Learning from Imperfect Patient-Reported Outcomes: An Observability-Aware Framework for Language-Based Clinical Phenotyping

Sheida Habibi[1], Dio Kavalieratos[2], Ravi B. Parikh[2], and Selen Bozkurt[1]

[1]*Biomedical Informatics Department, Emory University, Atlanta, GA 30322, USA*
[2]*Emory University School of Medicine, Atlanta, GA 30322, USA*
E-mail: sheida.habibi@emory.edu

Language models applied to clinical text are often evaluated against diagnosis codes or chart review, which reflect what clinicians recognized and documented rather than the full patient experience. Independently collected patient-reported outcomes reduce this dependence but create a challenge: when the reference standard and model input arise from different measurement channels, apparent errors may reflect either failure to recognize available evidence or absence of relevant evidence from the analyzed text. We introduce an observability-aware framework for evaluating clinical language models under this partial information. We demonstrate the framework in 1,725 palliative care patients without a prior depression diagnosis or antidepressant treatment, using Patient Health Questionnaire-8 (PHQ-8) scores ≥ 5 as the reference standard. Llama 3.3 70B Instruct classified clinical notes and portal messages, with predictions combined by logical OR. Clinical notes achieved sensitivity 0.67, specificity 0.58, and positive-class F1 0.64. Adding portal messages increased sensitivity to 0.71 and F1 to 0.65, recovering 34 additional positive patients while reducing specificity by 0.03. We defined lexicon-relative observability as the presence of at least one depressive-symptom domain in clinical notes. Approximately 80% of PHQ-8-positive patients were observable, with sensitivity 0.76 in this subgroup. Blinded review identified clinically meaningful evidence in 72.9% of lexicon-unobservable true positives versus 12.5% of lexicon-unobservable false negatives, with strong agreement ($\kappa = 0.85$). Content-conditioned abstention increased sensitivity and F1 but did not consistently reduce selective risk, indicating case enrichment rather than improved reliability. Observability-aware evaluation can help determine whether improvements should focus on stronger models, broader text sources, or direct outcome collection.

Keywords: clinical natural language processing; large language models; patient-reported outcomes; depression; palliative care.

1. Introduction

Depression is a leading cause of disability and is associated with reduced quality of life, increased healthcare utilization, suicide risk, and premature mortality.[1,2] Despite effective treatments, depressive symptoms remain frequently underrecognized because of stigma, limited access to mental health services, brief encounters, and reluctance to disclose emotional concerns.[3–5] This challenge is consequential in palliative care, where mood symptoms may be overlooked amid serious illness or difficult to distinguish from fatigue, sleep disturbance, appetite changes, and functional decline resulting from the underlying illness.[21,22]

Electronic health records (EHRs) offer an opportunity to identify depressive symptoms during routine care. EHR phenotyping based on diagnosis codes, medications, and billing data is useful for established disease but less effective before symptoms are formally recognized.

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

The authors have no conflicts of interest relevant to this article to disclose. This work was supported in part by a philanthropic gift from the Fraser Parker Foundation to Emory University's Division of Palliative Medicine. Dr. Bozkurt's effort was supported in part by the National Institute of Mental Health of the National Institutes of Health under award number R21MH139049.

Unstructured clinical text, including clinical notes, patients' portal messages and ambient transcripts of visits, can capture mood changes, emotional distress, psychosocial stressors, sleep problems, and functional impairment that precede diagnosis.[6–9] Advances in natural language processing (NLP), including transformer-based models and large language models (LLMs), have therefore enabled sophisticated depression phenotyping from clinical documentation.[10–15] These studies, however, have mostly relied on diagnosis codes, chart review, or expert-assigned labels as the reference standard. Such labels introduce a dependency that is analytically consequential: the reference standard and the model input derive from the same clinical encounter and the same clinician's judgment. Apparent model performance may therefore reflect the detection of documented clinical recognition rather than the patient symptoms, and the two are not interchangeable. Recognition is largely a precondition for coding, so patients whose symptoms are already recognized are, by construction, the patients for whom automated identification offers the least incremental value. Conversely, patients with unrecognized symptoms constitute the population of principal clinical interest, and this population cannot be operationally defined using a label derived from clinician recognition.

Independently collected patient-reported outcomes break this dependence. The Patient Health Questionnaire-8 (PHQ-8) measures symptom frequency over a defined recall period, whereas clinical text contains only information expressed, elicited, and documented within the available window. A patient may therefore meet the outcome threshold without corresponding depressive-symptom evidence in the model input. Conventional sensitivity treats such cases identically to those in which relevant evidence is present but missed, conflating limited textual observability with model-recognition failure. Thus, the patient-reported reference is not necessarily invalid but may be partially observable from the language modality being evaluated.

The text source further shapes what becomes observable. Clinical notes reflect information elicited during encounters and clinical interpretation but may omit concerns that patients do not disclose or clinicians do not document. Patient portal messages provide a complementary patient-generated channel for describing symptoms, emotional experiences, treatment concerns, and requests for support between visits. Prior studies have demonstrated their utility for identifying medication-related issues, adverse events, and mental health concerns, including depression-related content.[16–20] However, their incremental contribution beyond clinical notes remains insufficiently characterized.

Existing language-based phenotyping studies generally emphasize full-cohort discrimination without assessing if phenotype-relevant evidence is present in the analyzed text.[10–15] Few approaches quantify text-relative observability or determine whether restricting prediction to content-rich records reduces error or merely enriches the retained population for positive cases.

