Journal of Medical Systems (2026) 50:162
https://doi.org/10.1007/s10916-026-02486-y

RESEARCH

Decoding Complications: Large Language Model-Based Classification of Free-Text Entries in Clinical Stroke Registries

Hannah Asperger[1,2] · Johannes Weller[1,3,4] · Felix J Bode[1] · Taraneh Ebrahimi[1,2] · Lara Klehr[1] · Julia Layer[1,2] · Julius N Meißner[1] · Louisa Nitsch[1,3] · Omid Shirvani[1,2] · Christian S Thielscher[1,2] · Niklas M von Danwitz[1,2] · Sach Mukherjee[2] · Franziska Dorn[5] · Gabor C Petzold[1,2] · Sebastian Stösser[1] · on behalf of the GSR-ET Investigators

Received: 21 July 2026 / Accepted: 30 October 2026
© The Author(s) 2026

Abstract

Clinical registries capture adverse events (AEs) in free-text fields that are rarely analyzed. We developed a reproducible large language model (LLM)-assisted workflow using 11,216 free-text entries from 18,069 patients in the German Stroke Registry; 9,146 informative entries were retained after filtering. Embedding-based clustering showed poor separation, whereas an LLM-guided taxonomy identified 89 leaf labels. On the model-assisted 1,000-entry development set used for model and prompt selection, GPT-5.4 reached F1 92.4% (95% confidence interval [CI] 91.4-93.4). On 300 held-out entries, weighted F1 against clinician consensus was 96.5% (94.6-98.1) for frozen GPT-5.4 labels and 90.5% (87.5-93.5) for locally run Qwen 3.5 27B. For AE concepts with a comparable structured field, free text documented 1,345 additional instances (+19.6% over 6,867 structured instances); it also documented 7,546 instances of AE concepts without a structured equivalent, most often pneumonia, urinary tract infection and delirium. Beyond baseline covariates and the structured AE count, integrated complication concepts increased the area under the curve for prediction of good functional outcome by 0.025 (95% CI 0.021-0.028) and for death by 0.018 (0.014-0.021); gains were minimal after including the discharge National Institutes of Health Stroke Scale (NIHSS) score. On 1,979 public coronary-trial fragments from the Aggregate Analysis of ClinicalTrials.gov (AACT), the software ran end to end, demonstrating its technical applicability to a second clinical text corpus. In conclusion, this workflow enables structured analysis of clinical registry free text and complements structured adverse-event documentation.

Keywords Clinical information systems · Stroke registry · Large language models · Clinical natural language processing · Complication surveillance · Endovascular thrombectomy · Clinical information extraction · Adverse event detection

Introduction

Standardized data collection in clinical registries inevitably simplifies complex clinical observations. Free-text fields are how clinicians document what predefined categorical variables cannot capture, yet this unstructured information remains largely unused for systematic analysis. This is a practical gap for clinical information systems: registries already contain rich safety information, but the information is locked in a format that is difficult to aggregate, validate, and reuse across research and quality-monitoring workflows.

Stroke registries monitor the quality and outcomes of acute stroke care [1, 2]. With endovascular thrombectomy (EVT) established as standard of care for large-vessel

✉ Hannah Asperger
hannah.asperger@ukbonn.de

1 Division of Vascular Neurology, University Hospital Bonn, Bonn, Germany

2 German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

3 Department of Neuroimmunology, University Hospital Bonn, Bonn, Germany

4 Department of Neurooncology, University Hospital Bonn, Bonn, Germany

5 Department of Neuroradiology, University Hospital Bonn, Bonn, Germany

Published online: 16 November 2026

occlusion stroke [3–5], registries provide real-world safety data that complement trial evidence [6, 7]. Complications after EVT are common and include intracranial hemorrhage, infarct progression, emboli to new territories, arterial injury, and access-site complications [6, 7]. Reported rates range from 10% to 20%, although definitions vary [7]. Medical complications such as stroke-associated pneumonia are similarly frequent and worsen outcomes [2, 8, 9].

Beyond these headline events, prospective studies show a broad and persistent post-stroke complication burden that spans infections, falls, thromboembolism, pain, and neuropsychiatric sequelae, many of which remain common beyond hospital discharge [1, 2]; for common entities such as post-stroke infections, pooled incidence estimates vary widely, suggesting that measurement choices strongly shape reported complication profiles [8]. In registry settings, structured categorical variables enable aggregation but are limited to predefined categories, whereas rare, context-dependent, or complex events end up in free-text fields [7, 10]. The heterogeneity of these entries, with abbreviations, mixed languages, and inconsistent terminology, has so far precluded their systematic use for research or quality monitoring.

Natural language processing (NLP) can convert unstructured text into structured data. Biomedical language models such as BioBERT, PubMedBERT, and ClinicalBERT have been shown to outperform general-domain models on clinical tasks [11–13]. Embedding-based clustering groups semantically similar texts [14, 15] but struggles with the overlapping categories and imbalanced frequencies typical of complication data. Large language models (LLMs) can classify text with minimal task-specific training [16], although clinical applications require careful evaluation given the risks of inconsistent outputs [17, 18] and privacy constraints [10].

We developed and evaluated a reproducible LLM-assisted workflow for converting registry free text into an integrated complication surveillance layer. Using free-text data from the German Stroke Registry–Endovascular Treatment (GSR-ET), we (1) characterized the spectrum of complications beyond predefined categorical variables, (2) tested whether unsupervised clustering reveals clinically meaningful patterns, (3) benchmarked LLM-based classification, reporting the data path of every model call, and (4) quantified what taxonomy-derived concepts add to 90-day outcome discrimination once integrated with structured registry data. The intended contribution is not a fixed stroke-specific label list, but an auditable workflow that other registries can directly utilize, using their own text, taxonomy decisions, validation sets, and deployment constraints.

Methods

Data Source and Preparation

We analyzed data from GSR-ET (ClinicalTrials.gov: NCT03356392), an ongoing, investigator-initiated, multicenter, open-label registry that prospectively collects data on patients with intracranial large-vessel occlusion (LVO) stroke treated with endovascular treatment [19]. Data from 30 participating centers were included, comprising 18,069 patients enrolled between May 2015 and December 2023. AEs are documented at four clinical timepoints (acute treatment, 24 hours, hospital discharge, and 90-day follow-up); at each, common complications are captured through structured categorical variables, while free-text fields allow documentation of AEs not covered by predefined categories.

We extracted all non-null free-text entries across the four timepoints (11,216 entries). Although designated for AE documentation, these fields were also used to record clinical findings, procedures, comorbidities, and outcomes, so any taxonomy had to cover the full range of clinical information rather than AEs alone. Encoding correction, Unicode normalization, and whitespace standardization were applied for language detection, sufficiency filtering, and duplicate analysis, without case normalization; the language models received the original registry strings. We excluded entries consisting exclusively of binary responses without clinical descriptions and entries containing only ambiguous single characters, while short entries matching established medical abbreviations (e.g., SAH, DVT) were retained. English entries were kept in their original language.

Only the text of the free-text field was sent to models, without any identifier or other registry field; hosted models were called through the OpenRouter API with its zero-data-retention routing flag [20]; MedGemma 27B and, in the held-out validation, Qwen 3.5 27B ran locally. The data path and local or hosted deployment of each model are detailed in Supplementary Methods S1 and Supplementary Table 1 (Online Resource 1).

Embedding-Based Clustering Evaluation

Before developing the LLM-based approach, we tested whether embedding-based clustering could identify meaningful complication categories, using six biomedical, general-purpose, and multilingual embedding models (Supplementary Table 2) on original-language entries and English translations produced with GPT-4o-mini. K-Means and density-based HDBSCAN clustering were assessed by the Silhouette Score (values above 0.50 indicating adequate separation) and the HDBSCAN noise proportion, with 95% confidence intervals (CIs) from 100 bootstrap iterations (Supplementary Methods S2, Supplementary Figs. 1-2).

Taxonomy Development

We developed the taxonomy with a data-driven, bottom-up LLM pipeline in three stages: taxonomy discovery (broad clinical domains, then categories within each domain), labeling of all entries against the resulting flat taxonomy with review of residual "other" categories, and hierarchy construction (subdividing categories with at least 200 provisionally labeled entries, then tree consolidation and cross-domain ambiguity pruning); each discovery step combined multiple independent runs with different random seeds with LLM-guided consolidation. Qwen 3.5 27B performed discovery, consolidation, review, and hierarchy construction, and Qwen 3.5 9B the flat labeling of all 9,146 entries (Supplementary Methods S3). A neurology resident with several years of experience in vascular neurology reviewed every pipeline step, and a vascular neurology attending reviewed the final taxonomy before it was frozen. The original label names denote operational text-derived categories, which may include suspected diagnoses or findings beyond the named diagnosis; their scope is specified in Online Resource 2 and does not imply clinical adjudication. We mapped the 89 taxonomy leaves to SNOMED CT and the 40 AE leaves additionally to MedDRA (Supplementary Methods S4). The workflow is implemented in an open-source Python package (version 1.1.0) at https://github.com/haasp1/med_taxonomizer [21].

