Scientia Psychiatrica
Vol. 7, No. 4, 2026, pp. 226–237 · e-ISSN 2715-9736 journal homepage: scientiapsychiatrica.com

ORIGINAL RESEARCH ARTICLE

Explainable Temporal Reconciliation for Psychiatric Risk Phenotyping in Synthetic Longitudinal Clinical Notes

Dessy Agustina[1], Sarah Istiqomah[2*]
[1]*Department of Health Sciences, Sei Rambutan Community Health Center, Palembang, Indonesia*
[2]*Department of Neurology, Kenanga Medical Center, Demak, Indonesia*

ARTICLE INFO

Article history
Received: August 18, 2026
Revised: September 20, 2026
Accepted: October 14, 2026
Available online: October 26, 2026
Published: October 30, 2026

Keywords
Clinical natural language processing
Explainable artificial intelligence
Psychiatry
Synthetic data
Temporal reasoning

***Corresponding author**
Sarah Istiqomah
E-mail: Sarah.istiqomah@gmail.com
ORCID: 0009-0002-1316-0866

DOI
10.37275/scipsy.v7i4.224

ABSTRACT

Background. Psychiatric risk changes during hospitalization. Document-level natural language processing can retain historical high-risk language after the current clinical state has changed.

Objective. To evaluate an explainable approach that reconciles chronology, source authority, negation, and transition barriers across longitudinal records against a maximum-risk baseline.

Methods. This retrospective study used all 180 synthetic inpatient psychiatric chart packets of ClinAuthBench v1. A deterministic pipeline gave priority to the most recent suicide-risk screens, psychiatric progress notes, and social-work transition-barrier notes. The baseline was a maximum-risk-across-record strategy. The primary endpoint was current risk (low versus elevated/uncertain). Secondary endpoints were three-class risk, diagnosis, psychosis, substance use, transition barriers, and lower-level-of-care readiness. We used accuracy, Wilson 95% confidence intervals (CIs), macro-F1, Cohen's κ, paired bootstrap differences, and exact McNemar testing.

Results. Gold current risk comprised 72 low, 82 moderate, and 26 high cases. Temporal reconciliation classified 180/180 cases correctly on the binary endpoint, versus 108/180 for the baseline (difference 40.0 percentage points; bootstrap 95% CI 32.8–47.2; exact McNemar $p < 0.001$). Three-class accuracy was 88.9% (160/180; 95% CI 83.5–92.7), macro-F1 was 0.852, and κ was 0.824, with all errors at the moderate–high boundary. Five structured phenotype endpoints reached 100% accuracy.

Conclusion. In this synthetic benchmark, temporal and source-aware reconciliation kept historical high-risk language from dominating current-risk classification. The results show construct-aligned technical validity. They do not show clinical deployment validity, and external validation on real notes with clinician adjudication is required. For every packet the explainable trace retained source, time, screen status, barrier status, and context.

Copyright: © 2026 The Author(s). Published by HM Publisher in collaboration with CMHC Research Center on behalf of Scientia Psychiatrica. The authors retain copyright of their work.

Access licence: This is an open access article distributed under the terms of the Creative Commons Attribution–NonCommercial–ShareAlike 4.0 International Licence (CC BY-NC-SA 4.0), which permits non-commercial use, sharing, adaptation and reproduction in any medium, provided the original author and source are credited and any derivative work is distributed under the same licence.

1. Introduction

Electronic health records hold two kinds of information about psychiatric care. Structured fields record selected diagnoses, medications, observations, and scale responses. Free-text notes preserve the reasoning that links those observations to a patient's changing state. During psychiatric hospitalization this narrative layer matters a great deal, because risk is not a fixed attribute. Suicidal ideation may be intense at admission and recede later. Psychotic symptoms may fluctuate. Collateral information can confirm or challenge what a patient reports, and a safety plan may move from incomplete to workable. Housing, supervision, and follow-up arrangements may decide whether a lower level of care is feasible. Recent psychiatric clinical-NLP studies show the value of extracting phenotypes, psychosis episodes, transdiagnostic features, cohort outcomes, and occupational context from routine notes.[1–5] A system that finds a phrase without knowing when it applied, who recorded it, whether it was negated, or whether later evidence replaced it can produce a summary that is formally recognizable yet clinically misleading.

Clinical natural language processing (NLP) handles contextual phenomena through temporal normalization, retrieval context, and transparent scope rules. TimeX formalized phenotype-onset extraction.

Multi-corpus retrieval work showed that clinical context can rival model choice. ConText set out a durable framework for negation, experiencer, and temporality.[6–8] Interpretability is therefore not an obsolete property of simple systems. In safety-sensitive tasks an explicit rule hierarchy can be useful precisely because an investigator can trace which source, time point, and contextual trigger produced a label. Psychiatric risk extraction brings difficulties beyond ordinary entity recognition. First, current and historical risk often coexist in one packet. A prior attempt, an admission-level high-risk statement, or a copied-forward warning may remain relevant background without describing the present. Second, psychiatric notes carry clinically meaningful uncertainty. Declining to answer a suicide-intensity item is not the same as a negative answer, and an incomplete or malformed scale should never be turned into an invented score. Third, discharge or transition readiness is a synthesis, not a single-note fact. A favorable screen cannot by itself settle concerns about unstable housing, unreliable monitoring, an incomplete crisis plan, persistent psychosis, or substance-related risk. Recent research has shown that NLP can recover social determinants from free text and augment sparse structured fields, while also revealing sensitivity to extraction method and local documentation practice.[9–11] Fourth, the authority of a document depends on the question asked. A recent structured screen may be the best source for an item response. A psychiatric progress note may be best for current clinical interpretation. A social-work note may be best for unresolved transition barriers.

Detecting suicidality and supporting clinical risk assessment are different tasks. Contemporary studies have used weak supervision, rule-based systems, deep learning, cross-system portability analyses, and NLP-enriched electronic health records to identify suicidal ideation or improve risk prediction.[12–16] Portability remains a central concern, since performance can change across institutions, documentation cultures, and electronic-record systems. A benchmark result should therefore be described by the target it actually measured. Extracting a generator-defined label correctly from synthetic, templated notes shows that an algorithm matches the benchmark's constructs. It does not show that the same algorithm can diagnose, triage, discharge, or predict outcomes for real patients. Even so, synthetic clinical records are valuable in early technical work. They can be distributed without exposing protected health information. They allow controlled construction of difficult documentation patterns. They also support reproducible tests that would be impractical to share with real psychiatric records. ClinAuthBench v1 comprises dense 72-hour adult inpatient psychiatric packets that contain multiple note types, structured gold labels, trajectory metadata, transition barriers, evidence anchors, and explicit constraints on interpretation.[17] Its cases are built to test current-versus-historical risk, contradictions across documents, missing or stale evidence, and lower-level-of-care reasoning. These properties make the benchmark suitable for isolating temporal-reconciliation behavior before any claim of real-world performance is considered.