In this study we develop and apply an observability-aware evaluation of language-based depressive symptom detection against an independently collected patient-reported outcome. The analysis proceeds in three parts. First, we benchmark an instruction-tuned LLM applied to clinical notes and patient-generated portal messages and interpret its performance in relation to trivial, lexical, and supervised baselines commonly used in clinical phenotyping. Second, we partition outcome-positive patients by model prediction and by lexicon-observability, defined as the presence of explicit depressive symptom language in the analyzed text, which

separates cases missed despite available evidence from cases in which no such evidence was documented. We assessed the validity of this grouping through a blinded review of lexicon-unobservable cases. Third, we quantified the additional contribution of portal messages and, in a separate clinical-note-only analysis, evaluated whether the number of lexicon-detected symptom domains could serve as an abstention criterion across different coverage levels. We release the prompts, lexicon, and analysis code to support replication.

2. Methodology

Figure 1 summarizes the study workflow.

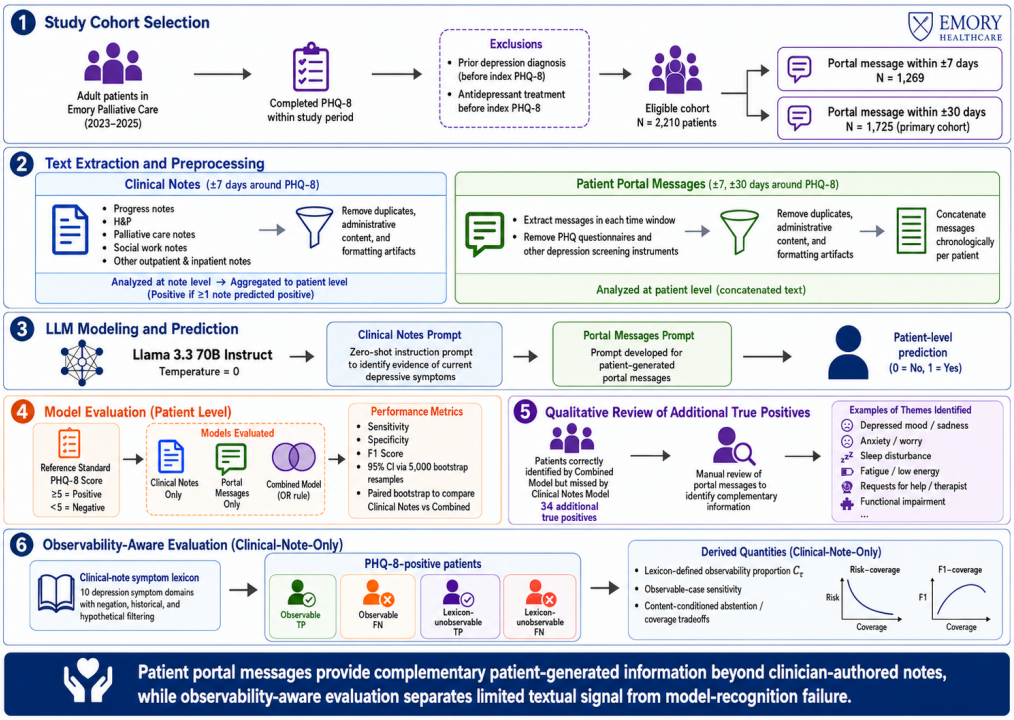


Fig. 1. Study workflow, including cohort selection, text preprocessing, LLM-based prediction, patient-level evaluation, and observability-aware error decomposition.

2.1. *Study design and cohort*

We included adult patients who received palliative care within Emory Healthcare and completed at least one PHQ-8 assessment between January 1, 2023, and December 31, 2025. The study was approved by the Emory University Institutional Review Board. Eligible patients had at least one PHQ-8 and at least one clinical note within ±7 days. For patients with multiple eligible assessments, the first was the index assessment.

To focus on unrecognized symptoms and to prevent the model from reading a diagnosis off the chart, we excluded 976 patients with a documented depression diagnosis (ICD-10-CM) or antidepressant exposure before the index PHQ-8, leaving 2,210 eligible patients. Among these, 1,269 had at least one portal message within ±7 days of the index assessment and 1,725 within ±30 days. The ±30-day portal-message cohort was designated as the primary analytic cohort

because it retained more eligible patients and therefore provided greater statistical precision, whereas the ±7-day portal-message cohort was used as a temporal sensitivity analysis.

2.2. *Reference Standard*

The independently collected PHQ-8 assesses eight depressive symptoms over the preceding two weeks, with items scored from 0 to 3 and total scores ranging from 0 to 24. Scores of 0–4 were classified as minimal or no symptoms and scores ≥5 as at least mild symptom burden, a threshold selected to prioritize sensitivity for early identification rather than clinical diagnosis.[23,24] Because the PHQ-8 and analyzed text represented distinct measurement channels, patients could meet the threshold without corresponding textual evidence or describe distress despite scoring below it. Performance therefore depended on both textual evidence availability and model recognition.

2.3. *Text Processing and Model Configuration*

We extracted clinical notes and patient-generated MyChart portal messages within the prespecified temporal windows. The symmetric ±7-day clinical-note window accommodated documentation that may be entered or finalized after the corresponding encounter while maintaining temporal proximity to the index assessment. Clinical notes included progress notes, palliative care notes, social work notes, and other documentation. Portal messages were ordered chronologically and concatenated into a patient-level document for each temporal window.

