

- Enabling Large-Scale Observational Research on Neonatal Outcomes by Standardising Unstructured APGAR Data into the OMOP CDM
- Building High-Quality, Research-Ready Pregnancy Episode Cohorts by Overcoming EHR Text Redundancy and Format Heterogeneity

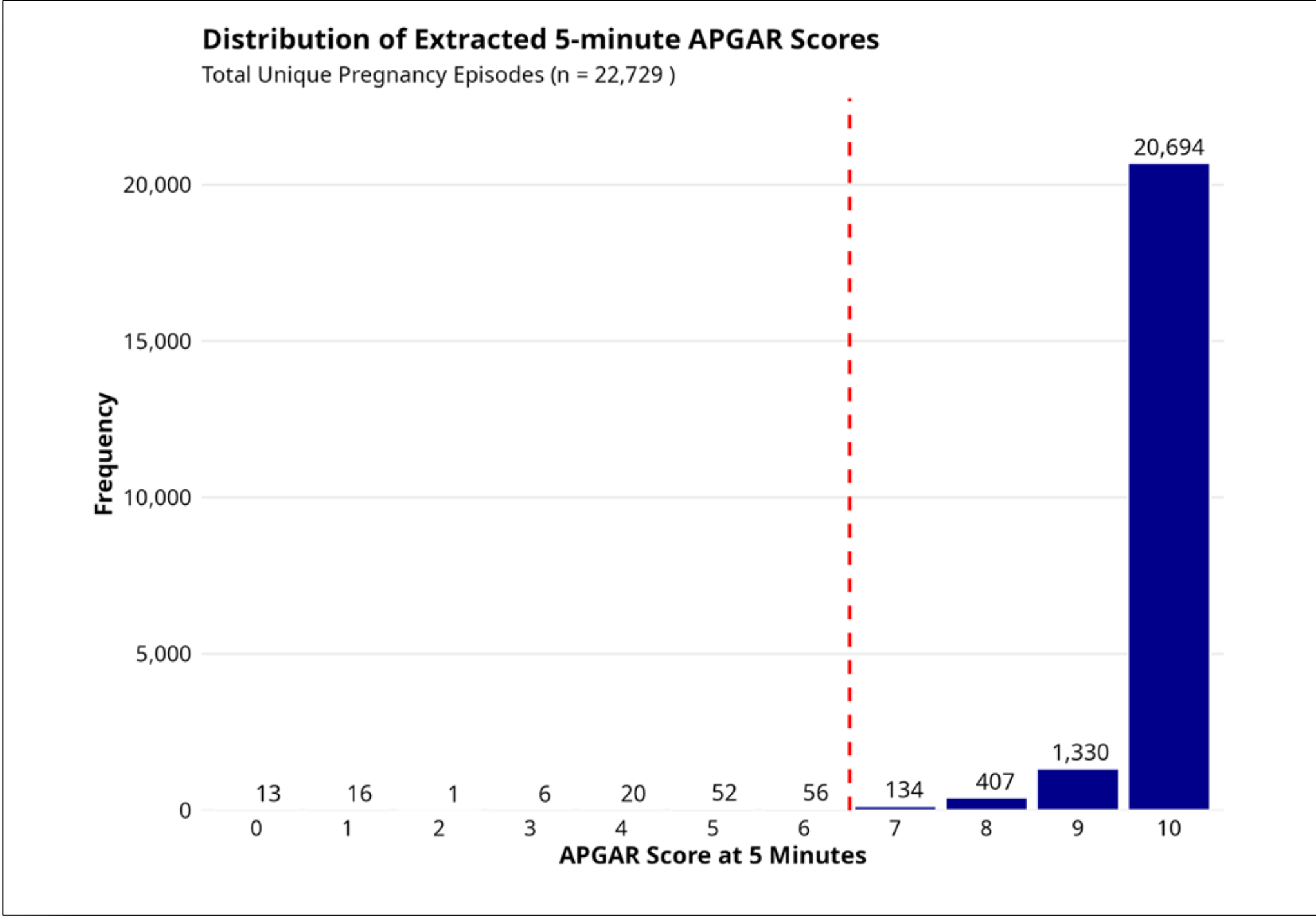
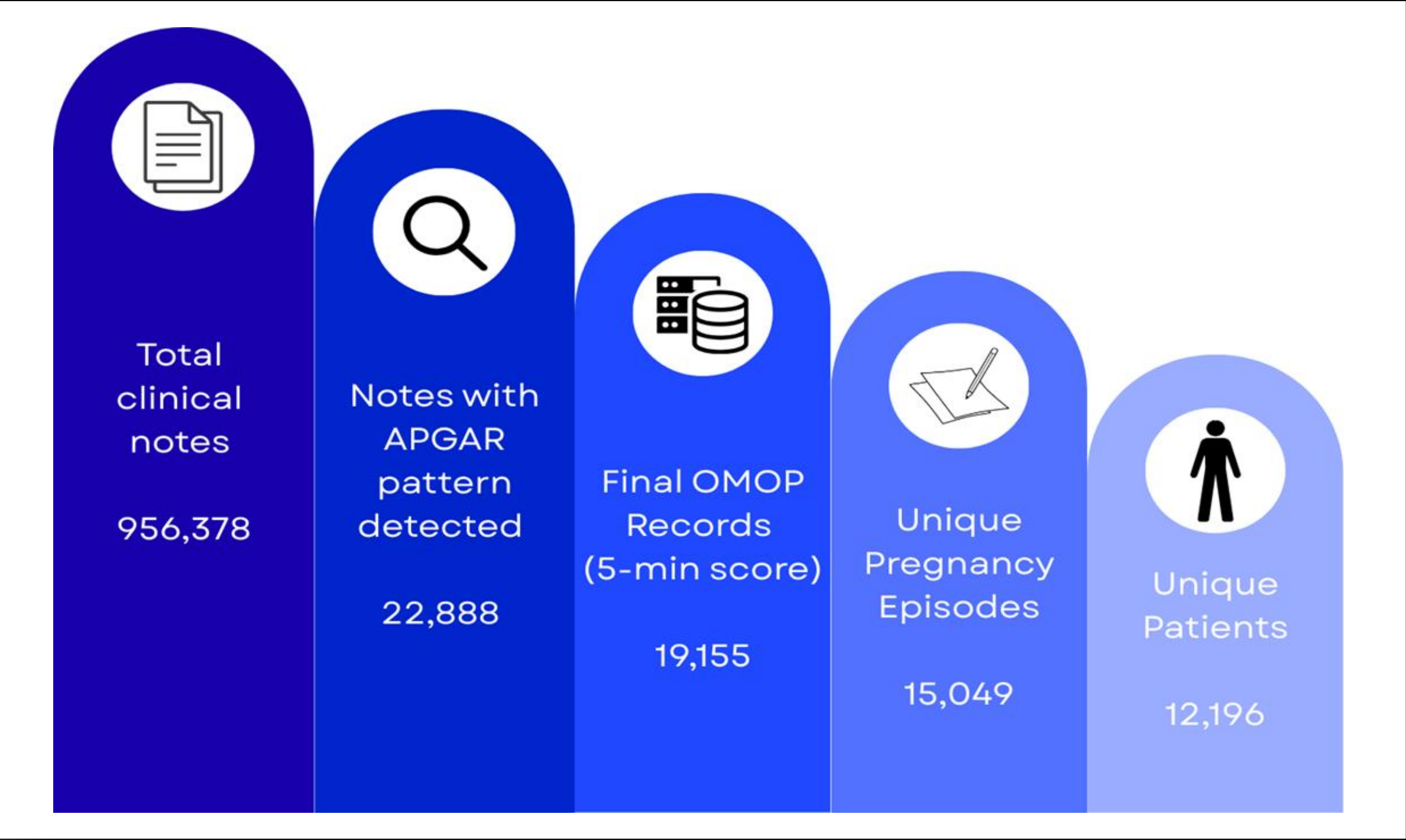
Scalable Extraction of APGAR Scores from Electronic Health Records Clinical Annotations and Standardization for OMOP CDM