Temporal representation is a broader challenge for clinical NLP. Order in clinical narratives cannot be reduced to dates alone; it also depends on relations among events, linguistic cues, source context, and inferential status.[6,8] In longitudinal psychiatric notes the problem becomes operational. A high-risk statement from admission should not be silently erased. Neither should it automatically override a later, authoritative assessment. The output must preserve both the historical evidence and the current state. This motivates an architecture that records document time, source type, extracted state, contextual trigger, and any unresolved barrier, rather than returning a decontextualized label.

Existing work establishes the importance of contextual status, transparent extraction, and external validation. It does not, however, directly quantify how far a source-aware temporal hierarchy changes risk classification in dense synthetic psychiatric packets relative to a maximum-risk strategy. The novelty of this study lies in a direct, paired evaluation of an auditable five-stage temporal-reconciliation architecture. The architecture preserves historical evidence and selects the current state through endpoint-specific source authority and unresolved-barrier logic. We aimed to compare temporal reconciliation with a maximum-risk-across-record baseline for binary current risk, to quantify three-class performance and boundary errors, to evaluate five structured psychiatric phenotypes, and to characterize performance across predefined documentation challenges and generator types.

2. Methods

2.1. Study design and setting

We conducted a retrospective secondary validation study of a fixed, publicly available synthetic benchmark. Two deterministic strategies were evaluated on the same 180 chart packets: an explainable temporal-reconciliation pipeline and a maximum-risk-across-record baseline. The unit of analysis was one 72-hour synthetic inpatient psychiatric packet. The analysis was designed as a technical validation of extraction and reconciliation behavior. It was not a diagnostic-accuracy study in patients. Nor was it a clinical prediction-model development study. Reporting was informed proportionately by the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis, artificial intelligence extension (TRIPOD+AI), concerning data provenance, target definition, evaluation populations, reference standards, performance estimates, uncertainty, and limitations.[18]

2.2. Participants

Dataset, provenance, and analytical population. ClinAuthBench v1 is an English-language synthetic benchmark for adult inpatient psychiatric authorization and longitudinal evidence reasoning.[17] The release was obtained from Hugging Face at revision

fb4eb98a8b0cbd79860c52dd184c4b140d13edc2. The JSON file had SHA-256 checksum 9c0f357159b37d25293a1fc952e5d3d5d17ee91ee50ea150cb9aa9a16102a07b, matching the file in the public GitHub v1.0.0 tag. Each record contained a stable case identifier, a short title, a synthetic chart packet, benchmark metadata, a synthetic Markov decision process-style trajectory, structured gold fields, evidence anchors, documentation-challenge tags, and quality checks.

All 180 records were included. There was no sampling, exclusion, imputation, outcome adjudication, or participant-level linkage. Cases 1–120 used a rule-based synthetic trajectory generator, whereas cases 121–180 used a probabilistic generator. The benchmark represented seven diagnosis families and four primary documentation challenges: current-versus-historical risk, cross-document contradiction, transition-barrier reasoning, and missing, invalid, or stale evidence. The parser then recovered 4,331 form instances: 169 packets contained 24 forms and 11 contained 25 forms. The corpus and study profile are detailed in Table 1.

Table 1. Corpus and study profile.

Characteristic	Value
Analytical population	180 synthetic adult inpatient psychiatric packets
Observation window	72 hours per packet
Form instances	4,331 across 19 form types
Forms per packet	169 packets with 24 forms; 11 packets with 25 forms
Current-risk reference	Low, 72 (40.0%); moderate, 82 (45.6%); high, 26 (14.4%)
Documentation challenges	Transition barrier, 81; current–historical, 45; missing/invalid/stale, 32; contradiction, 22
Synthetic generators	Rule-based, 120; probabilistic, 60

Note: All counts were recalculated from the fixed JSON source. Percentages use n = 180.

2.3. Variables and instruments

Gold-standard fields and outcome normalization. The benchmark metadata served as the reference standard. The current-suicide-risk field was narrative and not a pre-coded ordinal variable, so normalization was specified before metric recalculation. Labels beginning with "low current" were mapped to low risk, labels containing "moderate-to-high," "assumed high," or "high current risk" were mapped to high risk, and the remaining current-risk descriptions that indicated continued review or monitoring were mapped to moderate risk. The binary primary endpoint treated low risk as negative and combined moderate and high risk as elevated/uncertain. We chose this grouping because the operational question was whether a packet could be separated from cases that still required elevated-risk review, and the question was not whether a deterministic system could replace clinician judgment among moderate and high states.

Additional gold fields were primary diagnosis, psychosis present, substance-use categories, lower-level-of-care barriers, and lower-level-of-care readiness. Diagnosis was evaluated as seven-class exact agreement, whereas psychosis, substance use, any unresolved transition barrier, and lower-level-of-care readiness were evaluated as binary endpoints. Evidence-anchor coverage meant the availability and selection of the expected supporting form. It was not treated as token-level rationale faithfulness, nor as proof that an extracted span captured the full clinical reasoning.

Endpoints and performance measures. The primary endpoint was binary current-risk accuracy. Sensitivity, specificity, positive predictive value, negative predictive value, and F1 were calculated using elevated/uncertain risk as the positive class. The primary contrast was the paired difference in accuracy between temporal reconciliation and the maximum-risk baseline.

Secondary endpoints included three-class current-risk accuracy, class-specific precision, recall, F1, macro-F1, Cohen's κ, and a confusion matrix. Structured phenotype endpoints were summarized with accuracy and, where binary, sensitivity, specificity, and F1. Exploratory analyses stratified three-class accuracy by documentation challenge and generator. These strata were interpreted descriptively because some contained small samples or did not include all risk classes.

2.4. Data collection procedure

Form segmentation and temporal ordering. The parser segmented packets using the FORM: marker and extracted ISO-8601 creation timestamps. Forms were sorted chronologically within each packet, and repeated note types kept their original order. For each prediction the pipeline preserved the form name, timestamp, extracted content state, and rule trigger. A later screen could not delete historical evidence. Admission risk stayed available as context even when it no longer determined the current label.

For the primary risk task, the hierarchy gave priority to four sources: the most recent Columbia-Suicide Severity Rating Scale (C-SSRS)-related screen, the most recent psychiatric progress assessment, the most recent discharge-planning or social-work barrier note, and the interdisciplinary treatment summary. For transition readiness, the latest social-work evidence was treated as authoritative for housing, monitoring, transportation, follow-up, and unresolved placement barriers. Psychiatric readiness remained a separate clinical input. Repeated nursing and medication-response notes served as contextual evidence only. They did not automatically override a later specialized assessment.

Explainable temporal-reconciliation pipeline. The deterministic pipeline performed five ordered operations. First, it detected target concepts and structured responses, including suicide-screen answers, diagnostic phrases, psychosis indicators, substance-use mentions, barrier statements, and readiness statements. Second, it applied local contextual rules for negation, refusal, uncertainty, and historical status, and third, it assigned source authority according to the endpoint. Fourth, it reconciled evidence chronologically, retaining the latest authoritative state while preserving higher-risk historical evidence in the trace. Fifth, it emitted both a categorical prediction and a compact evidence record containing the selected form, time, screen state, barrier state, and rule path. The complete architecture is illustrated in Figure 1.

