

From Unstructured Clinical Notes to Actionable Insights: A Workflow for Standardising Critical ECOG Performance Status for OMOP-Based Real-World Oncology Research.

Unlocking the Patient's Functional Trajectory: Precision Rule-Based Text-Mining and Analysis-Ready Data on Vital Status for 11,833 Patients.

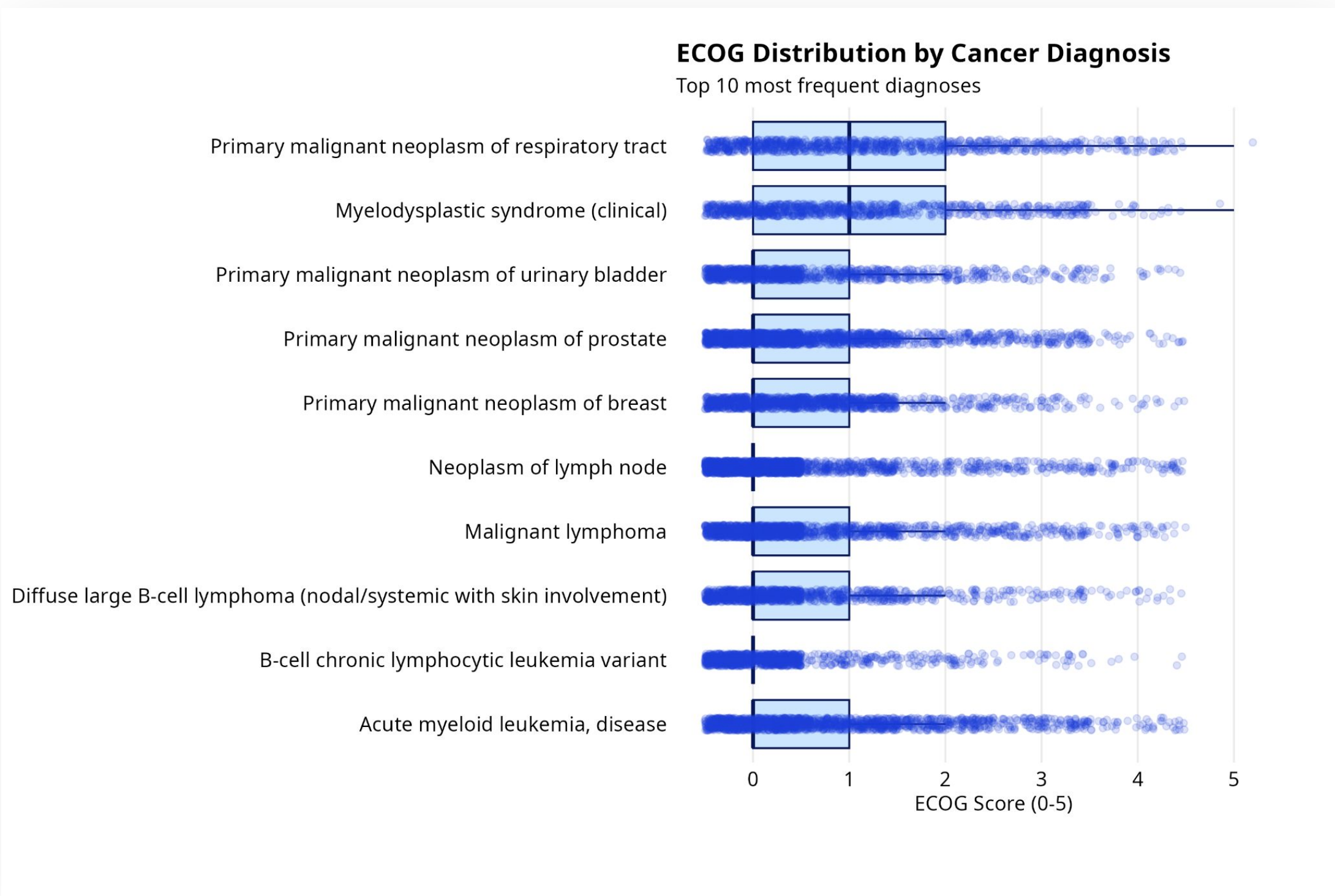
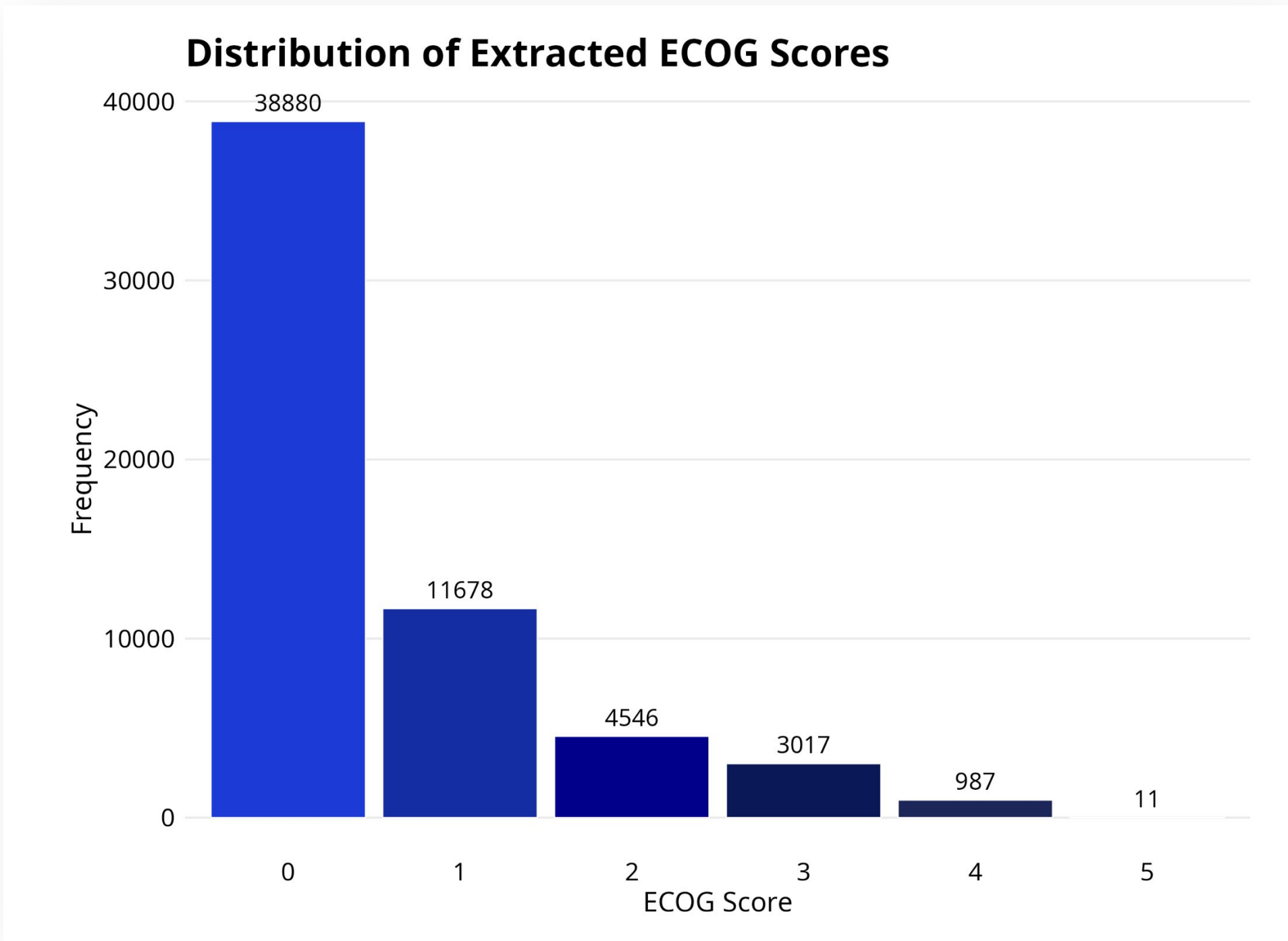
From Free Text to OMOP-Ready ECOG scores: Standardizing Performance Status with Rule-Based Text-Mining Pipeline in 14 Million Clinical Annotations