LLM Classification Benchmark and Development Reference Standard

For a 1,000-entry set drawn with a fixed random seed, the neurology resident annotated all entries in an interface that displayed GPT-5.2 labels, and the attending then reviewed the reference with model labels visible after benchmark predictions existed; there was no independent duplicate labeling, and because the same set was used to compare models and to refine the prompt, we refer to it as the development set (Supplementary Methods S5). Ten LLMs were first evaluated using the initial prompt (prompt v1), each classifying the entries three times against the full taxonomy tree with structured output constrained to valid leaf paths (model identifiers and inference settings in Supplementary Table 3). Labels were aggregated by majority vote; we report label-level precision, recall, and F1 with 95% bootstrap CIs, exact TP, FP, and FN counts, and single runs. GPT-5.4 and Qwen 3.5 27B were then re-evaluated on the same entries with an enhanced prompt (prompt v2) that added clinical inference rules, a registry abbreviation dictionary, and hierarchy-level guidance.

Held-Out Comparison with Clinician Consensus

We evaluated the frozen taxonomy and prompt v2 on 300 entries selected from the 8,146 entries outside the model-selection and prompt-refinement set, using stratified random sampling enriched for 22 rare predicted leaves (annotation manual in Online Resource 2). The two clinicians labeled every entry independently of each other, with each initial set saved before the GPT-5.4 labels for that entry were displayed. They then reached consensus. GPT-5.4 and locally run Qwen 3.5 27B were evaluated against this consensus (Supplementary Methods S5).

Full-Corpus Labeling and Source Comparison

GPT-5.4 with prompt v2 labeled all informative entries in a single run, reusing the first prompt v2 benchmark run for the development entries (Supplementary Methods S6). We linked labels to registry records and timepoints, mapped structured categorical AE fields to comparable taxonomy concepts, and counted patient × timepoint × concept instances, separating free-text-only instances of concepts with a directly comparable structured field from those of concepts without any structured equivalent (Supplementary Methods S7). Using development-set performance, we assessed sensitivity of the source counts to false-positive and missed labels and projected counts for the earlier prompt v1 classifier. Death was compared separately with a structured modified Rankin Scale (mRS) score of 6 at the same registry timepoint, and a registry-based symptomatic intracranial hemorrhage (sICH) indicator approximating the European Cooperative Acute Stroke Study II (ECASS II) criterion was derived from documented hemorrhage plus a National Institutes of Health Stroke Scale (NIHSS) increase of at least 4 points at the same or next available timepoint before 90 days [22].

Clinical Utility Analysis

We related early and in-hospital non-death complications (documented during treatment, at 24 hours, or at discharge) to 90-day good functional outcome (mRS 0-2 among patients with pre-stroke mRS 0-2) and death (mRS 6), adjusting for age, sex, pre-stroke mRS, baseline NIHSS, posterior circulation occlusion, IV thrombolysis, general anesthesia, successful reperfusion, and external/mothership admission status. We estimated adjusted associations for frequent individual complication concepts and compared outcome-model discrimination across the + Structured AE, + Integrated burden, + Integrated categories, and + Integrated

concepts feature sets, with time-updated comparators that added 24-hour NIHSS and individual structured AE indicators (and discharge NIHSS) (Supplementary Methods S8). Robustness to documentation intensity, center effects, and missing 90-day mRS was assessed with center fixed effects, a documentation stress test, restriction to high-follow-up centers, delta-adjusted multiple imputation, and leave-one-center-out validation.

Software Demonstration on Public AACT Trial Text

To show that the discovery workflow runs end to end on an independent public corpus, we reran it on 1,979 target-enriched fragments from 1,747 coronary-intervention trials in the Aggregate Analysis of ClinicalTrials.gov (AACT) [23], withholding the originating table from all prompts, and labeled them with GPT-5.4 under a source-blind and a cardiology-aware prompt. We compared the predicted domains with the originating AACT tables (Supplementary Methods S9).

Statistical Analysis

Analyses were performed in Python 3.12 with scikit-learn and statsmodels (Supplementary Methods S10). Concept-level associations were reported as odds ratios (ORs) with 95% CIs and interpreted descriptively. Discrimination was estimated by 5-fold stratified cross-validation (area under the curve [AUC], average precision, Brier score), and paired AUC differences were assessed with 5,000 paired bootstrap resamples of the out-of-fold predictions, with Benjamini-Hochberg adjustment across the prespecified comparisons. Missing predictors were imputed before fitting association models and within training folds for prediction models; outcomes were imputed only in the missing-outcome sensitivity analyses. GPT-5.4 (OpenAI) and Claude Opus 4.6 (Anthropic) were used as supporting tools for coding assistance and language editing.

Results

Data Preparation

Of the 11,216 extracted free-text entries, 9,146 (81.5%) contained informative medical descriptions and formed the corpus for all subsequent analyses; 2,068 binary "yes"/"no" responses and 2 ambiguous single characters were excluded, and encoding corrections were applied to 1,168 entries (10.4%) for filtering and duplicate analysis only (Fig. 1).

Embedding-Based Clustering

All six embedding models reached K-Means Silhouette Scores below 0.30, well under the 0.50 threshold for adequate clustering (best: ClinicalBERT, 0.255), with or without English translation, and the uniform manifold approximation and projection (UMAP) visualization showed no clear cluster separation (Fig. 2A-B; Supplementary Tables 4-6, Supplementary Figs. 1-2). HDBSCAN classified 62-71% of texts as noise (Fig. 2C-D), so registry entries did not separate into stable, clinically interpretable clusters, motivating an LLM-guided pipeline that discovers a taxonomy directly from the data.

Taxonomy Discovery

The LLM-guided pipeline (Fig. 3A) identified five clinical domains (adverse events [AEs], clinical findings, procedures, comorbidities, and other), each proposed by at least four of six independent runs; only one corresponded to the designated purpose of these fields. Two-step consolidation produced an intermediate flat taxonomy of 40 categories (13,025 labels on the 9,146 entries, mean 1.4 per entry), and hierarchy construction with cross-domain ambiguity pruning yielded 89 leaf labels across five domains and up to three levels (Fig. 3; Supplementary Fig. 3). The taxonomy contained clinically specific distinctions such as hemorrhagic transformation subtypes (petechial vs. parenchymal), procedure-related versus spontaneous vascular events, and pre-existing comorbidities versus acute AEs based on temporal qualifiers. Exact MedDRA matches were assigned to 14 of the 40 AE leaves; exact SNOMED CT matches were assigned to 36 of all 89 leaves (12 AE and 24 non-AE; Online Resource 3).

LLM Classification Benchmark

On the development set (1,787 reference labels on 88 of the 89 leaves), GPT-5.4 (F1 88.9%, 95% CI 87.7-90.0), Gemini 3.1 Pro (88.5%, 87.3-89.6), and Qwen 3.5 27B (87.2%, 85.9-88.5) had the highest F1 under prompt v1 (Fig. 4A; Supplementary Tables 7-8); MedGemma 27B reached 64.4%. Precision generally exceeded recall, and AEs were classified more reliably than clinical findings and comorbidities (Fig. 4B-C; Supplementary Figs. 4-5). Prompt v2 raised GPT-5.4 F1 to 92.4% (91.4-93.4) by trading precision (95.3% to 91.9%) for recall (83.3% to 92.9%) (Fig. 4D; Supplementary Tables 7, 9; Supplementary Fig. 6); Qwen 3.5 27B reached 91.0%. When the initial annotation was used as reference for the same 1,000 development entries, GPT-5.4 with prompt v2