To minimize information leakage, embedded PHQ questionnaires and other structured depression-screening instruments were removed before inference, along with duplicate records, administrative or automated content, and formatting artifacts. Depressive symptoms were then classified using Llama 3.3 70B Instruct without task-specific fine-tuning,[25] with temperature set to 0. Separate zero-shot prompts were developed for clinical notes and patient-generated portal messages. Text was classified as positive only when it contained evidence of current depressive symptoms; historical, hypothetical, or negated symptoms and antidepressant use without accompanying symptom evidence were insufficient for a positive classification.

We evaluated three patient-level prediction strategies: (1) clinical notes only, in which a patient was classified as positive if at least one note received a positive prediction; (2) portal messages only, in which a prediction was generated from the concatenated patient-level portal-message document; and (3) combined text sources, in which clinical note and portal message predictions were integrated using a logical OR rule. The combined strategy assessed whether patient-generated communication contributed information beyond clinical documentation.

2.4. *Manual Review of Complementary Portal-Message Cases*

We manually reviewed all additional PHQ-8-positive patients identified by the combined OR strategy but missed by the clinical-note-only model. Portal messages were compared with corresponding clinical notes to identify complementary information. Non-mutually exclusive themes were summarized as counts, and one de-identified representative portal-message excerpt was reported for each theme.

2.5. *Lexicon-Based Measurement of Textual Observability*

To distinguish absent detected evidence from recognition failure, we defined a text-only, label-independent proxy for whether each patient's notes contained explicit depressive-symptom content. The lexicon was built from established symptom constructs and documentation terminology, covering 10 domains: depressed mood, anhedonia, sleep disturbance, fatigue or low energy, appetite or weight change, worthlessness or guilt, impaired concentration, psychomotor change, mental-health help-seeking, and suicidality or self-harm. Terms and regular expressions were developed without access to PHQ-8 labels or model predictions. Mentions were excluded when preceded within a fixed left-context window by negation cues such as "denies," historical cues such as "history of," or hypothetical cues such as "monitor for".[26,27]

For patient i, c_i denoted the number of distinct symptom domains remaining after contextual filtering. Observability at threshold τ was defined as

$$O_i(\tau) = \begin{cases} 1, & \text{if } c_i \geq \tau, \\ 0, & \text{if } c_i < \tau. \end{cases} \tag{1}$$

Patients were lexicon-observable when at least τ domains were detected. The indicator was calculated from text alone and did not use the PHQ-8 label or model prediction.

Observability was lexicon-relative meaning notes without detected symptom terms could still contain clinically meaningful evidence of depression. Observability was assessed only in clinical notes from the 1,725-patient primary cohort.

2.6. *Decomposition of Outcome-Positive Cases*

Observability was used as an explanatory stratification and did not affect cohort inclusion. All 1,725 patients remained in full cohort performance estimates and selective prediction analyses.

Of the P PHQ-8-positive patients, each was assigned to one of four mutually exclusive groups defined jointly by the clinical-note prediction $\widehat{y}_i$ and the observability indicator $O_i(\tau)$:

$$\begin{aligned} TP_{\text{obs}} &= \sum_{i=1}^{N} \mathbb{1}\left(y_i = 1,\ \widehat{y}_i = 1,\ O_i(\tau) = 1\right) & FN_{\text{obs}} &= \sum_{i=1}^{N} \mathbb{1}\left(y_i = 1,\ \widehat{y}_i = 0,\ O_i(\tau) = 1\right) \\ TP_{\text{unobs}} &= \sum_{i=1}^{N} \mathbb{1}\left(y_i = 1,\ \widehat{y}_i = 1,\ O_i(\tau) = 0\right) & FN_{\text{unobs}} &= \sum_{i=1}^{N} \mathbb{1}\left(y_i = 1,\ \widehat{y}_i = 0,\ O_i(\tau) = 0\right). \end{aligned} \tag{2}$$

These groups satisfied:

$$P = TP_{\text{obs}} + FN_{\text{obs}} + TP_{\text{unobs}} + FN_{\text{unobs}}. \tag{3}$$

Here, FN_{obs} represented missed cases despite detected depressive content, whereas FN_{unobs} represented missed cases without detected explicit content. The number of lexicon-unobservable positive cases was

$$U_\tau = TP_{\text{unobs}} + FN_{\text{unobs}} = \sum_{i=1}^{N} \mathbb{1}\left(y_i = 1,\ O_i(\tau) = 0\right). \tag{4}$$

Accordingly, the observability proportion can also be defined as

$$C_\tau = 1 - \frac{U_\tau}{P}. \tag{5}$$

Thus, C_τ represented the proportion of PHQ-8-positive patients whose notes contained at least τ lexicon-detected symptom domains. For methods relying only on the explicit evidence captured by the lexicon, however, it could be interpreted as a lexicon-relative upper bound.

Sensitivity among lexicon-observable positive cases was compared descriptively with full-cohort sensitivity, computed as

$$\text{Sensitivity}_{\text{obs}} = \frac{TP_{\text{obs}}}{TP_{\text{obs}} + FN_{\text{obs}}}, \qquad \text{Sensitivity} = \frac{TP}{TP + FN}. \tag{6}$$

This comparison characterizes performance within the lexicon-defined observable subset but does not estimate sensitivity loss caused by absent textual evidence. We reported TP_{unobs} separately because positive predictions in lexicon-unobservable cases may reflect implicit evidence missed by the lexicon, nonspecific correlates, or unsupported predictions.