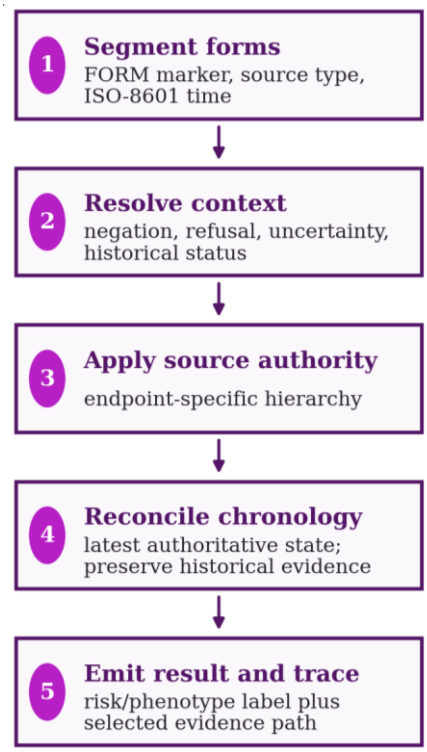


Figure 1. Five-stage temporal-reconciliation architecture. Endpoint-specific source authority is applied after contextual interpretation and before chronological reconciliation; the output preserves the current label and its evidence trace.

For current risk, an all-negative recent screen could support low risk only when the most recent transition evidence contained no unresolved inpatient-level barrier. Refusal of suicide-intensity items was retained as high/uncertain rather than numerically scored. Positive screening evidence was integrated with psychiatric context and unresolved barriers to distinguish moderate from high risk. The rule set did not infer a clinical disposition. A lower-level-of-care-ready label represented agreement with the benchmark's transition construct, not an independent discharge recommendation.

The structured phenotype rules were deliberately transparent. Primary diagnosis was mapped from the latest authoritative assessment. Psychosis was positive when the current diagnostic/assessment context contained a psychotic-spectrum condition or psychotic features. Substance use was positive only for affirmed mentions in diagnosis, history, or toxicology; negated and historical-only statements were retained as context. Transition barriers and readiness came from the latest dedicated social-work note. The rules were deterministic and called no generative model or external clinical service. The rules and the outcome normalization were specified on the same 180 records, and no holdout set was used.

Comparator. The baseline assigned the highest risk category mentioned anywhere in the packet, regardless of document date, negation, or later resolution. Because admission documents in this benchmark commonly contained high-risk language, this baseline operationalized a recognizable failure mode of longitudinal extraction: treating the record as an unordered bag of terms. It was not the strongest available machine-learning system. Its role was to isolate the value added by temporal and source-aware reconciliation relative to a deliberately simple, clinically unsafe heuristic.

2.5. Statistical analysis

Proportions were accompanied by two-sided 95% Wilson CIs. The paired accuracy difference was calculated as temporal accuracy minus baseline accuracy at the case level. Its 95% CI was estimated by nonparametric bootstrap sampling of cases with replacement for 5,000 replicates using seed 20260901. Exact two-sided McNemar testing compared discordant correct/incorrect pairs for the binary endpoint, and Cohen's κ estimated agreement beyond chance for three-class risk. Macro-F1 gave equal weight to the three risk classes. No adjustment for multiplicity was applied because secondary and subgroup analyses were

descriptive or exploratory and were not used to declare independent confirmatory findings.

2.6. Ethical approval

Ethical approval was granted by the CMHC Research Center Ethics Committee Indonesia (No. 2026/0301). Informed consent was not applicable because the study used only fully synthetic records and did not involve real participants or patient data. The analytical governance emphasized transparency, human oversight, traceability, equity, and prevention of premature clinical use.

Reproducibility and governance. The analysis was conducted locally. The case package preserves the source DOCX, fixed JSON, tagged repository snapshot, checksum, executable metric-validation script, executed audit notebook, machine-readable outputs, and publication figures. No record text was sent to an external clinical AI service. ClinAuthBench is distributed under CC BY 4.0, and the dataset states that all cases are synthetic and contain no real patient records, protected health information, real names, facilities, staff identifiers, medical-record numbers, phone numbers, or addresses.

3. Results

3.1. Corpus composition and audit

All 180 case identifiers were unique, all records carried the synthetic-data flag, and all required gold fields for the reported endpoints were present. The parser recovered 4,331 form instances across 19 form types. Gold current risk comprised 72 low-risk cases (40.0%), 82 moderate-risk cases (45.6%), and 26 high-risk cases (14.4%); the binary elevated/uncertain group therefore contained 108 cases (60.0%). Diagnosis groups were nearly balanced, with 25 or 26 cases in each of seven families. Documentation challenges comprised 81 transition-barrier cases, 45 current-versus-historical cases, 32 cases with missing/invalid/stale evidence, and 22 cross-document contradiction cases. The rule-based and probabilistic generators contributed 120 and 60 cases, respectively, as detailed in Table 1.

3.2. Primary binary endpoint

Temporal reconciliation correctly classified all 180 cases as low versus elevated/uncertain, yielding accuracy 100.0% (95% CI 97.9–100.0). Sensitivity, specificity, positive predictive value, negative predictive value, and F1 were each 1.000. The maximum-risk baseline correctly classified 108/180 cases, for accuracy 60.0% (95% CI 52.7–66.9). Baseline sensitivity for elevated/uncertain risk was 100.0%, but specificity was 0.0% because all 72 low-risk cases were classified as elevated. The primary and secondary endpoint estimates are detailed in Table 2.

Table 2. Performance of the maximum-risk baseline and temporal-reconciliation pipeline.

Endpoint	Maximum-risk baseline	Temporal reconciliation
Binary current risk, correct/n (%)	108/180 (60.0)	180/180 (100.0)
Three-class current risk, correct/n (%)	26/180 (14.4)	160/180 (88.9)
Three-class macro-F_1	0.084	0.852
Three-class Cohen's κ	0.000	0.824
Low precision / recall / F_1	Not estimable / 0.000 / 0.000	1.000 / 1.000 / 1.000
Moderate precision / recall / F_1	Not estimable / 0.000 / 0.000	0.943 / 0.805 / 0.868
High precision / recall / F_1	0.144 / 1.000 / 0.252	0.579 / 0.846 / 0.688
Diagnosis	Not assessed	180/180 (100.0)
Psychosis	Not assessed	180/180 (100.0)
Substance use	Not assessed	180/180 (100.0)
Transition-barrier status	Not assessed	180/180 (100.0)
Lower-level-of-care readiness	Not assessed	180/180 (100.0)

Note: Binary accuracy difference, 40.0 percentage points (bootstrap 95% CI, 32.8–47.2); exact two-sided McNemar $p < 0.001$. Three-class temporal accuracy 95% CI, 83.5–92.7. F_1, F1 score. CI, confidence interval; κ, Cohen's kappa.