Background: The APGAR is a quick, standardised assessment used right after birth to sum up how a newborn is doing and whether immediate support is needed. The score is based on five observable signs: appearance (skin colour), pulse (heart rate), grimace (reflex irritability in response to stimulation), activity (muscle tone), and respiration (breathing effort). Each component is rated 0, 1, or 2, and the total ranges from 0 to 10. In practice, a total of 7–10 is generally reassuring. To overcome the barrier of unstructured clinical text, we developed a scalable pipeline to extract and standardise it into the OMOP CDM.

Results

We processed 950,000 clinical notes, successfully filtering redundant documentation. Our temporal clustering logic (150-day window) identified 15,049 biologically distinct pregnancy episodes, covering the period 2008 to 2024.

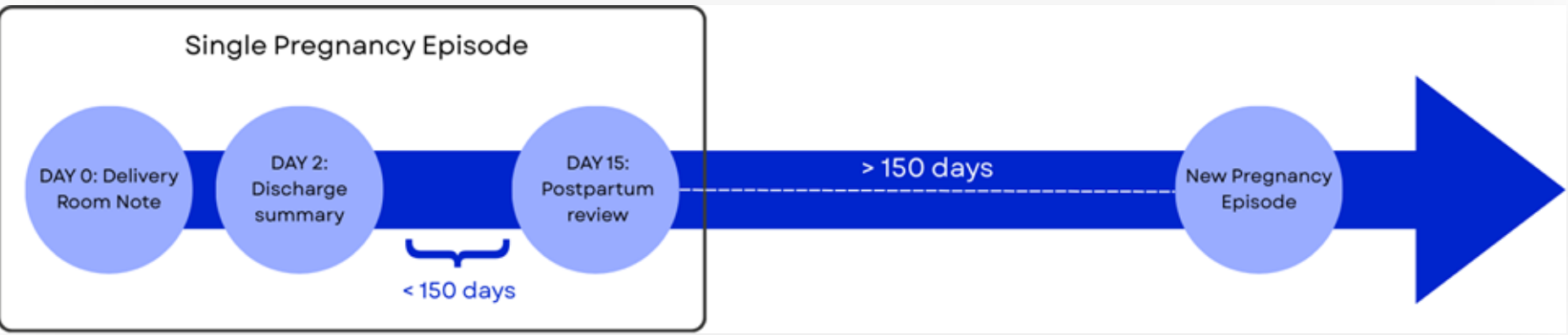
The resulting distribution of 5-minute APGAR scores aligns with established neonatal population norms. The left-skew (majority of scores ≥ 7) serves as an internal validation of the extraction accuracy and clinical plausibility of our rule-based engine.



Methods

- The extraction process consisted of four stages:
- 1 Regex Extraction & Parsing:** Developed robust Regular Expression patterns to identify APGAR sequences (e.g., "9/10/10") while handling heterogeneous formats and common character/OCR errors.
 - 2 Handling Multiple Births:** Detected linguistic markers (e.g., "twins", "fetus 2") to extract multiple sets of scores and preserve sibling relationships within the same clinical visit.

- 3 Episode Definition (Temporal Clustering):** Grouped records into 150-day windows to collapse redundant notes into unique and valid pregnancy episodes.
- 4 OMOP Standardization:** Mapped scores to the measurement table and the Perinatal Extension (PET) of the OMOP CDM to ensure interoperability and mother-child traceability.



Conclusions

We successfully standardised unstructured APGAR data into the OMOP CDM with 99.6% coverage. A temporal clustering logic was essential to eliminate EHR redundancy and accurately define unique pregnancy episodes. Future work will compare this rule-based pipeline against Large Language Models (LLMs) to improve edge-case extraction and provide an automated validation layer.



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