2.7. *Sensitivity Analysis of the Observability Definition*

We repeated the positive-case decomposition using observability thresholds of $\tau = 1\ , 2\ , 3$ and 4, corresponding to detection of at least one, two, three or four distinct symptom domains. At each threshold, we recalculated the four observability subgroups, U_τ, C_τ , and $Sensitivity_{obs}$.

The primary definition used $\tau = 1$ because it was the most inclusive. Stricter thresholds were evaluated only to assess sensitivity to the operational definition and were not interpreted as more accurate estimates of semantic observability, because increasing τdeliberately reclassified some patients with detected content as lexicon-unobservable.

2.8. *Content-Conditioned Abstention and Selective Prediction*

We evaluated the content count c_i as an input-side abstention criterion. At threshold t, patients with $c_i \geq$ t were retained, regardless of PHQ-8 label, and patients with $c_i <$ t were excluded only from performance calculations. The abstention rule used neither the PHQ-8 label, the model prediction, nor model correctness. Unlike conventional selective-classification approaches, in which rejection is typically based on model confidence or uncertainty,[28,29] our retention rule was based solely on the amount of phenotype-relevant content available in the input. We nevertheless adopted the standard risk–coverage formalism from selective prediction.[28,29]

The retained set and coverage were defined as

$$R(t) = \{i : c_i \geq t\}, \qquad \text{Coverage}(t) = \frac{|R(t)|}{N}, \tag{7}$$

where N was the size of the primary analytic cohort.

Selective risk, defined as the proportion of incorrect predictions among patients retained at threshold t, was calculated under 0–1 loss as

$$\text{Risk}(t) = 1 - \frac{TP(t) + TN(t)}{|R(t)|}. \tag{8}$$

Within each retained set, we calculated sensitivity, specificity, precision, positive-class F_1 score, and PHQ-8-positive prevalence, p, which represented the proportion of retained patients with PHQ-8 ≥ 5 and defined as

$$p(t) = \frac{\sum_{i \in R(t)} \mathbb{1}\,(y_i = 1)}{|R(t)|}. \tag{9}$$

Because retained-set prevalence could vary across thresholds, selective risks were not interpreted in isolation. At each threshold, we also calculated the risk of the optimal trivial constant classifier, that predicted the majority class for every retained patient, as

$$\mathrm{Risk}_{\mathrm{trivial}}(t) = \min\{p(t),\, 1 - p(t)\}\,. \tag{10}$$

The model's prevalence-adjusted advantage over trivial classification was then defined as

$$\Delta_{\mathrm{risk}}(t) = \mathrm{Risk}_{\mathrm{trivial}}(t) - \mathrm{Risk}(t). \tag{11}$$

Positive values indicated lower error than majority-class prediction in the retained set.

Thresholds were evaluated across c_i values, excluding those retaining fewer than 20 patients. Risk–coverage, F_1–coverage, and risk-advantage–coverage curves assessed performance across thresholds. Abstention was considered reliability-enhancing only if risk decreased while advantage over the threshold-specific trivial classifier was maintained or increased; otherwise, gains in sensitivity or F_1 were interpreted as case enrichment or screening prioritization.

2.9. *Performance Evaluation and Statistical Analysis*

All analyses were conducted at the patient level. Notes-only, portal-only, and combined strategies were evaluated using accuracy, sensitivity, specificity, precision, positive-class F_1, and macro F_1; abstention analyses additionally reported coverage, selective risk, and retained-set prevalence. Ninety-five percent confidence intervals were estimated using 5,000 patient-level bootstrap resamples. Differences between strategies were assessed using 5,000 paired bootstrap resamples that preserved both predictions within each resample, with two-sided bootstrap p-values calculated for differences in sensitivity, specificity, and positive-class F_1.[30]

2.10. *Blinded Semantic Audit of Lexicon-Unobservable Cases*

To assess the semantic validity of lexicon-unobservable cases, two reviewers independently evaluated all 48 true positives and a random sample of 48 lexicon-unobservable false negatives, blinded to PHQ-8 scores, LLM predictions, lexicon output, and subgroup membership. Cases were categorized as containing clinically meaningful, ambiguous or indirect, or no evidence of current depressive symptoms. Disagreements were resolved by consensus, and inter-rater agreement was assessed using Cohen's κ. The audit was designed to assess the unobservable designation among PHQ-8-positive cases; it does not establish the validity of the full four-cell decomposition.

3. Results

Table 1 reports cohort characteristics overall and by PHQ-8 status.