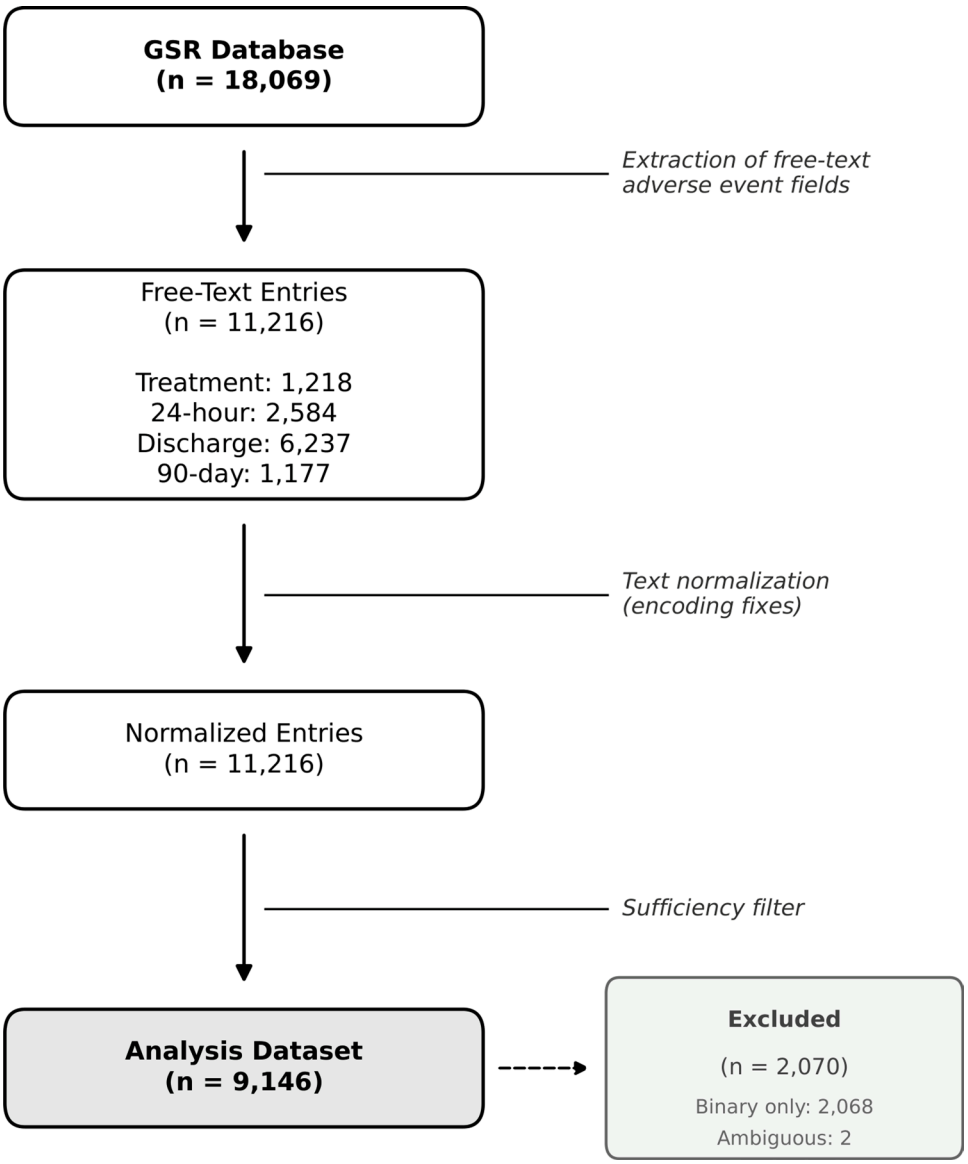


Fig. 1 Data preparation pipeline. Overview of the free-text extraction, normalization, and filtering process

reached micro F1 88.3% across all taxonomy domains. Paired prompt differences are reported in Supplementary Table 8. Structural failures and hierarchy-level errors and performance by label frequency are reported in Supplementary Tables 10 and 11, respectively. On 300 held-out entries, the two clinicians established a consensus reference following independent initial annotation (inter-annotator agreement in Supplementary Table 12A). Against this consensus, frozen GPT-5.4 corpus labels reached stratum-weighted F1 96.5% (95% CI 94.6-98.1), while locally run Qwen 3.5 27B reached 90.5% (87.5-93.5) by majority vote across three runs (full performance metrics in Supplementary Table 12B).

Full-Dataset AE Profile

Applying GPT-5.4 with prompt v2 in a single run to all 9,146 informative entries produced 16,727 labels (mean 1.8 per entry), of which 10,919 (65.3%) were AEs; at least one AE label was assigned to 8,042 entries (87.9%). Across the four registry timepoints, the AE profile shifted from procedure-related events during treatment toward late medical complications by 90 days (Fig. 5). During treatment, procedural events, hemorrhage, and thromboembolism together accounted for 81.4% of AE labels, led by intracranial hemorrhage (20.9%). By 24 hours the profile shifted toward acute brain injury, with hemorrhage (27.9%) and neurological

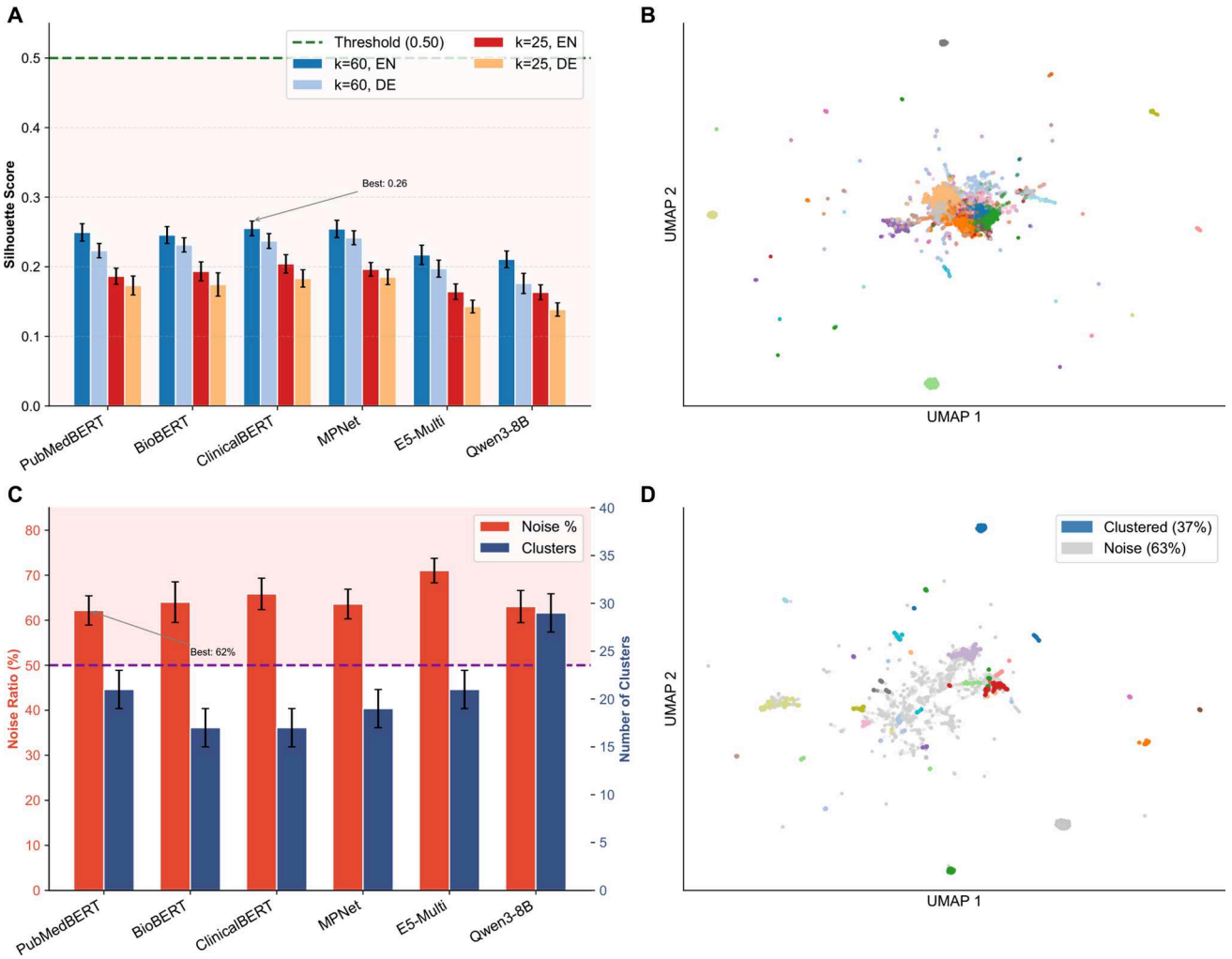


Fig. 2 Embedding-based clustering evaluation across six embedding models. (**A**) K-Means Silhouette Scores comparing Silhouette-optimal k (k=55-60) and Elbow-optimal k (k=20-25) for both English translations and original-language entries. All conditions fell below the 0.50 threshold indicating adequate clustering. ClinicalBERT achieved the highest score (0.255 at k=60). (**B**) UMAP visualization of ClinicalBERT embeddings (k=25) showing overlapping, poorly separated clusters. (**C**) HDBSCAN noise ratio and number of identified clusters across the six embedding models, with the purple dashed line marking the 50% noise threshold. (**D**) UMAP visualization of PubMedBERT embeddings with HDBSCAN clustering. Gray points represent noise (62% of data); colored points show identified clusters

events (22.4%) predominating and pneumonia emerging (13.7%). By discharge, infections dominated (44.0% of AE labels), with pneumonia the single most frequent AE (28.2%), followed by death (14.2%) and urinary tract infection (8.7%). At 90-day follow-up the profile diversified, with infections still most common (30.3%), documented mortality of 16.1%, and falls, injuries, and seizures together contributing 12.7%.

Structured Categorical Variables and Free Text Capture Complementary Information

Of 16,359 non-death AE concept-instances, 8,891 (54.3%) were documented only in free text, 6,736 (41.2%) only in structured variables, and 732 (4.5%) in both (Supplementary Fig. 7A). These free-text-only instances comprise two distinct findings. First, for the 8 AE concepts with a directly comparable structured field, free text documented 1,345 instances that the structured field did not record, an increase of 19.6% over the 6,867 structured instances of the same concepts (Supplementary Fig. 7C). Second, 7,546 free-text-only instances concerned AE concepts for which no structured field exists, most often pneumonia, urinary tract infection, other infection, delirium, and reocclusion (Supplementary Fig. 7D). Concept-level classification performance and the sensitivity of source counts to classification error are reported in Supplementary Table 13A and B, respectively. Of 1,088 free-text death labels (1,035 patients), 1,048 (96.3%) coincided with an mRS of 6 recorded at the same registry timepoint (Supplementary Table 14).

