

The ER Dataset for AI Training

Physician-Validated Curation for High-Performance Clinical AI

1. Dataset Overview

The ER Dataset provides 32,667 physician-verified clinical records specifically curated to eliminate noise, bias, and irrelevant metadata from medical literature, making it the ideal training set for LLMs and classifiers. Specialties include: Oncology (790), Cardiovascular (2,158), Sepsis (1,122), Trauma (29,387), Stroke, and Toxicology.

2. The 10 Curation Rules That Filter the Noise

1. ER Applicability Score - Scored 0-10 for emergency department utility.
2. Guideline Alignment - Flags deviations from ACLS-ATLS standard-of-care guidelines.
3. Statistical Integrity - Enforces cohort size ≥ 30 and p-value < 0.05 .
4. Outcome Relevance - Prioritizes patient-centered outcomes over surrogate biomarkers.
5. Bias Detection - Identifies and flags industry-sponsored or single-center studies.
6. Clinical Plausibility - Flags extreme or implausible statistical claims.
7. Actionability - Categorizes records by workflow priority (STAT vs. Routine).
8. Evidence Grade - Grades quality from A (Meta-Analysis) to F (Adverse Events).
9. Population Fit - Matches study demographics to emergency patient cohorts.
10. Recency Weight - Applies time-decay weight favoring articles from the last 5 years.

3. Native Formats & Specifications

- Standard CSV: Simple tabular data import for machine learning pipelines.
- JSON: Rich nested schema containing individual rule breakdowns and physician feedback.
- UDS (Universal Document): Cryptographically sealed file format ensuring end-to-end data integrity.

4. Licensing & Pricing

- Mini (Free Sample): 10 fully-reviewed records for evaluation.
- Starter (\$2,000): Complete dataset, standard commercial usage, single seat license.
- Growth (\$3,500): Unlimited commercial license, multi-seat team access.