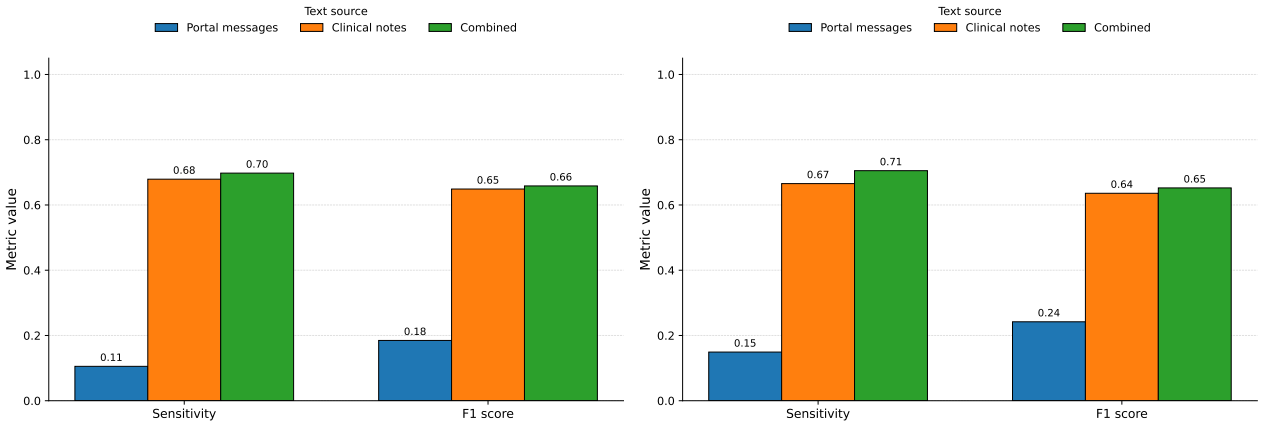
3.1. *Model Performance*

Figure 2 summarizes performance across both portal-message windows. Clinical notes outperformed portal messages alone, while combining sources increased sensitivity and positive-class F_1 but reduced specificity (Table 2). In the primary ±30-day cohort, sensitivity rose from 0.67 to 0.71 and F_1 from 0.64 to 0.65, while specificity fell from 0.58 to 0.55.

Table 1. Characteristics of the primary analytic cohort overall and stratified by PHQ-8 status.

Characteristic	**Overall** ($N = 1{,}725$)	**PHQ-8** < 5 ($N = 867$)	**PHQ-8** ≥ 5 ($N = 858$)
Age, years	60.8 (14.7)	61.0 (14.5)	60.7 (15.0)
Clinical-note count per patient	13.1 (15.2)	12.6 (16.6)	13.6 (13.6)
Portal-message count per patient	26.9 (26.5)	25.2 (25.1)	28.6 (27.8)
Gender			
Female	1,028 (59.6%)	505 (58.2%)	523 (61.0%)
Male	697 (40.4%)	362 (41.8%)	335 (39.0%)
Race			
American Indian or Alaska Native	5 (0.3%)	2 (0.2%)	3 (0.3%)
Asian	73 (4.2%)	28 (3.2%)	45 (5.2%)
Black or African American	609 (35.3%)	326 (37.6%)	283 (33.0%)
Native Hawaiian/Pacific Islander	2 (0.1%)	2 (0.2%)	0 (0.0%)
White	957 (55.5%)	479 (55.2%)	478 (55.7%)
More than one race	9 (0.5%)	6 (0.7%)	3 (0.3%)
Other/unknown	70 (4.1%)	24 (2.8%)	46 (5.4%)
Ethnicity			
Hispanic or Latino	83 (4.8%)	37 (4.3%)	46 (5.4%)
Not Hispanic or Latino	1,585 (91.9%)	805 (92.8%)	780 (90.9%)
Unknown or declined	57 (3.3%)	25 (2.9%)	32 (3.7%)

Values are mean (SD) for continuous variables and n (%) for categorical variables.



(a) Using ±7-Day Portal-Message Window (b) Using ±30-Day Portal-Message Window

Fig. 2. Model Performance Comparison for Clinical Notes, Portal Messages, and Combined Sources

3.2. *Portal Messages Contribution to Additional True-Positive Detection*

Among the 34 additional PHQ-8-positive cases, portal messages most commonly described depressed mood (14 cases), anxiety or emotional distress (13), and fatigue or low energy (10) (Table 3). Compared with notes, which often emphasized physical symptoms, treatment, or disease management, portal messages contained more explicit psychosocial information.

3.3. *Lexicon-Defined Observability and Positive-Case Decomposition*

At the primary threshold of $\tau = 1$, 685 of the 858 PHQ-8-positive patients met the lexicon-defined observability criterion, yielding an observability proportion of $C_{\tau=1} = 0.80$. The remain-

Table 2. Patient-level performance of the clinical-note-only and combined OR strategies across portal-message windows.

±7-Day Portal-Message Window				
Metric	**Clinical notes**	**Combined**	**Difference with 95% CI**	**Bootstrap p-value**
Accuracy	0.63	0.63	0.006 (-0.001 to 0.012)	0.114
Precision	0.62	0.62	0.002 (-0.004 to 0.007)	0.456
Sensitivity	0.68	0.70	0.019 (0.009 to 0.030)	< 0.001
Specificity	0.57	0.56	-0.008 (-0.016 to -0.002)	0.010
F_1 score	0.65	0.66	0.009 (0.003 to 0.017)	0.004
Macro F_1	0.62	0.63	0.005 (-0.001 to 0.011)	0.129
±30-Day Portal-Message Window				
Metric	**Clinical notes**	**Combined**	**Difference with 95% CI**	**Bootstrap p-value**
Accuracy	0.62	0.63	0.005 (-0.003 to 0.014)	0.288
Precision	0.61	0.61	-0.002 (-0.010 to 0.006)	0.599
Sensitivity	0.67	0.71	0.040 (0.027 to 0.053)	< 0.001
Specificity	0.58	0.55	-0.03 (-0.040 to -0.019)	< 0.001
F_1 score	0.64	0.65	0.016 (0.007 to 0.026)	< 0.001
Macro F_1	0.62	0.62	0.004 (-0.005 to 0.012)	0.415

Table 3. Themes identified during review of portal messages from the 34 additional PHQ-8-positive patients.

