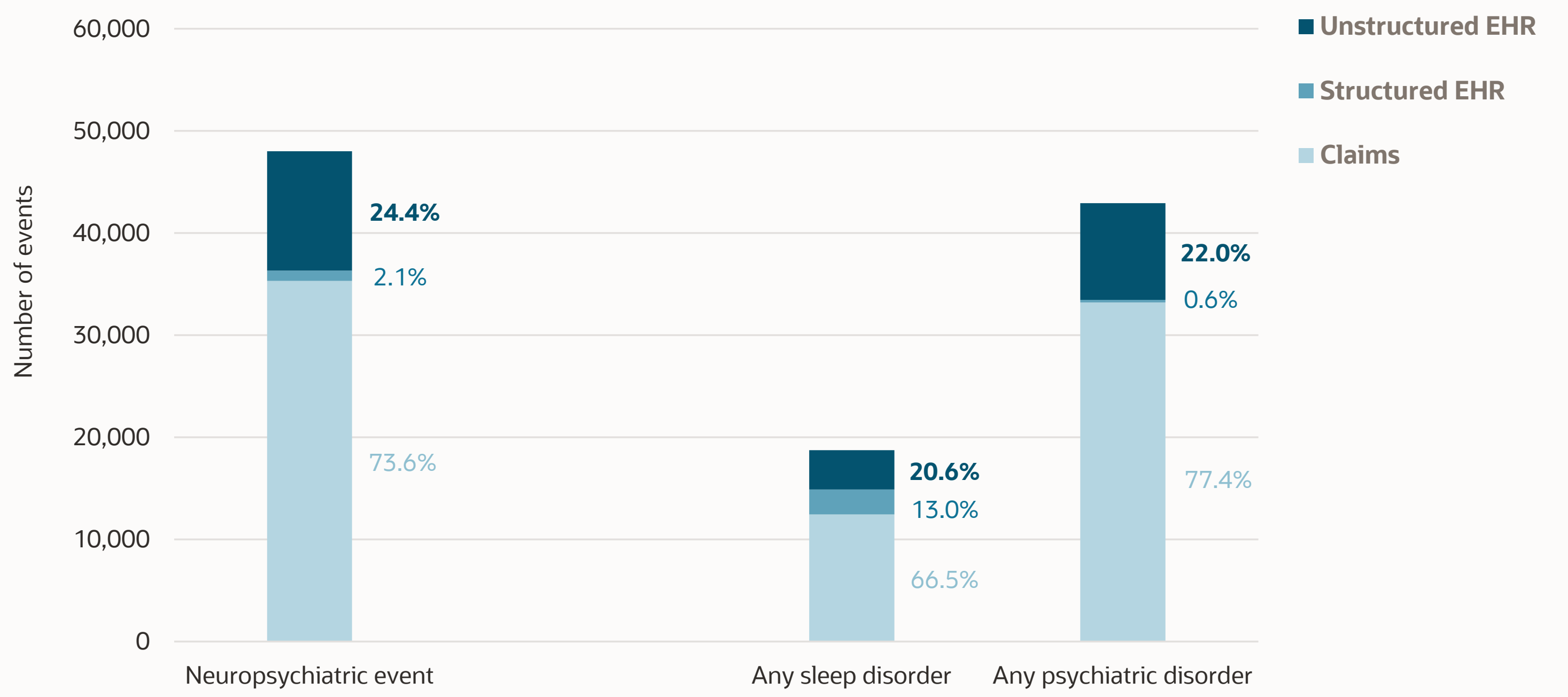
| | Claims | Claims + Structured EHR | Claims + Structured EHR + Unstructured EHR |
|---|---|---|---|
| **Number of patients** | 76,016 | 71,620 | 71,244 |
| **Events per person** | | | |
| Mean (SD) | 2.53 (1.53) | 2.64 (1.67) | 2.62 (1.73) |
| Median (IQR) | 2 (1-3) | 2 (1-4) | 2 (1-4) |

Abbreviations: IQR, interquartile range; SD, standard deviation

Matched study patients had 2.5 events/person when utilizing structured data to identify outcomes, and 2.6 events/person with the addition of unstructured clinical notes (Table 2).