The paired difference in accuracy was 40.0 percentage points. Across 5,000 bootstrap resamples, the 95% CI was 32.8–47.2 percentage points. Seventy-two cases were correct only under temporal reconciliation, whereas no case was correct only under the baseline. The exact two-sided McNemar test gave $p < 0.001$ (4.24×10^{-22}). The result indicates that the baseline's apparent sensitivity was obtained by labeling every packet as elevated, whereas temporal reconciliation recovered low-risk current states without losing elevated-risk cases within this benchmark.

3.3. Three-class current-risk classification

The maximum-risk baseline assigned every packet to the high-risk class. Its three-class accuracy was therefore 14.4%, macro-F1 was 0.084, and Cohen's κ was 0.000. Temporal reconciliation correctly classified 160/180 cases, for accuracy 88.9% (95% CI 83.5–92.7), macro-F1 0.852, and κ 0.824. Low-risk classification was perfect. Moderate-risk recall was 80.5% with precision 94.3% and F1 0.868. High-risk recall was 84.6%, but precision was 57.9% and F1 was 0.688 because moderate cases were more often escalated to

high than high cases were reduced to moderate. The complete confusion matrices are visualized in Figure 2.

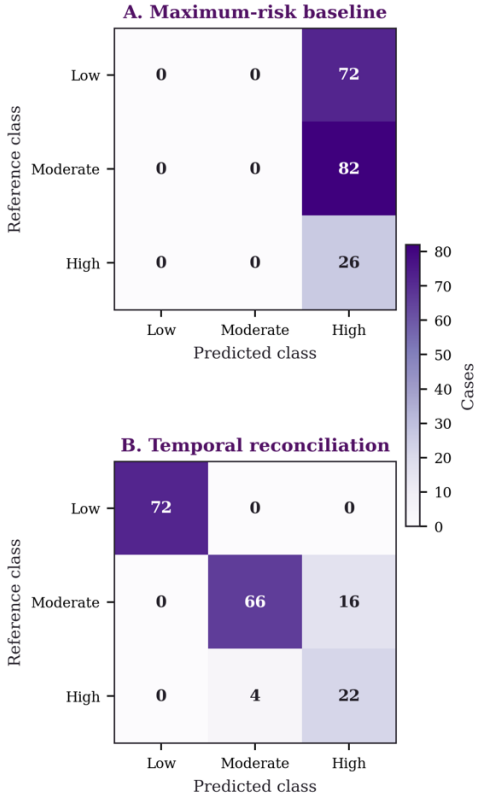


Figure 2. Three-class confusion matrices for the maximum-risk baseline and temporal-reconciliation pipeline. The baseline assigned all 180 packets to high risk; temporal reconciliation correctly classified all 72 low-risk packets, 66 of 82 moderate-risk packets, and 22 of 26 high-risk packets.

All 20 errors occurred at the moderate–high boundary: 16 gold-moderate cases were classified as high and four gold-high cases were classified as moderate. There were no low-to-elevated or elevated-to-low errors. Thus, collapsing moderate and high into the primary elevated/uncertain category produced perfect internal binary agreement even though clinically meaningful ordinal uncertainty remained.

3.4. Structured phenotype extraction

Primary diagnosis, psychosis, substance use, transition-barrier status, and lower-level-of-care readiness each achieved 180/180 exact agreement with benchmark labels (accuracy 100.0%; Wilson 95% CI 97.9–100.0). These results should be read together with how the corpus was generated. The benchmark uses controlled form structures, repeated linguistic patterns, and gold fields constructed from the same synthetic state representation. Perfect agreement therefore demonstrates that the deterministic rules recovered the intended benchmark constructs. It does not establish that equivalent concepts can be extracted perfectly from real electronic health records, where abbreviations, note styles, copying, local templates, contradictory assessments, and incomplete documentation may differ substantially.

3.5. Exploratory subgroup performance

Three-class accuracy was 100.0% in the 45 current-versus-historical cases. It was 86.4% in 81 transition-barrier cases, 84.4% in 32 missing/invalid/stale-evidence cases, and 81.8% in 22 cross-document contradiction cases. Accuracy was similar for records produced by the rule-based generator (89.2%; 107/120) and probabilistic generator (88.3%; 53/60). These endpoint and subgroup estimates are compared in Figure 3. The subgroup estimates were descriptive. Differences were not interpreted as causal effects of generator or challenge type, and macro-F1 was not compared across strata lacking the same class composition.

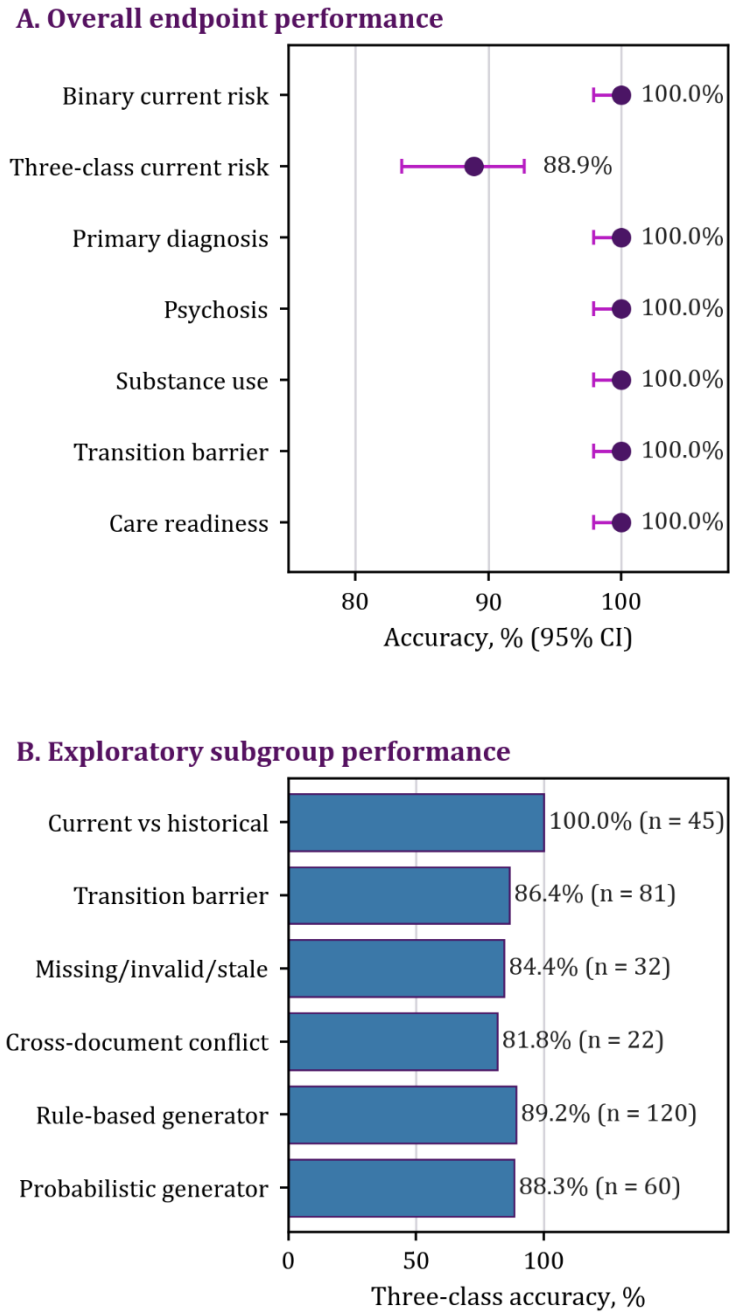


Figure 3. Endpoint and subgroup performance. Panel A shows overall accuracy with 95% confidence intervals where applicable; panel B shows descriptive three-class accuracy by documentation challenge and synthetic generator.