Type of Information	Count	Short Portal Example (De-identified)
Depressed mood	14	"I'm getting extremely depressed and looking for solutions"
Emotional distress	13	"Unfortunately I've been having panic attacks and heightened anxiety lately... with nothing to combat it with."
Fatigue or low energy	10	"I feel fatigued. I slept all weekend and that is not like myself."
Explicit request for mental health help	9	"I need a therapist to talk to about my cancer and being terminally ill. Can you please put me in touch with someone?"
Sleep disturbance	8	"I haven't slept in 4 nights. This has never happened before."
Appetite / weight changes	7	"I have a loss of appetite and barely eat."
Feelings of hopelessness or worthlessness	3	"I have no plan from you... I passed around like a rag doll."
Loss of interest or pleasure (anhedonia)	1	"He has been anhedonic and mostly sitting around doing nothing... doesn't seem to have the motivation or interest in doing anything."
Suicidal thoughts or self-harm concerns	1	"It's like I'm having so much anxiety that it's crippling me and I just want to kill myself!"
Financial or occupational stress	1	"The bill is unexpected and I can't afford to pay. I have already had choose between eating and paying bills."

ing 173 (20.2%) patients had no lexicon-detected depressive symptom domains. Among the 685 lexicon-observable patients, the LLM correctly identified 523, corresponding to an observable-case sensitivity of 0.76. Of the model's 287 false negatives, 125 were lexicon-unobservable, whereas 162 contained detectable depressive content but were missed by the model. The LLM also produced positive predictions for 48 PHQ-8-positive patients whose notes contained no

lexicon-detected symptom domains, as presented in Table 4.

3.4. *Sensitivity to the Operational Definition of Observability*

In addition to the primary decomposition at $\tau = 1$, we repeated the analysis using thresholds of two, three, and four detected symptom domains (Table 4). As the threshold increased, lexicon-defined observability decreased, whereas sensitivity among lexicon-observable patients increased because the subgroup was restricted to patients with more explicit symptom content. Correctly predicted PHQ-8-positive patients classified as lexicon-unobservable increased from 48 at $\tau = 1$ to 437 at $\tau = 4$. Thus, lower C_τ values at stricter thresholds partly reflected reclassification of patients with fewer detected domains rather than the absence of lexicon-detected depressive content. We retained $\tau = 1$ as the primary threshold because it provided the most inclusive definition of documented symptom content.

Table 4. Decomposition of PHQ-8-positive patients across lexicon-defined observability thresholds.

Threshold	$TP_{\text{obs}}/FN_{\text{obs}}$	$TP_{\text{unobs}}/FN_{\text{unobs}}$	**Unobservable positives**	C_τ	**Sensitivity**$_{\text{obs}}$
≥ 1 domain	523 / 162	48 / 125	173	0.80	0.76
≥ 2 domains	415 / 84	156 / 203	359	0.58	0.83
≥ 3 domains	286 / 42	285 / 245	530	0.38	0.87
≥ 4 domains	134 / 12	437 / 275	712	0.17	0.92

3.5. *Content-Conditioned Abstention and Coverage Tradeoffs*

As shown in Figure 3, by increasing the content threshold, coverage decreased and the retained set became increasingly enriched for PHQ-8-positive patients. Sensitivity and positive-class F_1 improved, whereas specificity declined. However, selective risk did not consistently decrease, and the model's advantage over the threshold-specific trivial classifier narrowed and became slightly negative at the lowest coverage. These findings indicate that content-conditioned abstention improved case detection primarily through class enrichment rather than through greater overall predictive reliability.

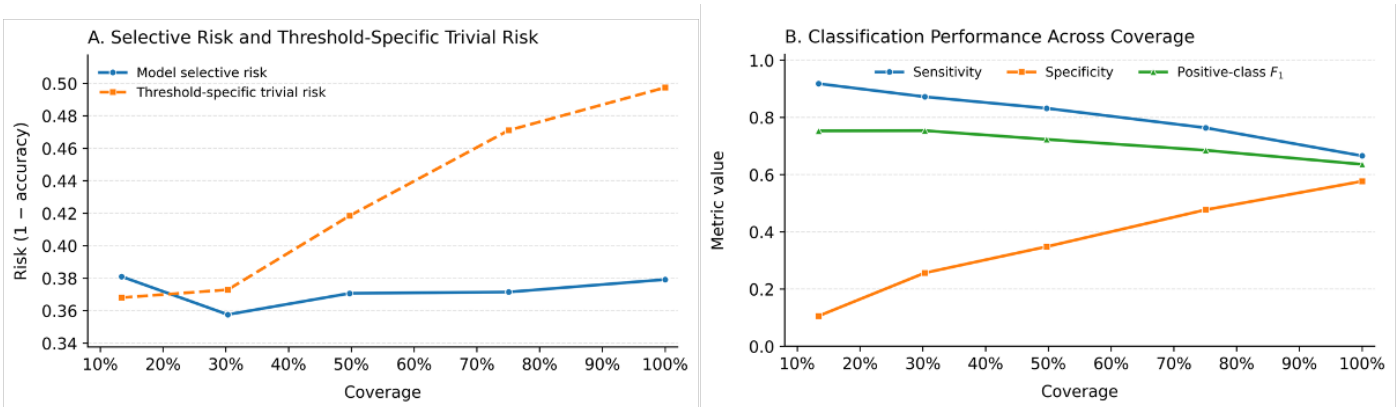


Fig. 3. Content-conditioned abstention across coverage levels. Panel A shows model and trivial-classifier risk; Panel B shows retained-set sensitivity, specificity, and positive-class F_1.

