Leveraging AI-driven natural language processing to enhance symptom capture in primary biliary Cholangitis (PBC) and primary Sclerosing Cholangitis (PSC)

RWD#129

Anthony D Perez, PhD[1] **Ashley Jaksa, MPH**[1]

[1]**Pedestal Health**

Background and Objectives

- Fatigue and pruritus are hallmark symptoms experienced by patients with cholestatic liver diseases, Primary Biliary Cholangitis (PBC) and Primary Sclerosing Cholangitis (PSC). These symptoms have a negative impact on patient quality of life, work impairment and disease burden.
- Despite their clinical relevance, these symptoms are often under-captured in structured real-world data (RWD), such as Electronic Health Records (EHR).
- Unstructured clinical narratives (e.g., provider notes) frequently contain rich symptom descriptions that often go unquantified in traditional RWD analysis. Natural Language Processing (NLP) offers an approach to unlocking this latent information at scale.
- This study evaluated the extend to which NLP of unstructured clinical narratives can improve real-world prevalence estimates for:
  - Pruritus - in PBC and PSC cohorts
  - Fatigue - in PSC cohort
- We compare structured EHR code-based identification alone against a multimodal approach integrating both structured and unstructured data.

Methods

- Patients with PBC (ICD10: K74.3) and PSC (ICD10: K83) were identified within the TARGET-LD and TARGET-GASTRO cohorts
- Baseline symptom prevalence was calculated using structured EHR data (e.g., diagnosis codes, problem lists, coded clinical findings).
- For patients with available unstructured clinician notes, a section-aware NLP pipeline was deployed. The NLP accounted for EHR vendor (Epic, Cerner, and others) and health-care-network-specific note architecture. Symptom-specific sections (e.g., Past Medical History, History of Present Illness, etc) were targeted: pruritus and fatigue were detected and assigned a polarity score (affirming vs. negating; Box 1).
- NLP-derived insights were integrated with structured data to generate multimodal prevalence estimates.
- Multimodal estimates compared against code-only identification to quantify the incremental yield of unstructured data.

Box 1: Examples of of affirmation and negation of symptoms

AFFIRMING Pruritus - "...ongoing jaundice, pruritus, weight loss that did not respond to cholestyramine..." "...saw his PCP with c/o increasing pruritus and yellow eyes... main complaint is the itching..."

NEGATING Pruritus - "...she has not had jaundice, pruritus, but has had some stabbing right upper quadrant pain..." "...she denied any pain today or pruritus..."

Results:

- Preliminary analyses included 668 patients with PBC and 3,958 with PSC.
- Relying solely on structured data yielded low apparent prevalence for pruritus (11% PBC; 17% PSC) and fatigue (20% PSC). The integration of unstructured data significantly increased capture across both cohorts. Multimodal prevalence for pruritus rose to 37% in PBC and 53% in PSC. In the PSC cohort, fatigue identification increased to 66% through the addition of NLP-derived insights.

Conclusions

- Diagnostic codes alone substantially underestimated the real-world burden of pruritus and fatigue in PBC and PSC, by a factor of 3 or more in these analysis.
- NLP-processed unstructured data markedly increased the sensitivity of symptom identification without requiring additional data collection or patient contact.
- This multimodal approach provides a more robust framework for understanding the natural history and real-world burden of symptoms in cholestatic liver diseases.
- These findings have direct implications for drug developers, payers, HTA bodies, providers and patients evaluating the burden of illness and treatment benefits in PBC and PSC.
- Structured EHR data can be treated as a floor, not a ceiling for symptom prevalence where provider documentation patterns drive under capture.

Accurate symptom burden estimates are foundational to understanding the value of new therapeutics. This multimodal methodology is directly applicable across TARGET-LD and TARGET-GASTRO to support evidence generation for novel interventions targeting PBC and PSC symptom relief — providing real-world context that complements clinical trial endpoints."

Pedestal Health

PATIENT POPULATIONS · PRELIMINARY ANALYSIS

PBC Cohort

668

patients

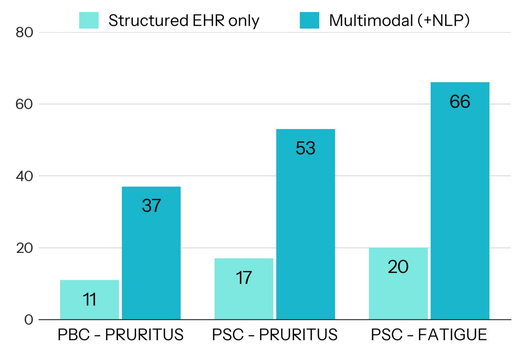
Disease: Primary Biliary Cholangitis
ICD-10: K74.3
Data Source: TARGET-LD & TARGET-GASTRO
Symptom studied: Pruritus

PSC Cohort

3,958

patients

Disease: Primary Sclerosing Cholangitis
ICD-10: K83.0
Data Source: TARGET-LD & TARGET-GASTRO
Symptom studied: Pruritus & Fatigue



PBC · Pruritus
STRUCTURED EHR ONLY
11%
MULTIMODAL (+ NLP)
37%
+26 pp percentage point increase · 3.4× increase

PSC · Pruritus
STRUCTURED EHR ONLY
17%
MULTIMODAL (+ NLP)
53%
+36 pp percentage point increase · 3.1× increase

PSC · Fatigue
STRUCTURED EHR ONLY
20%
MULTIMODAL (+ NLP)
66%
+46 pp percentage point increase · 3.3× increase

Acknowledgements and Disclosures: AP and AJ are employees of Pedestal Health. Pedestal Health is a health evidence solutions company headquartered in Durham, NC. An artificial intelligence tool was utilized to assist the authors in the writing process, specifically for spelling verification and sentence reformulation to enhance clarity in English.

ISPOR: Philadelphia, PA May 17-May 20, 2026