3.6. Error patterns and evidence trace

The dominant error pattern was moderate-to-high escalation. In these packets, a positive recent screen combined with psychosis or substance-related context triggered a high rule even though the benchmark retained a moderate label. This behavior increased caution but reduced high-class precision. The four high-to-moderate errors occurred when the terminal screen or psychiatric wording appeared less acute while cravings, placement barriers, or other contextual factors sustained a high gold label. The complete error distribution is detailed in Table 3. These errors show that form priority and chronology are necessary but not sufficient: the rules also require a calibrated conflict-resolution policy across symptoms, screening results, and transition evidence.

Table 3. Residual error distribution for three-class current-risk classification.

Error pattern	n	Interpretation
Low reference classified as moderate or high	0	No low-to-elevated error
Moderate reference classified as high	16	Predominant false escalation
High reference classified as moderate	4	False reduction across the moderate–high boundary
High or moderate reference classified as low	0	No elevated-to-low error
Total three-class errors	20	All errors confined to the moderate–high boundary

Note: Reference counts were low, 72; moderate, 82; and high, 26.

Each temporal prediction retained the selected screening state, latest barrier status, and rule path, preserving the distinction between current and historical evidence. Token-level rationale fidelity and clinician comprehension were not evaluated.

4. Discussion

4.1. Principal findings

This study shows a large effect of temporal and source-aware reconciliation on psychiatric risk phenotyping in a controlled longitudinal benchmark, and the rule path of every prediction can be traced. The maximum-risk strategy found high-risk language in every packet. It therefore reached perfect sensitivity for the binary elevated/uncertain class but could not recognize improvement: specificity was 0%, three-class accuracy was 14.4%, and κ was 0. Temporal reconciliation changed that behavior by separating the evidence that represented the current state from the evidence that should stay as historical context, and on the binary endpoint it recovered all 72 low-risk cases while keeping all 108 elevated/uncertain cases. The paired gain of 40.0 percentage points was more than an aggregate difference. Every discordant pair favored the temporal approach.

The three-class result adds a qualification. Accuracy of 88.9%, macro-F1 of 0.852, and κ of 0.824 indicate strong internal agreement. Yet 20 packets remained uncertain at the moderate–high boundary. High-class precision was 57.9%, so caution-oriented escalation produced a meaningful number of false high labels relative to the benchmark, and in a safety-sensitive setting whether that trade-off is acceptable depends on intended use, downstream action, alert burden, and clinician oversight. The study supports a hierarchy for organizing evidence. It does not support a claim that ordinal suicide risk can be automated from chart text.

4.2. Comparison with the existing literature

Why chronology changed the result. The baseline failed because it answered a different question from the one the endpoint asked. It effectively asked whether high-risk language had ever appeared during the 72-hour packet. The target asked which risk state the most recent authoritative evidence supported once unresolved barriers were considered. Both questions are legitimate. They are not interchangeable. Historical risk matters clinically and should stay visible, but it should not quietly become the current state just because a system lacks a temporal model.

This distinction agrees with research on clinical context and temporal representation.[6–8] Lexical detection alone cannot tell whether an event is present, past, hypothetical, copied forward, or superseded, and psychiatric documentation sharpens the problem because risk statements are repeated on purpose for safety and continuity. A system that treats repetition as independent corroboration may grow more confident exactly when text has been copied and not newly assessed. The temporal pipeline avoided this by preserving chronology and selecting endpoint-specific sources.

Source authority mattered as much as time. The latest document is not always the most relevant one. A late nursing note may record behavior without replacing a psychiatric assessment of suicide risk, and a C-SSRS response may describe item endorsement without resolving whether housing, monitoring, or a crisis plan supports transition. A social-work note may be authoritative for placement barriers but not for diagnosis. The architecture therefore paired temporal order with a source hierarchy, and it did not apply a universal "last mention wins" rule.

Relation to psychiatric NLP evidence. The results fit a wider literature in which suicide-related clinical NLP covers domain-specific vocabulary development, social-determinant enrichment, post-discharge prediction, adolescent inpatient outcomes, landmark modeling, first-attempt prediction, and multimodal comparison with clinician assessment and self-report.[19–25] These studies show that unstructured documentation can add clinically relevant signal. They also show that target definition and intended use decide what a model output means. The present study adds a focused demonstration: the value of an interpretable system may lie less in finding a term than in reconciling several true statements across time.

The findings also complement recent evaluations of personalized decision support, language-model detection of self-injury, multimodal pediatric surveillance, privacy-preserving local models, preprocessing strategies, psychotherapy-service characterization, sentiment lexicons, and multisite readmission analyses.[26–33] Read together with cross-system portability evidence, these studies explain why perfect benchmark extraction cannot be assumed to carry over to another corpus. Templates, local abbreviations, note ownership, assessment routines, and population characteristics all contribute to dataset shift.

Suicide prediction and suicide-language extraction must also be kept apart conceptually. Work on trustworthy crisis prediction, network-scale NLP validation, tailored clinical pipelines, symptom-outcome trajectories, serious-mental-illness trajectories, antidepressant response, opioid phenotyping, and scalable incident detection shows how varied the tasks grouped under clinical NLP are.[34–42] This study did not predict suicide attempts, mortality, self-harm, readmission, or treatment response. It compared model output with a synthetic reference label for current-risk reconciliation. The correct clinical reading is narrow. Temporal and source-aware rules matched the benchmark's current-risk construct better than a maximum-risk heuristic did.

4.3. Interpretation and mechanism

Meaning of perfect structured-field performance. Five structured endpoints reached 100% accuracy. Such results need explanation, because they can be persuasive and misleading at once. In this corpus, controlled templates aligned with structured metadata produced the diagnoses, psychosis indicators, substance-use mentions, barrier statements, and readiness language. A transparent mapping rule could therefore retrieve them with little linguistic ambiguity. Hence the result is expected. The lower Wilson confidence bound of 97.9% expresses binomial uncertainty within the 180 benchmark cases. It says nothing about uncertainty under changes in documentation style, population, institution, language, or annotation policy.