3.6. *Manual Validation of Lexicon-Unobservable Cases*

In the blinded semantic audit, clinically meaningful depressive-symptom evidence was identified in 35 of 48 (72.9%) lexicon-unobservable true positives and 6 of 48 (12.5%) lexicon-unobservable false negatives. In true-positive cases, this evidence often consisted of indirect or contextual expressions not captured by the lexicon. Inter-rater agreement was strong ($\kappa = 0.85$). These findings indicate that lexicon-defined unobservability was instrument-relative: some notes contained meaningful symptom information beyond the lexicon, whereas cases negative by both the lexicon and model more often contained limited evidence.

4. Discussion

This study introduces a lexicon-based observability-aware framework that helps to distinguish failure to recognize available phenotype-relevant evidence from limited availability of that evidence in the analyzed text. We demonstrated the framework in palliative care, a setting in which depressive symptoms may overlap with serious illness and treatment effects, although the empirical estimates remain specific to this population and institution.

Conventional sensitivity treats all missed outcome-positive cases identically. Among the 287 PHQ-8-positive patients missed by the clinical-note model, 125 had no lexicon-detected depressive-symptom domains, whereas 162 contained lexicon-detected content but were still missed. Sensitivity was therefore higher among lexicon-observable positive cases than in the full cohort, 0.76 versus 0.67. This decomposition indicates whether improvement efforts should target model recognition or broader availability of phenotype-relevant evidence.

The blinded audit showed that the lexicon was informative but incomplete. Clinically meaningful depressive-symptom evidence was substantially more common among lexicon-unobservable true positives than false negatives, suggesting that the LLM recognized indirect or contextual evidence beyond explicit symptom terminology, whereas cases classified as negative by both the lexicon and model more often reflected limited documentation. Although the audit estimates remain imprecise, their confidence intervals did not overlap and inter-rater agreement was strong. Accordingly, $C_\tau = 0.80$ should be interpreted as a lexicon-defined observability proportion rather than a universal ceiling on text-based model performance.

Portal messages provided a complementary measurement channel. Adding them to clinical notes increased sensitivity from 0.67 to 0.71 and identified 34 additional PHQ-8-positive patients, often through psychosocial information that was absent from or insufficiently represented in the notes. However, specificity decreased from 0.58 to 0.55, suggesting that the logical OR strategy is better suited to screening and case prioritization than diagnostic classification.

Content-conditioned abstention did not consistently improve predictive reliability. Restricting evaluation to notes with more lexicon-detected symptom domains increased sensitivity and positive-class F1 but reduced specificity, enriched the retained population for PHQ-8-positive patients, and did not lower selective risk. These gains therefore reflected case enrichment rather than improved error control, making content-based restriction more useful for prioritization than reliability-enhancing rejection.

Observability-aware evaluation is applicable when reference standards and model inputs arise from different measurement channels, because conventional metrics alone cannot dis-

tinguish unavailable evidence from model-recognition failure. Studies using patient-reported outcomes, quality-of-life measures, or other independently collected labels should define and validate an input-specific observability instrument, jointly stratify outcome-positive cases by observability and model prediction, and report observable-subset performance alongside full-cohort performance. This decomposition can indicate whether improvements should focus on stronger models, broader or more appropriate text sources, or direct collection of the target outcome. Although the instrument must be adapted to each phenotype, setting, and text source, performance should be interpreted in relation to the information available to the model.

When structured or patient-reported outcomes are unavailable, agreement and disagreement between lexicon and model predictions may help prioritize patients for formal depressive-symptom screening or additional review. Concordant negative signals may indicate limited documented evidence but cannot exclude depressive symptoms. This text-only triage application was not evaluated and requires prospective validation.

Several limitations should be considered. Observability depended on the terms, symptom domains, and contextual filters included in the rule-based lexicon, and the semantic audit included only 48 lexicon-unobservable true positives and 48 lexicon-unobservable false negatives. Requiring portal-message availability may have introduced selection related to digital access, health literacy, healthcare engagement, and socioeconomic factors. Findings from this palliative-care population may not generalize to settings in which depressive symptoms are expressed or documented differently. Finally, this retrospective, single-system study evaluated one instruction-tuned LLM, prompting configuration, phenotype, and PHQ-8 threshold. Although the framework is independent of these choices, external and prospective studies are needed across institutions, models, text sources, populations, and phenotyping tasks.

5. Conclusion

Evaluation of language-based clinical phenotyping against independently collected patient-reported outcomes faces an interpretability problem: conventional sensitivity cannot distinguish cases missed despite relevant textual evidence from cases with little or no detectable evidence. Our observability-aware framework addresses this problem by stratifying outcome-positive patients by model prediction and lexicon-defined evidence availability. Portal messages broadened case detection by contributing patient-generated symptom information insufficiently represented in clinical notes. Content-conditioned abstention enriched the retained population for positive cases but did not reduce overall prediction error, supporting screening prioritization rather than reliability-enhancing rejection. Observability-aware analysis therefore provides a practical framework for interpreting language-based phenotyping when reference standards and model inputs arise from different measurement channels.