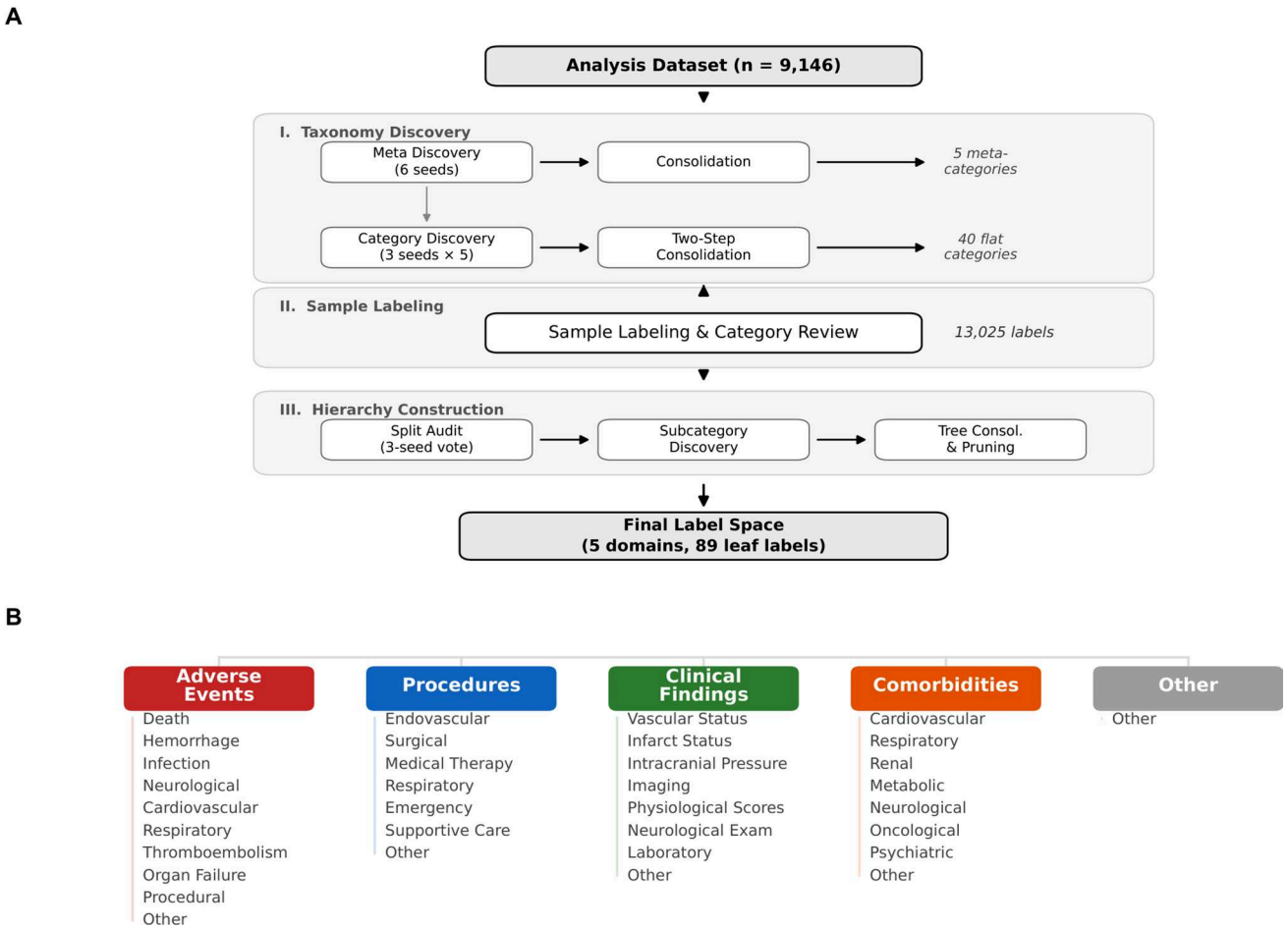


Fig. 3 LLM-guided taxonomy discovery and resulting classification. (**A**) The discovery pipeline comprises three stages: Stage I applied a multi-seed ensemble approach at two levels (six independent runs for domain discovery, three runs per domain for category discovery), each followed by LLM-guided consolidation. Stage II classified all 9,146 entries against the resulting flat taxonomy with multi-label assignment and reviewed residual "other" categories. Stage III identified categories with sufficient diversity for subdivision through three-seed majority voting, discovered subcategories, and refined the hierarchical tree through consolidation and cross-domain ambiguity pruning. (**B**) Overview of the resulting label space with five domains (color-coded) and their root-level categories. The complete hierarchical taxonomy with all 89 leaf labels is provided in Supplementary Fig. 3

Integrated Complication Capture and Outcome Patterns

Overall, 8,293 of 18,069 patients (45.9%) had at least one integrated non-death complication during treatment, at 24 hours, or by discharge (Table 1).

The most frequent integrated non-death complications across the full cohort were intracranial hemorrhage (14.1%), pneumonia (11.6%), recurrent stroke (8.0%), space-occupying infarction/elevated intracranial pressure (ICP; 5.9%), and derived sICH (4.6%) (Fig. 6A, Table 2; full list in Supplementary Table 15). Structured variables dominated for mapped concepts (for example, 94.3% of recurrent stroke instances were structured-only), whereas pneumonia and urinary tract infection were documented exclusively in free text and vasospasm only from structured fields.

In exploratory baseline-adjusted analyses, space-occupying infarction/ICP and derived sICH had the largest associations, with much lower odds of good functional outcome (OR 0.04 and 0.09) and higher odds of death (OR 7.30 and 5.38); intracranial hemorrhage, recurrent stroke, clot migration, and dissection/perforation were also unfavorable, whereas vasospasm showed no clear association (Fig. 6B).

Observed 90-day mRS was less complete among patients without documented complications (74.7% vs. 86.8%; Supplementary Table 16). Among patients with pre-stroke mRS 0-2 and observed 90-day mRS (n=12,225), the proportion with 90-day mRS 0-2 fell from 53.4% (no documented complications) to 11.3% (three or more complications), while death rose from 17.6% to 46.1% (Fig. 6C).

Adding the structured complication count to baseline clinical and procedural covariates increased cross-validated