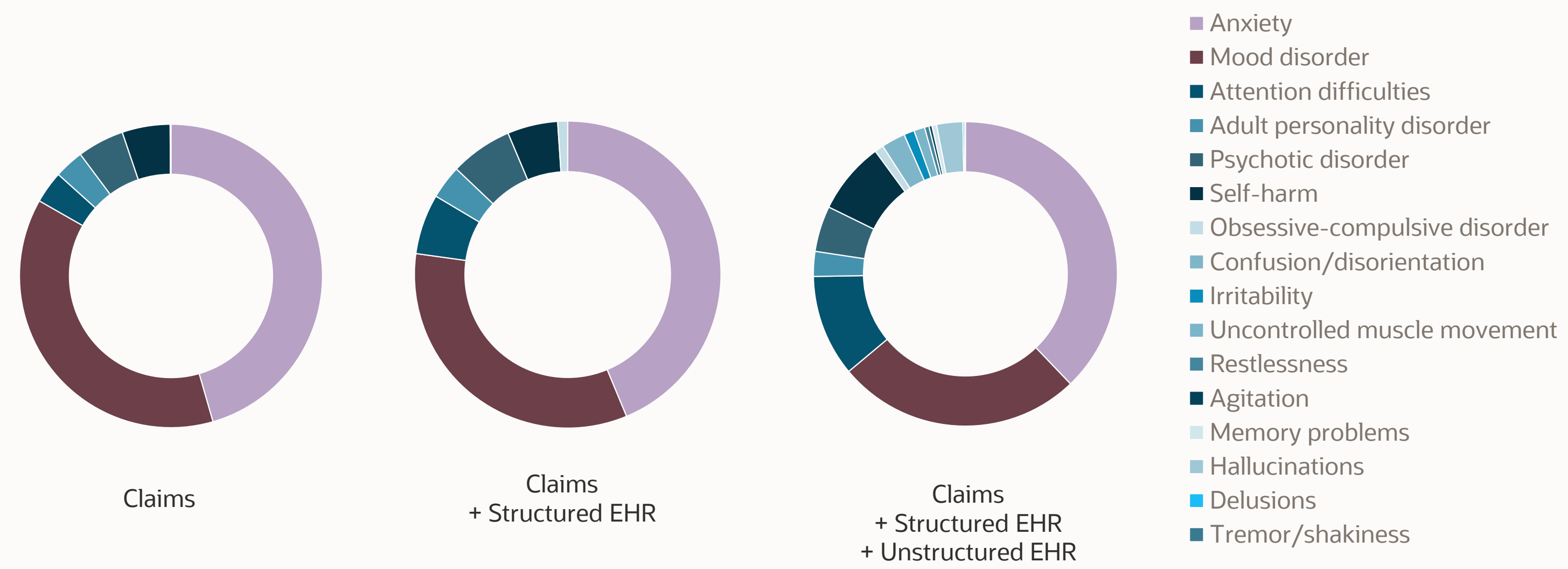
Figure 3 presents the contribution of neuropsychiatric outcomes according to data source in the final analysis that included claims and structured and unstructured EHR data. Compared to outcomes identified from claims only, adding structured EHR data resulted in only a modest increase in numbers of events identified for neuropsychiatric events, with the majority of the additional events being sleep disorders. Unstructured data added an additional 20%+ of outcome events.

**Figure 3. Additional Contribution of Neuropsychiatric Events in the Overall Matched Cohort Analysis to Claims Data Using Structured and Unstructured EHR Data**



Anxiety and mood disorder were the most frequently documented neuropsychiatric events in all sources of data. Many events, including agitation, muscle problems, hallucinations, and delusions were not identified at all in the structured data (Fig 4).

**Figure 4. Distribution of Neuropsychiatric Events in the Overall Matched Cohort by Data Source**



## Conclusion

This study found that neuropsychiatric events may be undercounted using only structured data from EHR and claims, as the number of observed suicidality/self-harm events doubled with the addition of unstructured EHR data. Further, events such as irritability, agitation, and memory problems were only detected in unstructured data. This study illustrates the importance of unstructured data especially related to mental health outcomes.

This method is limited by the time required to annotate training data and the model's ability to identify and train on rare events, such as stuttering. Future work using large language models and hybrid methods may be able to overcome these limitations.

## References

[1] Young et al 2023. https://doi.org/10.1016/j.jadr.2023.100507

[2] Haerian et al 2012. https://pmc.ncbi.nlm.nih.gov/articles/PMC3540459/

[3] FDA. Accessed April 17, 2023. https://www.fda.gov/drugs/drug-safety-and-availability/fda-requires-boxed-warning-about-serious-mental-health-side-effects-asthma-and-allergy-drug

[4] Mosaic-NLP 2024. https://www.sentinelinitiative.org/sites/default/files/documents/MOSAIC-NLP_AnnotationGuidelines_v1.0_0.pdf

Presented at PHUSE CSS 2025 Utrecht, the Netherland 20–21 May

**Funding Statement**:

This project was supported by Task Order 75F40119F19002 under Master Agreement 75F40119D10037 from the US Food and Drug Administration (FDA). FDA coauthors reviewed the study protocol, statistical analysis plan, and the manuscript for scientific accuracy and clarity of presentation. Representatives of the FDA reviewed a draft of the manuscript for presence of confidential information and accuracy regarding statement of any FDA policy. The views expressed are those of the authors and not necessarily those of the US FDA.