This is construct alignment, not clinical validity. A benchmark can show whether an implementation processes faithfully the conditions it was designed to contain. By itself it cannot establish transportability to real notes, the validity of the underlying clinical construct, fairness across patient groups, or better patient outcomes. Recent work on synthetic-to-real drift, social-determinant documentation, standardized addiction notes, psychosis-severity extraction, depression-severity estimation, shared-task annotation, and multilingual clinical corpora shows why realism, language, institution, and annotation policy must be tested explicitly.[43–50] Synthetic regularity can make extraction easier than natural clinical text. The most useful response to perfect performance is a precise statement of what was and was not tested.

Statistical and practical importance. The 40-point binary gain corresponds to 72 packets whose improving current state the baseline could not recognize. Its bootstrap CI stayed far from zero, and every McNemar discordance favored temporal reconciliation. This statistical separation reflects a structural contrast: the baseline cannot resolve improvement when every packet contains high-risk admission language. At the ordinal level, false-high classifications could create alert burden, while high-to-moderate classifications could understate severity if treated as dispositive. The evidence trace is therefore a safety mechanism, and future evaluation should test whether clinicians understand it and use it appropriately.

Alternative explanations. Several alternative explanations deserve attention. First, performance may reflect template recognition more than temporal reasoning. Repeated phrases and form names in the benchmark make source identification unusually reliable. Second, the rules were specified and evaluated on the same corpus, so design choices could align with benchmark regularities. Third, synthetic states generated the narrative gold labels, and blinded clinicians did not adjudicate them independently. Fourth, binary perfection partly results from collapsing moderate and high risk, the boundary where every observed error occurred. Fifth, evidence-anchor coverage measured whether an expected form was available, not semantic faithfulness at the span level.

None of these explanations negates the observed paired difference. The baseline still failed precisely because it ignored a property the benchmark was designed to test. They do, however, change the general claim. The fitting conclusion is that the benchmark provides a successful proof of concept for an architecture, not a validated clinical instrument. A stronger test would freeze the rules before exposure to an external corpus and blind adjudicators to model output. It would also define current-risk time anchors prospectively and report performance by institution, note type, diagnosis, language, and demographic group.

4.4. Implications for practice and policy

Implications for system design. Four design implications follow. First, a psychiatric NLP system should store event time and document time separately where possible. A note written today may describe an event from last year. A copied statement may carry neither a new event nor a new assessment. Second, endpoint-specific source authority should be explicit, since diagnostic, screening, behavioral, and social-transition questions may need different hierarchies. Third, refusal and missingness must stay distinct from negative findings. Turning uncertainty into a favorable score is unsafe and analytically invalid. Fourth, output should include the evidence path and unresolved conflicts, not only a categorical label.

Intrinsic transparency has advantages here. Decision rules can be reviewed, versioned, tested against known edge cases, and shown to clinicians in a form that supports correction. A rule path is valuable only if it faithfully reflects the computation and helps a qualified user understand the evidence. The present study established computational traceability. It did not test explanation effectiveness.

An operational system would also need explicit abstention, action-specific error costs, multilingual validation, and reproducibility controls. Missing authoritative screens, irresolvable note conflicts, or unreliable timestamps should yield “insufficient or conflicting evidence,” with coverage and review burden reported next to accuracy. Local adaptation should rely on locally authored notes and clinicians, because psychiatric meaning travels through idiom and institutional practice. Reproducibility should preserve

the dataset revision, preprocessing, rule configuration, software environment, seeds, and output schema. External studies should keep the maximum-risk heuristic for mechanistic continuity and add contemporary rule-based and model-based comparators under identical patient-level splits and adjudication.

External validation pathway. A staged validation program is required before any clinical use. The first stage should be retrospective external validation on real, deidentified notes from at least one institution that took no part in rule specification. A psychiatric corpus would need dedicated annotation of current risk, historical risk, time anchor, negation, refusal, barrier status, and evidence spans.

The reference standard should rely on at least two independent psychiatric clinicians working under a predefined adjudication process. Inter-rater agreement should be reported before model evaluation, because ambiguous moderate–high cases may reflect genuine clinical disagreement and not algorithm error. Rules should be frozen before labels are opened, and evaluation should separate patients or sites. Performance should include class-specific sensitivity, specificity, precision, calibration where relevant, error severity, abstention, and confidence intervals. Subgroup analyses should be prespecified and sized to assess fairness across age, sex, language, diagnostic group, race/ethnicity where lawful and appropriate, and site.

The next stage should be prospective silent-mode evaluation, in which the system processes live documentation without affecting care. Such a stage can measure latency, missing forms, unexpected note types, rule firing, drift, and disagreement with clinician assessments. Only once safety thresholds and governance are defined should an interface study ask whether evidence traces improve clinician review without increasing automation bias. A clinical-impact study would be justified only after the technical, external, fairness, and human-factor gates are met.

Governance and safety. Psychiatric risk tools work in a domain where errors may influence observation, liberty, stigma, resource allocation, and access to care. Governance must, as a first step, define intended use, prohibited use, accountable users, monitoring, appeal, and shutdown procedures. For the present architecture these safeguards imply three things. Labels should not be the sole grounds for diagnosis, discharge, involuntary treatment, or triage. Qualified clinicians must retain authority. Every output should be open to challenge against its evidence.

4.5. Strengths and limitations

Strengths. The study has several strengths. It used the complete fixed benchmark and not a convenience subset. It preserved the dataset revision and checksum and compared both systems on identical packets. It reported paired uncertainty and exact testing, and presented class-specific errors alongside accuracy. It examined predefined documentation challenges and generators and kept an interpretable evidence path.

Limitations. The limitations mark the boundary of inference. All records were synthetic, English-language, adult inpatient psychiatric packets produced from controlled templates. The gold standard was generator-derived and not adjudicated by blinded clinicians. Rules were developed and evaluated on the same benchmark. There was no independent holdout, external institution, prospective protocol, or live workflow, and the primary endpoint collapsed moderate and high risk. Evidence fidelity was measured at form level, not at token span or causal rationale, and no real demographic attributes were available for fairness analysis. The source-referenced custom case-level prediction output was unavailable for independent case assignment, although its aggregate confusion matrices and paired counts were internally consistent and fully recalculated. These conditions do not prevent technical interpretation. They do preclude claims of clinical validity or deployment readiness.

5. Conclusion

Across 180 synthetic longitudinal inpatient psychiatric packets, explicit temporal and source-aware reconciliation outperformed a maximum-risk-across-record strategy. The main benefit was not better detection of risk words. It was the correct assignment of those words to a current or historical state. Binary accuracy rose from 60.0% to 100.0%, and three-class accuracy reached 88.9%, with every remaining error confined to the moderate–high boundary. Structured phenotype extraction matched benchmark labels perfectly because of controlled templates and construct alignment.

These findings support an explainable architecture that preserves chronology, source authority, negation, refusal, unresolved transition barriers, and an auditable evidence path. They do not support autonomous diagnosis, discharge decisions, observation changes, or triage for real patients. The next scientific step is a prospectively specified external validation using real deidentified psychiatric notes, independent clinician adjudication, span-level evidence assessment, fairness analysis, and prospective silent-mode testing. This study is therefore a transparent and reproducible proof of concept with an explicit clinical safety boundary.