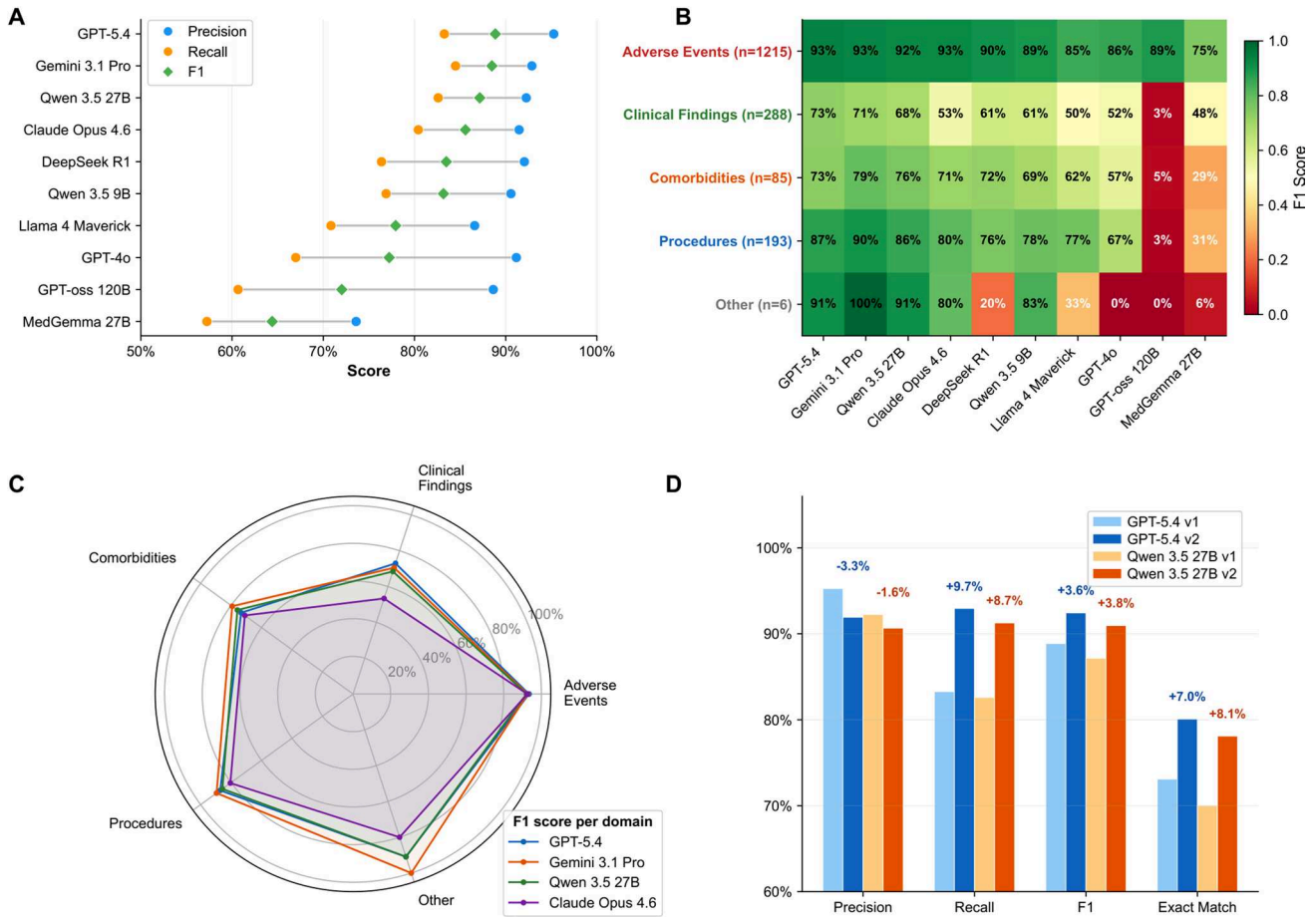


Fig. 4 LLM classification benchmark results. (**A**) Precision, recall, and F1 score for each model on the 1,000-entry development set, scored against the model-assisted clinical reference standard, using majority voting across three independent runs (runs without a label for an entry excluded from the vote). Models are sorted by F1; paired comparisons are reported in Supplementary Table 8. (**B**) F1 score per taxonomy domain and model. Sample counts (n) indicate the number of reference labels per domain. All models performed well on AEs but showed substantially lower performance on clinical findings and comorbidities. (**C**) Radar plot of the F1 score for the top-4 models across taxonomy domains, illustrating model-specific performance profiles. (**D**) Effect of prompt engineering: comparison of prompt v1 (baseline) vs. prompt v2 (enhanced with clinical inference rules, a registry abbreviation dictionary, and multi-level annotation guidance) for GPT-5.4 and Qwen 3.5 27B. Delta annotations show the absolute change; the largest gains were in recall (+9.7% and +8.7%) and exact match (+7.0% and +8.1%)

AUC from 0.797 to 0.819 for good functional outcome and from 0.795 to 0.815 for death, and indicators for the 20 most frequent individual complications (+ Integrated concepts) gave the highest discrimination (0.843 and 0.833; Fig. 6D, Supplementary Table 17). Compared with the structured AE count alone, this model increased AUC by 0.025 (95% CI 0.021-0.028) for good functional outcome and by 0.018 (95% CI 0.014-0.021) for death (both adjusted p<0.001; Supplementary Table 18). Discrimination was driven mainly by baseline NIHSS, age, and premorbid mRS; among complication-derived features, complication burden and space-occupying infarction/ICP contributed most (Supplementary Table 19). In time-updated sensitivity analyses the increment shrank to 0.004 (good functional outcome) and 0.005 (death) after 24-hour NIHSS and individual structured AE indicators were included, and to 0.002 and 0.000 (p=0.88 for death) after discharge NIHSS (Supplementary Table 20).

Documentation and complication capture varied across centers (Supplementary Table 21). The discrimination gain persisted in leave-one-center-out validation (Supplementary Table 22). Most associations changed little after center adjustment or restriction to centers with high follow-up completeness, whereas additional documentation adjustment attenuated several associations and reversed some (Supplementary Table 23). Missing-outcome sensitivity analyses are reported in Supplementary Table 24.

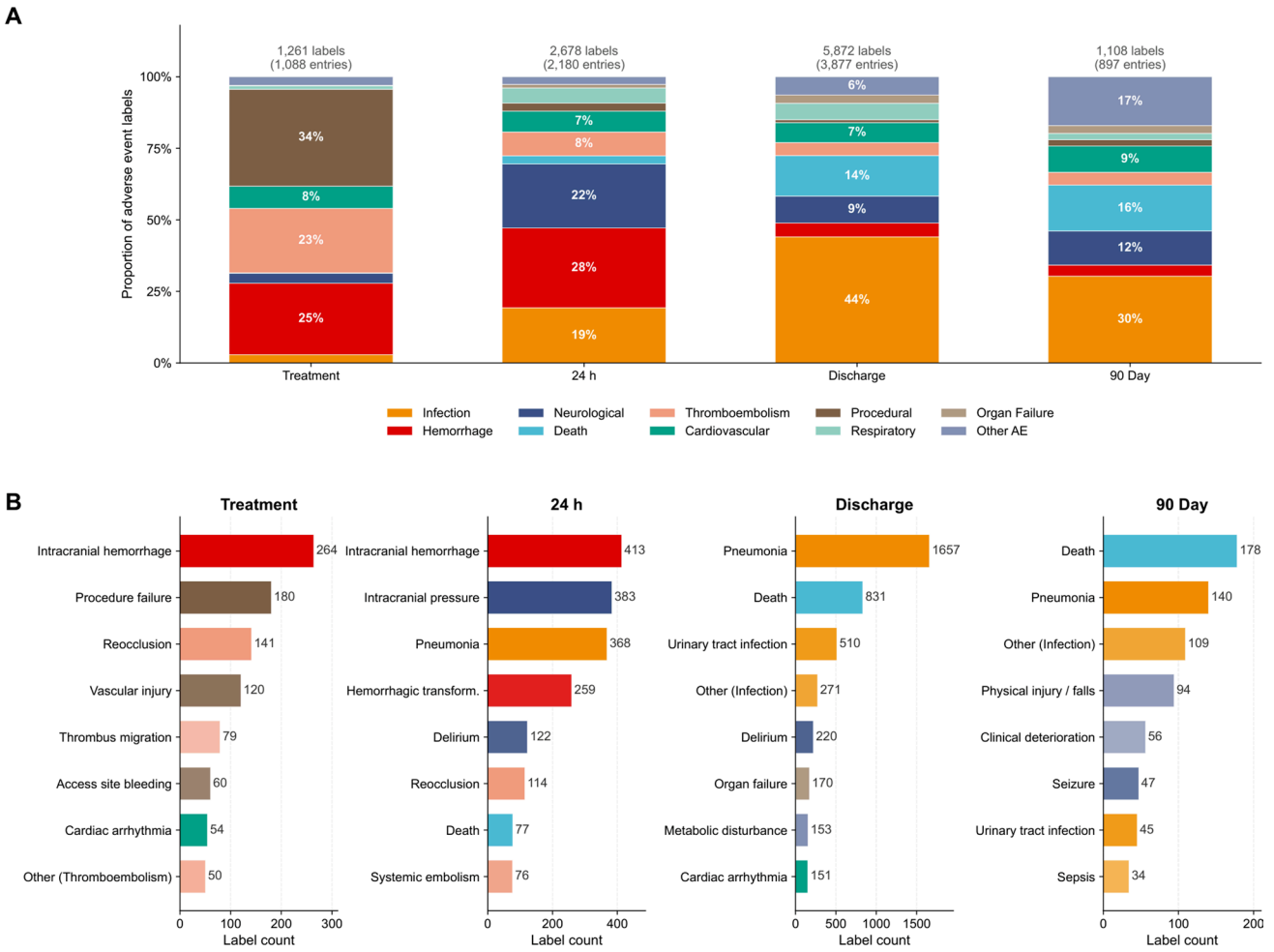


Fig. 5 Temporal AE profile across four clinical timepoints. (**A**) Proportional distribution of AE categories at each timepoint. Numbers above bars indicate total AE labels and number of free-text entries. (**B**) Top eight specific AEs per timepoint, colored by parent category as in (A). The complication profile shifts from procedure-related events during treatment (hemorrhage, thromboembolism, procedural complications: 81.4% of AE labels) through acute brain injury at 24 hours (hemorrhage and neurological events: 50.3% of AE labels) to infection-dominated patterns at discharge (44.0% of AE labels) and a heterogeneous mix of late complications at 90-day follow-up

Software Demonstration on Public AACT Coronary-Intervention Text

On 1,979 fragments from 1,747 coronary-intervention trials, the workflow induced, without access to the source-table labels, a hierarchy of six meta domains (AEs, conditions, interventions, outcomes, study design, and other), 36 subcategories, and 74 sub-subcategories (Supplementary Fig. 8). Source-role distributions and the prompt-context comparison are reported in Supplementary Table 25.

Discussion

This study shows that free-text AE fields in a large real-world stroke thrombectomy registry contain clinically relevant information that is not captured by predefined categorical variables alone. Registries are essential for real-world safety surveillance [6, 24], but standardized data elements necessarily simplify complex bedside documentation [7, 10]. The free-text fields were not merely incomplete structured complication fields; they contained a pragmatic clinical narrative spanning AEs, clinical findings, procedures, comorbidities, and outcomes, a mismatch also documented across registries, real-world data networks, and electronic health records [25–30].