Background: ECOG Performance Status is a clinician-rated scale (from 0 to 5) used in oncology to describe a patient’s overall functional status, specifically how a disease affects daily living abilities, self-care, and physical activity, and is vital for oncology. These scores are often buried in unstructured clinical notes. We developed a scalable pipeline to extract, clean these scores to map to the **OMOP CDM**, overcoming challenges like ranges, typos, and cumulative note duplication.

Results

Pipeline Performance: Successfully recovered **59,119** valid mentions from **11,833** unique patients.
Score Distribution: Most patients were rated ECOG 0 or 1, providing a baseline for functional status in our cancer cohort.

Clinical Insights by Diagnosis: Patients with respiratory tract neoplasms showed the highest functional impairment (mean 1.25), while chronic lymphocytic leukaemia patients showed the lowest (mean 0.30). The most frequent mentions were for lymph nodes, prostate, and acute myeloid leukemia.



Methods

The extraction process consisted of four stages:

- 1 Extraction Logic & Range Resolution:** robust patterns to identify ECOG sequences (e.g., "ECOG 1", "ECOG 1-2") while handling typographical errors like "COG PS" and non-standard formats.
- 2 Data Cleaning:** We implemented a *slice_tail* strategy, keeping the latest clinical observation performed on the same day.
- 3 Temporal Validation:** Grouped records into 5-day windows to collapse cumulative clinical notes. This logic effectively reduced noise from repeated assessments during the same clinical episode.
- 4 OMOP Standardization:** Mapped the curated scores to the Measurement domain of the OMOP CDM, ensuring international interoperability and analysis-readiness.

Conclusions

We successfully standardised unstructured ECOG data into the OMOP CDM, providing a high-resolution view of patient functional status. Addressing EHR redundancy via temporal clustering was essential for data quality. The distribution aligns with clinical expectations for oncology outpatients. Future work will integrate these scores with treatment response data and survival outcomes within the OHDSI framework.



Angela Leis, Carla Arxé, Juan Manuel Ramírez-Anguita, Antonio Arcas, Jan Carreras, Ricard García-Isern, Miguel-Angel Mayer