Declarations

Use of Generative Artificial Intelligence. The authors declare that no generative artificial intelligence tool was used in the conception, research, analysis, drafting, revision or preparation of this manuscript.

Author Contributions. D.A.: conceptualization, methodology, software, validation, formal analysis, investigation, data curation, visualization, writing – original draft. S.I.: conceptualization, resources, supervision, project administration, writing – review & editing. Both authors have read and approved the final version of the manuscript and agree to be accountable for all aspects of the work.

Funding. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest. The authors declare that they have no conflict of interest.

Ethical Approval. Ethical approval was granted by the CMHC Research Center Ethics Committee Indonesia (No. 2026/0301).

Informed Consent. Informed consent was not applicable because all records were fully synthetic and no real participants or patient data were used.

Data Availability. ClinAuthBench v1 is publicly available from Hugging Face at https://huggingface.co/datasets/Shivi1982/clin-auth-bench. The analyzed file was fixed at revision fb4eb98a8b0cbd79860c52dd184c4b140d13edc2 with SHA-256 checksum 9c0f357159b37d25293a1fc952e5d3d5d17ee91ee50ea150cb9aa9a16102a07b. Public schema and generation resources are available at https://github.com/Shivi1982/clin-auth-bench (tag v1.0.0).

Acknowledgements. None.

References

1. De La Hoz JF, Frydman-Gani C, Arias A, et al. Leveraging Natural Language Processing for Psychiatric Phenotyping from Spanish Electronic Health Records: Enabling the Investigation of Transdiagnostic Symptom Profiles at Scale. Complex Psychiatry. 2025;11(1):99-112. doi:10.1159/000546480.
2. Hua Y, Blackley SV, Shinn AK, et al. Identifying psychosis episodes in psychiatric admission notes via rule-based methods, machine learning, and pre-trained language models. Transl Psychiatry. 2025;15(1):520. doi:10.1038/s41398-025-03629-4.
3. Turner RJ, Coenen F, Roelofs F, et al. Information extraction from free text for aiding transdiagnostic psychiatry: constructing NLP pipelines tailored to clinicians' needs. BMC Psychiatry. 2022;22(1):407. doi:10.1186/s12888-022-04058-z.
4. Hansen L, Enevoldsen KC, Bernstorff M, et al. The PSYchiatric clinical outcome prediction (PSYCOP) cohort: leveraging the potential of electronic health records in the treatment of mental disorders. Acta Neuropsychiatr. 2021;33(6):323-330. doi:10.1017/neu.2021.22.
5. Chilman N, Song X, Roberts A, et al. Text mining occupations from the mental health electronic health record: a natural language processing approach using records from the Clinical Record Interactive Search (CRIS) platform in south London, UK. BMJ Open. 2021;11(3):e042274. doi:10.1136/bmjopen-2020-042274.
6. Chen F, Jiang S, Nguyen QM, et al. TimeX: Phenotype Onset Extraction from Clinical Narratives. Npj Health Syst. 2026;3(1):49. doi:10.1038/s44401-026-00099-8.
7. Mikkelsen Y. Clinical Context Variables Collectively Rival Model Choice in Embedding-Based Retrieval: Multi-Corpus Benchmark Study. JMIR Med Inform. 2026;14:e94241. doi:10.2196/94241.
8. Harkema H, Dowling JN, Thornblade T, et al. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. J Biomed Inform. 2009;42(5):839-851. doi:10.1016/j.jbi.2009.05.002.
9. Han S, Zhang RF, Shi L, et al. Classifying social determinants of health from unstructured electronic health records using deep learning-based natural language processing. J Biomed Inform. 2022;127:103984. doi:10.1016/j.jbi.2021.103984.
10. Lybarger K, Dobbins NJ, Long R, et al. Leveraging natural language processing to augment structured social determinants of health data in the electronic health record. J Am Med Inform Assoc. 2023;30(8):1389-1397. doi:10.1093/jamia/ocad073.
11. Wang B, Kabir D, Clark CR, et al. Extracting Social Determinants of Health From Electronic Health Records: Development and Comparison of Rule-Based and Large Language Model Methods. JMIR Med Inform. 2026;14:e89534. doi:10.2196/89534.
12. Cusick M, Adekkanattu P, Campion TR, et al. Using weak supervision and deep learning to classify clinical notes for identification of current suicidal ideation. J Psychiatr Res. 2021;136:95-102. doi:10.1016/j.jpsychires.2021.01.052.
13. Xie F, Ling Grant DS, Chang J, et al. Identifying Suicidal Ideation and Attempt From Clinical Notes Within a Large Integrated Health Care System. Perm J. 2022;26(1):85-93. doi:10.7812/tpp/21.102.
14. Vance LA, Way L, Kulkarni D, et al. Natural language processing to identify suicidal ideation and anhedonia in major depressive disorder. BMC Med Inform Decis Mak. 2025;25(1):20. doi:10.1186/s12911-025-02851-w.
15. Cusick M, Velupillai S, Downs J, et al. Portability of natural language processing methods to detect suicidality from clinical text in US and UK electronic health records. J Affect Disord Rep. 2022;10:100430. doi:10.1016/j.jadr.2022.100430.
16. Levis M, Levy J, Dent KR, et al. Leveraging Natural Language Processing to Improve Electronic Health Record Suicide Risk Prediction for Veterans Health Administration Users. J Clin Psychiatry. 2023;84(4):22m14568. doi:10.4088/jcp.22m14568.
17. Bhatia S. ClinAuthBench v1: a synthetic benchmark for inpatient health authorization reasoning. Hugging Face; 2026 [cited 2026 Sep 1]. Available from: https://huggingface.co/datasets/Shivi1982/clin-auth-bench
18. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378.
19. Morrow D, Zamora-Resendiz R, Beckham JC, et al. A case for developing domain-specific vocabularies for extracting suicide factors from healthcare notes. J Psychiatr Res. 2022;151:328-338. doi:10.1016/j.jpsychires.2022.04.009.
20. Mitra A, Pradhan R, Melamed RD, et al. Associations Between Natural Language Processing-Enriched Social Determinants of Health and Suicide Death Among US Veterans. JAMA Netw Open. 2023;6(3):e233079. doi:10.1001/jamanetworkopen.2023.3079.
21. Mitra A, Chen K, Liu W, et al. Post-discharge suicide prediction among US veterans using natural language processing-enriched social and behavioral determinants of health. Npj Ment Health Res. 2025;4(1):8. doi:10.1038/s44184-025-00120-2.
22. Carson NJ, Yang X, Mullin B, et al. Predicting adolescent suicidal behavior following inpatient discharge using structured and unstructured data. J Affect Disord. 2024;350:382-387. doi:10.1016/j.jad.2023.12.059.
23. Sheu YH, Sun J, Lee H, et al. An efficient landmark model for prediction of suicide attempts in multiple clinical settings. Psychiatry Res. 2023;323:115175. doi:10.1016/j.psychres.2023.115175.
24. Tsui FR, Shi L, Ruiz V, et al. Natural language processing and machine learning of electronic health records for prediction of first-time suicide attempts. JAMIA Open. 2021;4(1):ooab011. doi:10.1093/jamiaopen/ooab011.