The taxonomy was useful because it matched this documentation structure: derived from the registry text itself, it retained a clinically interpretable hierarchy and allowed multi-label assignment when a single entry documented several concurrent problems. Structured variables captured procedure-oriented concepts such as recurrent stroke and vasospasm, whereas free text documented pneumonia, urinary tract infection, other infections, delirium, seizures, and

Table 1 Baseline characteristics by early/in-hospital integrated complication status. Patients were grouped by whether at least one non-death complication concept was documented during treatment, at 24 hours, or at discharge after integrating structured categorical AE fields with taxonomy-classified free-text adverse events. "No complication" and "≥1 complication" refer to early/in-hospital integrated complication status. Values are median [IQR] or n/N (%). Outcomes are based on observed 90-day mRS values

Characteristic	Overall	No complication	≥1 complication
N	18,069	9,776	8,293
Age, years	76 [65-83]	76 [65-83]	77 [66-83]
Female sex	9,349/18,051 (51.8)	5,115/9,763 (52.4)	4,234/8,288 (51.1)
Premorbid mRS	0 [0-2]	0 [0-2]	0 [0-1]
Baseline NIHSS	14 [8-18]	13 [8-18]	15 [10-19]
Posterior circulation occlusion	2,043/18,069 (11.3)	1,087/9,776 (11.1)	956/8,293 (11.5)
IV thrombolysis	8,311/17,858 (46.5)	4,485/9,608 (46.7)	3,826/8,250 (46.4)
General anesthesia	13,709/16,906 (81.1)	7,285/8,964 (81.3)	6,424/7,942 (80.9)
Successful reperfusion	14,573/16,978 (85.8)	8,017/8,994 (89.1)	6,556/7,984 (82.1)
External admission	7,268/18,069 (40.2)	3,819/9,776 (39.1)	3,449/8,293 (41.6)
90-day mRS available	14,500/18,069 (80.2)	7,304/9,776 (74.7)	7,196/8,293 (86.8)
90-day mRS 0-2	5,003/14,500 (34.5)	3,427/7,304 (46.9)	1,576/7,196 (21.9)
90-day mRS 3-5	4,925/14,500 (34.0)	2,265/7,304 (31.0)	2,660/7,196 (37.0)
90-day mRS 6	4,572/14,500 (31.5)	1,612/7,304 (22.1)	2,960/7,196 (41.1)

reocclusion. The source comparison contains two different findings: for concepts that both sources can record, free text added to structured capture, whereas most free-text-only instances reflected concepts for which the case-report form has no field; the relevant unit of analysis is therefore the integrated complication profile.

The limited performance of embedding-based clustering marks a boundary of single-vector representations for short, abbreviated, multi-event registry text rather than a general limitation of embeddings [14, 15, 29, 30]. Large general-purpose LLMs performed best and the domain-specific medical model substantially worse, because the task required instruction following, multilingual robustness, and multi-label reasoning in addition to clinical vocabulary; this mirrors recent evidence on general-purpose LLMs in medicine [31], which still require task-specific validation before clinical use [17, 18, 32]. The documented AE spectrum, with neurovascular and procedural events early and infections and medical complications by discharge and 90 days, aligns with the EVT and stroke-complication literature while adding granularity [1, 2, 6, 8, 33–38].

These findings support a hybrid registry strategy. Structured categorical variables remain necessary for high-priority complications with stable definitions and audit relevance [7], while free-text processing preserves clinically important detail without continually expanding case-report forms; infections, delirium, and seizures may be candidates for future structured capture, whereas vasospasm should remain structured. Death is better treated as an outcome, and derived phenotypes such as registry-based sICH can be constructed from existing hemorrhage and NIHSS data. Such a workflow fits offline registry curation rather than bedside decision support, with registry teams periodically auditing the taxonomy and classified text [18].

The outcome analyses linked the integrated complication profile to 90-day outcomes but should not be interpreted causally. Integrated concept-level features added discrimination beyond baseline covariates and structured AE capture, but the increment was attenuated after 24-hour NIHSS and minimal after discharge NIHSS, consistent with prior GSR-ET modeling [39]. Derived sICH also incorporates post-baseline neurological deterioration; the increment therefore reflects prognostic information available later in the clinical course and cannot be attributed solely to free-text capture. "No documented complication" partly marked less completely documented records. Adjusting for text availability partly conditions on exposure measurement, so its attenuation and reversal of some associations require caution; undocumented complications cannot be recovered by these analyses. The strongest signals came from space-occupying infarction/ICP and derived sICH, consistent with the prognostic role of malignant edema and post-thrombectomy hemorrhage [40]; as post-baseline events they may be mediators or markers of severity, so their value is for prognosis, audit, and risk-adjustment rather than as a standalone prediction model [41].

The reusable contribution is the workflow rather than this exact taxonomy: other registries should sample their own text, discover and validate labels, and integrate the result with structured variables, and the open-weight Qwen 3.5 27B can be deployed locally where hosted services are not permitted. The AACT run demonstrates software execution on another corpus, but does not validate the GSR-ET taxonomy, classifier accuracy, or clinical generalizability.

We derived the taxonomy from the registry text rather than adopting MedDRA, SNOMED CT, the UMLS, or ICD [42–44], because the fields mixed AEs with procedures,