References

1. World Health Organization. Depression. Geneva, Switzerland: World Health Organization; 2023. Accessed July 22, 2026. https://www.who.int/news-room/fact-sheets/detail/depression.
2. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry. 2022;9(2):137-150. doi:10.1016/S2215-0366(21)00395-3.
3. Mitchell AJ, Vaze A, Rao S. Clinical diagnosis of depression in primary care: a meta-analysis. Lancet. 2009;374(9690):609-619. doi:10.1016/S0140-6736(09)60879-5.
4. Clement S, Schauman O, Graham T, et al. What is the impact of mental health-related stigma on help-seeking? A systematic review of quantitative and qualitative studies. Psychol Med. 2015;45(1):11-27. doi:10.1017/S0033291714000129.
5. US Preventive Services Task Force. Screening for depression and suicide risk in adults: US Preventive Services Task Force recommendation statement. JAMA. 2023;329(23):2057-2067. doi:10.1001/jama.2023.9297.
6. Bozkurt S, Habibi S, Oltmann C, Pathak R, Kavalieratos D. Identifying depression in palliative care patients' messages using large language models. J Pain Symptom Manage. 2026;71(6):e902–e903.
7. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform. 2008:128-144. doi:10.1055/s-0038-1638592.
8. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012;13(6):395-405. doi:10.1038/nrg3208.
9. Zeng Z, Deng Y, Li X, Naumann T, Luo Y. Natural language processing for EHR-based computational phenotyping. J Biomed Inform. 2019;97:103272. doi:10.1016/j.jbi.2019.103272.
10. Koleck TA, Dreisbach C, Bourne PE, Bakken S. Identifying symptom information in clinical notes using natural language processing. Nurs Res. 2021;70(3):173-183. doi:10.1097/NNR.0000000000000488.
11. Panaite V, Devendorf AR, Finch D, Bouayad L, Luther SL, Schultz SK. The value of extracting clinician-recorded affect for advancing clinical research on depression: proof-of-concept study applying natural language processing to electronic health records. JMIR Form Res. 2022;6(5):e34436. doi:10.2196/34436.
12. Leis A, Mayer MA, Gauna S, et al. Exploring the association of cancer and depression in electronic health records: combining encoded diagnosis and mining free-text clinical notes. JMIR Cancer. 2022;8(3):e39003. doi:10.2196/39003.
13. Alsentzer E, Murphy JR, Boag W, et al. Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. Association for Computational Linguistics; 2019:72-78. doi:10.18653/v1/W19-1909.
14. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. arXiv. Published online April 10, 2019. doi:10.48550/arXiv.1904.05342.
15. Grothman A, Ma WJ, Tickner KG, Martin EA, Southern DA, Quan H. Case identification of depression in inpatient electronic medical records: scoping review. JMIR Med Inform. 2024;12:e49781. doi:10.2196/49781.
16. Kruse CS, Argueta DA, Lopez L, Nair A. Patient and provider attitudes toward the use of patient portals for the management of chronic disease: a systematic review. J Med Internet Res. 2015;17(2):e40. doi:10.2196/jmir.3703.
17. van Buchem MM, de Hond AAH, Fanconi C, et al. Applying natural language processing to patient messages to identify depression concerns in cancer patients. J Am Med Inform Assoc.

2024;31(10):2255-2262. doi:10.1093/jamia/ocae188.
18. Duvall M, North F, Leasure W, Pecina J. Patient portal message characteristics and reported thoughts of self-harm and suicide: a retrospective cohort study. J Telemed Telecare. 2021;27(8):501-508. doi:10.1177/1357633X19887262.
19. Huang M, Khurana A, Mastorakos G, et al. Patient portal messaging for asynchronous virtual care during the COVID-19 pandemic: retrospective analysis. JMIR Hum Factors. 2022;9(2):e35187. doi:10.2196/35187.
20. Mermin-Bunnell K, Zhu Y, Hornback A, et al. Use of natural language processing of patient-initiated electronic health record messages to identify patients with COVID-19 infection. JAMA Netw Open. 2023;6(7):e2322299. doi:10.1001/jamanetworkopen.2023.22299.
21. Lloyd-Williams M. Screening for depression in palliative care patients: a review. Eur J Cancer Care (Engl). 2001;10(1):31-35.
22. Noorani NH, Montagnini M. Recognizing depression in palliative care patients. J Palliat Med. 2007;10(2):458-464. doi:10.1089/jpm.2006.0099.
23. Kroenke K, Strine TW, Spitzer RL, Williams JBW, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord. 2009;114(1-3):163-173. doi:10.1016/j.jad.2008.06.026.
24. Kroenke K, Spitzer RL, Williams JBW, Lowe B. The Patient Health Questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. Gen Hosp Psychiatry. 2010;32(4):345-359. doi:10.1016/j.genhosppsych.2010.03.006.
25. Grattafiori A, Dubey A, Jauhri A, et al. The Llama 3 herd of models. arXiv. Published online July 31, 2024. doi:10.48550/arXiv.2407.21783.
26. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform. 2001;34(5):301-310. doi:10.1006/jbin.2001.1029.
27. Harkema H, Dowling JN, Thornblade T, Chapman WW. 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.
28. El-Yaniv R, Wiener Y. On the foundations of noise-free selective classification. J Mach Learn Res. 2010;11:1605-1641.
29. Geifman Y, El-Yaniv R. Selective classification for deep neural networks. In: Advances in Neural Information Processing Systems 30. Curran Associates; 2017:4878-4887.
30. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. New York, NY: Chapman & Hall/CRC; 1993.