25. Nock MK, Millner AJ, Ross EL, et al. Prediction of Suicide Attempts Using Clinician Assessment, Patient Self-report, and Electronic Health Records. JAMA Netw Open. 2022;5(1):e2144373. doi:10.1001/jamanetworkopen.2021.44373.
26. Gericke F, Voorspoels W, Peeters E, et al. Personalised machine-learning decision support for suicidal thoughts and behaviours in the psychiatric emergency department. Psychiatry Res. 2025;352:116698. doi:10.1016/j.psychres.2025.116698.
27. Martinez-Romo J, Araujo L, Reneses B. Guardian-BERT: Early detection of self-injury and suicidal signs with language technologies in electronic health reports. Comput Biol Med. 2025;186:109701. doi:10.1016/j.compbiomed.2025.109701.
28. Edgcomb JB, Saha A, Klomhaus A, et al. Detecting Pediatric Emergency Service Use for Suicide and Self-Harm: Multimodal Analysis of 3828 Encounters. JMIR Ment Health. 2026;13:e82371. doi:10.2196/82371.
29. Kormilitzin A, Joyce DW, Tsiachristas A, et al. Detection of Self-Harm in Electronic Mental Health Records Using Privacy-Preserving Local Language Models: Methodological Study. JMIR Ment Health. 2026;13:e87586. doi:10.2196/87586.
30. Li S, Levis M, DiMambro M, et al. Preprocessing of natural language process variables using a data-driven method improves the association with suicide risk in a large veterans affairs population. Comput Biol Med. 2025;189:109939. doi:10.1016/j.compbiomed.2025.109939.
31. Levis M, Dimambro M, Levy J, et al. Using Natural Language Processing to Evaluate Differences in Psychotherapeutic Services for Posttraumatic Stress Disorder in a Suicide-Risk-Stratified Veteran Sample. J Clin Psychiatry. 2026;87(2):25m15904. doi:10.4088/jcp.25m15904.
32. Bittar A, Velupillai S, Roberts A, et al. Using General-purpose Sentiment Lexicons for Suicide Risk Assessment in Electronic Health Records: Corpus-Based Analysis. JMIR Med Inform. 2021;9(4):e22397. doi:10.2196/22397.
33. Cliffe C, Cusick M, Vellupillai S, et al. A multisite comparison using electronic health records and natural language processing to identify the association between suicidality and hospital readmission amongst patients with eating disorders. Int J Eat Disord. 2023;56(8):1581-1592. doi:10.1002/eat.23980.
34. Msosa YJ, Grauslys A, Zhou Y, et al. Trustworthy Data and AI Environments for Clinical Prediction: Application to Crisis-Risk in People With Depression. IEEE J Biomed Health Inform. 2023;27(11):5588-5598. doi:10.1109/jbhi.2023.3312011.
35. Wang Y, Hilsman J, Li C, et al. Development and validation of natural language processing algorithms in the national ENACT network. J Clin Transl Sci. 2025;9(1):e199. doi:10.1017/cts.2025.10116.
36. Kocaman V, Cheng FY, Bonis J, et al. Exploring Named Entity Recognition Potential and the Value of Tailored Natural Language Processing Pipelines for Radiology, Pathology, and Progress Notes in Clinical Decision Support: Quantitative Study. JMIR AI. 2025;4:e59251. doi:10.2196/59251.
37. Patel R, Irving J, Brinn A, et al. Associations of presenting symptoms and subsequent adverse clinical outcomes in people with unipolar depression: a prospective natural language processing (NLP), transdiagnostic, network analysis of electronic health record (EHR) data. BMJ Open. 2022;12(4):e056541. doi:10.1136/bmjopen-2021-056541.
38. De la Hoz JF, Arias A, Service SK, et al. Characterisation of serious mental illness trajectories through transdiagnostic clinical features. Br J Psychiatry. 2026;228(5):446-453. doi:10.1192/bjp.2025.107.
39. Sheu YH, Magdamo C, Miller M, et al. Phenotyping antidepressant treatment response with deep learning in electronic health records. Transl Psychiatry. 2026;16(1):525. doi:10.1038/s41398-026-04266-1.
40. Taylor RA, Gilson A, Schulz W, et al. Computational phenotypes for patients with opioid-related disorders presenting to the emergency department. PLoS One. 2023;18(9):e0291572. doi:10.1371/journal.pone.0291572.
41. Poulsen MN, Freda PJ, Troiani V, et al. Classifying Characteristics of Opioid Use Disorder From Hospital Discharge Summaries Using Natural Language Processing. Front Public Health. 2022;10:850619. doi:10.3389/fpubh.2022.850619.
42. Walsh CG, Wilimitis D, Chen Q, et al. Scalable incident detection via natural language processing and probabilistic language models. Sci Rep. 2024;14(1):23429. doi:10.1038/s41598-024-72756-7.
43. Cornelius J, Rinaldi F. Measuring the gap: correlating synthetic-to-real drift with PHI de-identification performance. Genomics Inform. 2026;24(1):10. doi:10.1186/s44342-026-00072-9.
44. Bazoge A, Constant Dit Beaufils P, Hmitouch M, et al. Improving social determinants of health documentation in French electronic health records using large language models. Sci Rep. 2025;15(1):45427. doi:10.1038/s41598-025-29987-z.
45. Gireesh H, Shukla L, Shivaprakash P, et al. Language Models for Standardising Clinical Notes and Information Extraction in Addiction Psychiatry-An Empirical Study. Drug Alcohol Rev. 2026;45(1):e70059. doi:10.1111/dar.70059.
46. Matsumura M, Nishida K, Toyoda K, et al. Quantifying improvement of psychotic symptoms in clozapine-treated schizophrenia: clinical note analysis with large language models. Sci Rep. 2026;16(1):8835. doi:10.1038/s41598-026-39676-0.
47. Li K, Zirikly A, Collica SC, et al. Measuring Depression Severity With Clinical Global Impression-Severity Scale Scores From Clinical Notes Using Large Language Models: Validation Study. JMIR Form Res. 2026;10:e86906. doi:10.2196/86906.
48. McCoy TH, Castro VM, Perlis RH. Estimating depression severity in narrative clinical notes using large language models. J Affect Disord. 2025;381:270-274. doi:10.1016/j.jad.2025.04.014.
49. Lybarger K, Yetisgen M, Uzuner Ö. The 2022 n2c2/UW shared task on extracting social determinants of health. J Am Med Inform Assoc. 2023;30(8):1367-1378. doi:10.1093/jamia/ocad012.
50. Oliveira LESE, Peters AC, da Silva AMP, et al. SemClinBr - a multi-institutional and multi-specialty semantically annotated corpus for Portuguese clinical NLP tasks. J Biomed Semantics. 2022;13(1):13. doi:10.1186/s13326-022-00269-1.