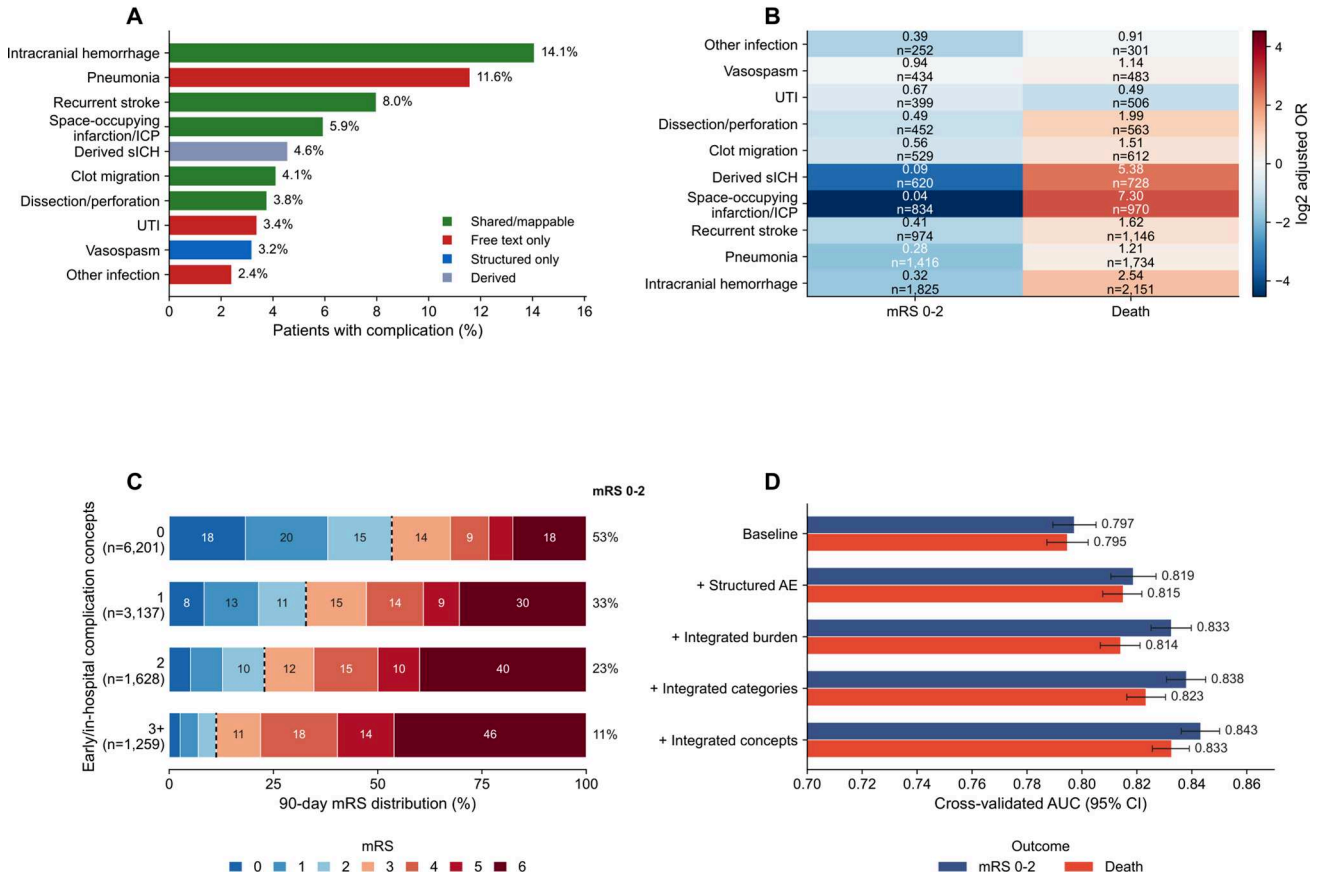


Fig. 6 Clinical utility of integrated taxonomy-based complication capture. Structured categorical AE concepts and taxonomy-classified free-text AEs were integrated into patient × timepoint × concept records. Death was excluded from complication counts and analyzed separately as an outcome. (**A**) Most frequent integrated non-death complications across the full registry cohort (n=18,069), shown as the percentage of patients with at least one corresponding event. Colors indicate whether concepts were mappable between structured and free-text sources, free-text-only, structured-only, or derived from a registry rule. (**B**) Exploratory baseline-adjusted ORs for the same frequent concepts in relation to good functional outcome and death. Models adjusted for age, sex, pre-stroke mRS, baseline NIHSS, posterior circulation occlusion, IV thrombolysis, general anesthesia, successful reperfusion, and external/mothership status. For mRS 0-2, ORs below 1 indicate lower odds of good functional outcome. For death, ORs above 1 indicate higher odds of death. (**C**) 90-day mRS distribution by early/in-hospital integrated complication burden, restricted to patients with pre-stroke mRS 0-2 and observed 90-day mRS. (**D**) Incremental 5-fold cross-validated logistic-regression discrimination for good functional outcome (90-day mRS 0-2 among pre-stroke mRS 0-2 patients) and death (90-day mRS 6). Error bars and bracketed intervals are 95% bootstrap confidence intervals from 1,000 resamples of out-of-fold predictions. Each model extends the baseline clinical and procedural covariates with early/in-hospital complication information at increasing resolution. + Structured AE adds the number of distinct complications recorded in the structured categorical fields. + Integrated burden adds the number of distinct complications after also including taxonomy-classified free text (a count only, without complication type). + Integrated categories adds this integrated count plus binary indicators for which broad complication categories were present (e.g., hemorrhage, infection, thromboembolism) and a derived sICH indicator. + Integrated concepts adds this integrated count plus the derived sICH indicator and binary indicators for the 20 most frequent individual complications (e.g., pneumonia, intracranial hemorrhage), without the category-level indicators

findings, and comorbidities in abbreviated German at bedside granularity. The terminology mapping supports interoperability, but broader or composite matches can lose clinical qualifiers, some leaves lack an equivalent, and MedDRA assignments require specialist confirmation. Clinical review during construction was not an independent per-label validation.

This study has limitations. First, the development reference standard was created with model labels visible, without independent duplicate labeling, and was reused for model selection and prompt refinement, so its performance estimates are subject to anchoring and model-selection bias. The held-out comparison used new entries, but its consensus was also formed with GPT-5.4 labels visible. The high agreement with that consensus does not eliminate possible model anchoring or establish an unbiased GPT–Qwen comparison; initial inter-annotator disagreement also indicates uncertainty in the human reference. Second, the

Table 2 Most frequent integrated non-death complications across the full registry cohort. Patient counts refer to unique patients with at least one event for the concept at any registry timepoint; percentages use the full cohort denominator (n=18,069). Names denote operational text categories (definitions in Online Resource 2), not adjudicated diagnoses. Instances count patient × timepoint × concept records after integrating structured categorical variables, taxonomy-classified free text, and derived sICH

Complication	Patients	% of cohort	Instances
Intracranial hemorrhage	2,542	14.1	3,192
Pneumonia	2,094	11.6	2,185
Recurrent stroke	1,441	8.0	1,553
Space-occupying infarction/ICP	1,072	5.9	1,313
Derived sICH	824	4.6	992
Clot migration	743	4.1	745
Dissection/perforation	680	3.8	692
UTI	610	3.4	627
Vasospasm	575	3.2	575
Other infection	435	2.4	448
Delirium	356	2.0	369
Myocardial infarction	353	2.0	373
Groin complications	351	1.9	403
Reocclusion	301	1.7	309
Cardiac arrhythmia	280	1.5	294
Systemic embolism	244	1.4	254
Systemic bleeding	232	1.3	243
Organ failure	230	1.3	236
Respiratory failure	207	1.1	217
Seizure	201	1.1	211
Procedure failure	197	1.1	198
Metabolic disturbance	183	1.0	183
Sepsis	176	1.0	180
Physical injury	171	0.9	173
Heart failure	154	0.9	164

benchmark mainly assesses frequent labels; performance for rare events is imprecise and their labels may be coarse. Third, the taxonomy was induced from a single national EVT registry and remains a registry-specific instrument that requires validation in other registries. Fourth, short, context-poor entries limit the separation of pre-existing from new conditions, and counts represent documented capture rather than adjudicated incidence. Some operational categories include findings beyond the named diagnosis, such as troponin elevation under myocardial infarction and isolated raised infection markers under infection; their counts should not be interpreted as confirmed diagnoses. Error-adjusted counts also assume that development-set performance applies to the corpus and does not depend on structured-field status. Fifth, the outcome analyses quantify incremental information value, not a deployable prognostic model or causal effects.

Future work should assess the taxonomy with independent clinicians and terminology experts and evaluate the workflow in other registries with locally deployed models.

Conclusion

Free-text AE entries in stroke registry data contain clinically relevant information that is not captured by structured categorical variables alone. A data-driven, hierarchical, multi-label taxonomy applied with LLMs can extract this information at scale, integrate it with existing structured categorical variables, and identify complication patterns associated with good functional outcome and death. The results support a hybrid registry strategy in which standardized categorical variables capture core complications while taxonomy-assisted free-text processing preserves clinically important granularity.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10916-026-02486-y.

Acknowledgements We acknowledge the contribution of all study nurses and extend our gratitude to the patients and their relatives for participating in the study.

Author Contributions HA and SS conceptualized and designed the study. OS, FJB, LN, SS, TE, NvD, HA, JL, JNM, LK, FD, GCP, CST, and JW acquired data. HA and SS analyzed data. HA created the development reference standard; SS reviewed it. HA and SS performed the held-out annotation. SM reviewed the taxonomy and annotation methodology from a machine-learning perspective. HA wrote the original draft. SS provided critical feedback on the manuscript. All authors reviewed and approved the final version of the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. This study was conducted without specific funding from public, commercial, or non-profit organizations.

Data Availability The data from the GSR-ET registry supporting the results of this study are available from the corresponding author upon reasonable request and pending authorization of the study sites and ethics committees. The AACT public-transfer dataset is derived from the publicly available Aggregate Analysis of ClinicalTrials.gov (AACT) database.

Declarations

Conflicts of interest FD: Consultant for Microvention, Balt, Johnson & Johnson; received scientific grants from Johnson & Johnson; received speakers honoraria from Johnson & Johnson, Q'Apel, Tonbridge, Asahi, Acandis, Stryker, Microvention, Medtronic, Penumbra. FJB: received speakers honoraria from AstraZeneca. All other authors report no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Langhorne, P., Stott, D. J., Robertson, L., MacDonald, J., Jones, L., McAlpine, C., Dick, F., Taylor, G. S., and Murray, G., Medical complications after stroke: a multicenter study. *Stroke*, 31 (6): 1223–1229, 2000. https://doi.org/10.1161/01.STR.31.6.1223.
2. Kumar, S., Selim, M. H., and Caplan, L. R., Medical complications after stroke. *The Lancet Neurology*, 9(1):105–118, 2010. https://doi.org/10.1016/S1474-4422(09)70266-2.
3. Goyal, M., Menon, B.K., van Zwam, W.H., et al., Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomised trials. *Lancet*, 387(10029):1723–1731, 2016. https://doi.org/10.1016/S0140-6736(16)00163-X.
4. Campbell, B.C.V., Mitchell, P.J., Kleinig, T.J., et al., Endovascular therapy for ischemic stroke with perfusion-imaging selection. *New England Journal of Medicine*, 372(11):1009–1018, 2015. https://doi.org/10.1056/NEJMoa1414792.
5. Saver, J.L., Goyal, M., Bonafe, A., et al., Stent-retriever thrombectomy after intravenous t-PA vs. t-PA alone in stroke. *New England Journal of Medicine*, 372(24):2285–2295, 2015. https://doi.org/10.1056/NEJMoa1415061.
6. Happi Ngankou, E., Gory, B., Marnat, G., Richard, S., Bourcier, R., Sibon, I., Dargazanli, C., Arquizan, C., Maïer, B., Blanc, R., Lapergue, B., Consoli, A., Vannier, S., Spelle, L., Denier, C., Boulanger, M., Gauberti, M., Saleme, S., Macian, F., Clarençon, F., Rosso, C., Naggara, O., Turc, G., Ozkul-Wermester, O., Papagiannaki, C., Viguier, A., Cognard, C., Lebras, A., Evain, S., Wolff, V., Pop, R., Timsit, S., Gentric, J.-C., Bourdain, F., Veunac, L., Eugène, F., Finitsis, S., and ETIS Registry Investigators., Thrombectomy complications in large vessel occlusions: Incidence, predictors, and clinical impact in the etis registry. *Stroke*, 52(12):e764–e768, 2021. https://doi.org/10.1161/STROKEAHA.121.034865.
7. Mortimer, A., Minks, D., Bhogal, P., Young, V., Macdonald, J., Sastry, A., Bosnell, R., McClelland, S.B., and White, P., Standardising mechanical thrombectomy complication reporting: A delphi consensus study to support guidance for national audit. *Interventional Neuroradiology*, 2025. https://doi.org/10.1177/15910199251364119. Online ahead of print.
8. Westendorp, W.F., Nederkoorn, P.J., Vermeij, J.-D., Dijkgraaf, M.G., and van de Beek, D. Post-stroke infection: a systematic review and meta-analysis. *BMC Neurology*, 11:110, 2011. https://doi.org/10.1186/1471-2377-11-110.
9. Ingeman, A., Andersen, G., Hundborg, H.H., Svendsen, M.L., and Johnsen, S.P., Processes of care and medical complications in patients with stroke. *Stroke*, 42(1):167–172, 2011. https://doi.org/10.1161/STROKEAHA.110.599738.
10. Berenspöhler, S., Minnerup, J., Dugas, M., and Varghese, J. Common data elements for meaningful stroke documentation in routine care and clinical research: Retrospective data analysis. *JMIR Medical Informatics*, 9(10):e27396, 2021. https://doi.org/10.2196/27396.
11. Lee, J., Yoon, W., Kim, S., et al., BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020. https://doi.org/10.1093/bioinformatics/btz682.
12. Gu, Y., Tinn, R., Cheng, H., et al., Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, 2022. https://doi.org/10.1145/3458754.
13. Alsentzer, E., Murphy, J.R., Boag, W., et al., Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 72–78, 2019. https://doi.org/10.18653/v1/W19-1909.
14. Reimers, N., and Gurevych, I., Sentence-BERT: sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pp. 3982–3992, 2019. https://doi.org/10.18653/v1/D19-1410.
15. McInnes, L., Healy, J., and Astels, S., hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2:205, 2017. https://doi.org/10.21105/joss.00205.
16. Brown, T.B., Mann, B., Ryder, N., et al., Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.
17. Singhal, K., Azizi, S., Tu, T., et al., Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023. https://doi.org/10.1038/s41586-023-06291-2.
18. Patel, S., Lam, K., et al., Safety and reliability of large language models in clinical decision support. *npj Digital Medicine*, 7:112, 2024. https://doi.org/10.1038/s41746-024-01102-y.
19. Alegiani, A.C., Dorn, F., Gemmler, H., et al., Systematic evaluation of stroke thrombectomy in clinical practice: the German stroke registry endovascular treatment. *International Journal of Stroke*, 14(4):372–380, 2019. https://doi.org/10.1177/1747493018806199.
20. OpenRouter. Zero Data Retention, 2026. https://openrouter.ai/docs/guides/features/zdr. Accessed 24 Sept 2026.
21. Asperger, H., Med Taxonomizer, version 1.1.0, 2026. https://github.com/haasp1/med_taxonomizer/releases/tag/v1.1.0. Concept https://doi.org/10.5281/zenodo.21275129.
22. Hacke, W., Kaste, M., Fieschi, C., von Kummer, R., Davalos, A., Meier, D., Larrue, V., Bluhmki, E., Davis, S., Donnan, G., Schneider, D., Diez-Tejedor, E., and Trouillas, P., Randomised double-blind placebo-controlled trial of thrombolytic therapy with intravenous alteplase in acute ischaemic stroke (ECASS II). *The Lancet*, 352(9136):1245–1251, 1998. https://doi.org/10.1016/S0140-6736(98)08020-9.
23. Tasneem, A., Aberle, L., Ananth, H., Chakraborty, S., Chiswell, K., McCourt, B.J., and Pietrobon, R., The database for aggregate analysis of ClinicalTrials.gov (AACT) and subsequent regrouping by clinical specialty. *PLoS ONE*, 7(3):e33677, 2012. https://doi.org/10.1371/journal.pone.0033677.
24. Cadilhac, D.A., Kim, J., Lannin, N.A., et al., National stroke registries for monitoring and improving the quality of hospital care: A systematic review. *International Journal of Stroke*, 11(1):28–40, 2016. https://doi.org/10.1177/1747493015607523.
25. Bun, R., Adisasmito, W.B., and Siswanto, B.B., Optimizing electronic medical records for an acute decompensated heart failure registry: A mixed-methods analysis of data completeness and coding practices at a national cardiovascular center in indonesia. *Cureus*, 18(1):e100876, 2026. https://doi.org/10.7759/cureus.100876.
26. Cheligeer, C., Southern, D.A., Yan, J., Wu, G., Pan, J., Lee, S., Martin, E.A, Jafarpour, H., Eastwood, C.A., Zeng, Y., and Quan, H. Utilizing large language models for detecting hospital-acquired conditions: an empirical study on pulmonary embolism. *Journal of the American Medical Informatics Association*, 32(5):876–884, 2025. https://doi.org/10.1093/jamia/ocaf048.
27. Blacketer, C., Schuemie, M.J., Moinat, M., Voss, E.A., Camprubi, M., Rijnbeek, P.R., and Ryan, P.B., Advancing real-world evidence through a federated health data network (EHDEN): Descriptive study. *Journal of Medical Internet Research*, 27:e74119, 2025. https://doi.org/10.2196/74119.
28. Murff, H.J., FitzHenry, F, Matheny, M.E., et al., Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA*, 306(8):848–855, 2011. https://doi.org/10.1001/jama.2011.1204.
29. Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S.F., Forshee, R., Walderhaug, M., and Botsis, T., Natural

language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *Journal of Biomedical Informatics*, 73:14–29, 2017. https://doi.org/10.1016/j.jbi.2017.07.012.
30. Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S., and Liu, H., Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, 77:34–49, 2018. https://doi.org/10.1016/j.jbi.2017.11.011.
31. Vishwanath, K., Alyakin, A., Ghosh, M., et al., General-purpose large language models outperform specialized clinical AI tools on medical benchmarks. *Nature Medicine*, 2023. https://doi.org/10.1038/s41591-026-04431-5.
32. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., et al., Large language models in medicine. *Nature Medicine*, 29(8):1930–1940, 2023. https://doi.org/10.1038/s41591-023-02448-8.
33. Behme, D., Gondecki, L., Fiethen, S., Kowoll, A., Mpotsaris, A., and Weber, W., Complications of mechanical thrombectomy for acute ischemic stroke-a retrospective single-center study of 176 consecutive cases. *Neuroradiology*, 56(6):467–476, 2014. https://doi.org/10.1007/s00234-014-1352-0.
34. Shapiro, M., Raz, E., Nossek, E., et al., Access site complications in endovascular treatment of acute ischemic stroke. *Journal of NeuroInterventional Surgery*, 12(5):475–478, 2020. https://doi.org/10.1136/neurintsurg-2019-015223.
35. Awere-Duodu, A., Darkwah, S., Osman, A.H., et al., A systematic review and meta-analysis show a decreasing prevalence of post-stroke infections. *BMC Neurology*, 24:479, 2024. https://doi.org/10.1186/s12883-024-03968-7.
36. Smith, C.J., Kishore, A.K., Vail, A., et al., Diagnosis of Stroke-Associated Pneumonia: Recommendations From the Pneumonia in Stroke Consensus Group. *Stroke*, 46(8):2335–2340, 2015. https://doi.org/10.1161/STROKEAHA.115.009617.
37. de Jonge, J. C., et al., Temporal Profile of Pneumonia After Stroke. *Stroke*, 53(1):53–60, 2022. https://doi.org/10.1161/STROKEAHA.120.032787.
38. Schumann-Werner, B., Becker, J., Nikoubashman, O., Wiesmann, M., Schulz, J.B., Reich, A., Pinho, J., and Werner, C.J., The relationship between neurogenic dysphagia, stroke-associated pneumonia and functional outcome in a cohort of ischemic stroke patients treated with mechanical thrombectomy. *Journal of Neurology*, 270:5958–5965, 2023. https://doi.org/10.1007/s00415-023-11940-7.
39. Shirvani, O., Warnat-Herresthal, S., Savchuk, I., Bode, F.J., Nitsch, L., Stösser, S., Ebrahimi, T., von Danwitz, N., Asperger, H., Layer, J., Meissner, J., and GSR-ET Investigators. Machine learning models for outcome prediction in thrombectomy for large anterior vessel occlusion. *Annals of Clinical and Translational Neurology*, 11(10):2696–2706, 2024. https://doi.org/10.1002/acn3.52185.
40. Qu, H.L., Sun, X.Y., Xu, X.M., and Zhang, X.B., Malignant cerebral edema after large-core anterior circulation stroke in the thrombectomy era: Prediction and neurocritical care pathways. *Neurocritical Care*, 2026. https://doi.org/10.1007/s12028-026-02518-y.
41. He, S., Gao, Y., Tan, Q., An, Y., Zhu, G., Lin, Y., Zhang, M., Mao, F., Deng, H., Chen, X., Chen, K., Hao, J., Yang, J., and Wang, X., Machine learning models for outcome prediction of patients with ischaemic stroke undergoing reperfusion therapy: a systematic review and meta-analysis. *Stroke and Vascular Neurology*, 2026. https://doi.org/10.1136/svn-2025-004020.
42. Brown, E.G., Wood, L., and Wood, S., The medical dictionary for regulatory activities (MedDRA). *Drug Safety*, 20(2):109–117, 1999. https://doi.org/10.2165/00002018-199920020-00002.
43. Donnelly, K., SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in Health Technology and Informatics*, 121:279–290, 2006.
44. Bodenreider, O., The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–D270, 2004. https://doi.org/10.1093/nar/gkh061.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.