Before the Labels: How Dataset Construction Shapes Suicidality Detection in Clinical Text

Priyanshi Garg[†*], **Ishita Rao**[†*], **Jieqiong Ding**[‡*], **Amandalynne Paullada**[†]
[†]Department of Linguistics, University of Washington, Seattle, WA, USA
[‡]Human Centered Design & Engineering, University of Washington, Seattle, WA, USA
{pgarg2, ishirao, jding07, paullada}@uw.edu

Abstract

Clinical NLP increasingly relies on electronic health record (EHR) data to detect suicidal behaviors, treating clinical documentation as more reliable ground truth than social media. We argue that this framing obscures how EHR-based suicidality datasets encode a particular operationalization of suicidality, shaped by who authors the data, how episodes are bounded, and how ambiguity is resolved. We ground this argument in a case study of the ScAN dataset (Rawat et al., 2022), built over MIMIC-III clinical notes. We show how governance constraints, ICD-based cohort selection, single-annotator labeling, and hospital-stay-level aggregation produce labels that reflect clinician-documented judgments, treat suicidality as a bounded episode, and assume that intent can be reliably inferred from documentation. A linguistic analysis demonstrates that identical labels subsume heterogeneous clinical framings differing in temporality, negation, and uncertainty. We argue that clinical NLP should examine the assumptions embedded in suicidality datasets before interpreting their labels as ground truth.

1 Introduction

Suicide is a leading cause of death worldwide, and detecting suicidal behavior from text is a high-stakes task in clinical NLP. Recent work has increasingly turned to electronic health records (EHRs) as a data source, motivated by well-documented concerns with social media data around consent, self-report reliability, and generalization (Chancellor et al., 2019; Ernala et al., 2019; Harrigian et al., 2021). EHRs offer structured clinical documentation, multi-source corroboration, and longitudinal patient histories.

However, systematic reviews have found that predictive models for suicide perform only marginally better than chance across a range of clinical settings (Belsher et al., 2019; Franklin et al., 2017; Large et al., 2016). This poor predictive performance warrants scrutiny of, among other factors, the training data and the assumptions encoded within the labels. EHRs should not be treated as direct access to suicidality itself. Clinical notes are documentation-mediated records: they synthesize patient report, clinician observation, prior records, collateral information, institutional requirements, and risk-management practices. This mediation is especially important for suicidality, where disclosure and documentation depend on the clinical interaction in which risk is assessed. Prior work shows that healthcare professionals vary in how they ask about suicide risk, including question designs that may invite denial or reduce elaboration (McCabe et al., 2017). Mental health professionals also vary in how consistently they assess suicidal thoughts and behaviors, with assessment practices shaped in part by comfort working with suicidal individuals (Roush et al., 2018). Thus, the absence, presence, or wording of suicidality evidence in EHR text reflects both clinical phenomena and the conditions under which those phenomena become documented.

Prior work has also examined ethical tensions in mental health inference, demographic bias in diagnosis and clinical documentation, and documentation quality in clinical NLP. For example, racial disparities in psychiatric diagnosis can shape what enters the clinical record (Schwartz, 2014), and stigmatizing language in EHR notes has been shown to vary across patient groups (Sun et al., 2022). These concerns are often treated separately. We propose *operationalization* as a unifying lens: dataset construction decisions collectively determine what counts as evidence, how it is categorized, what distinctions are preserved, and what forms of uncertainty are collapsed.

Our position is that EHR-based suicidality

[*]Equal contribution.

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026), pages 119–127
July 4, 2026 ©2026 Association for Computational Linguistics

datasets encode a specific operationalization of suicidality:

> *Documentation-mediated*, reflecting suicidality as represented in clinician-authored records rather than as an unmediated patient state; *episodic*, bounding suicidality to discrete hospital stays rather than modeling it as a longitudinal process; and *intent-resolved*, imposing binary distinctions between suicidal and non-suicidal self-harm even when the clinical text is ambiguous.

This is not an argument that clinical documentation is unreliable, nor that patient self-report is a bias-free alternative. Suicidality assessment is clinically difficult, relational, and uncertain. Rather, our claim is that when clinical documentation is converted into NLP labels, those labels should be interpreted as structured operationalizations of documented clinical evidence, not as neutral ground truth.

We ground this argument in a case study of the ScAN dataset (Rawat et al., 2022), examining how access constraints, cohort design, and annotation decisions shape the labels and how those choices appear in the linguistic structure of annotated spans.

Our contributions are:

- A position characterizing EHR-based suicidality datasets as encoding a documentation-mediated, episodic, and intent-resolved operationalization of suicidality
- A case study of ScAN tracing how governance, data sourcing, and annotation design give rise to this operationalization
- Empirical evidence that these construction choices are associated with systematic variation in linguistic framing and labeling patterns

2 ScAN as a Case Study

ScAN (Rawat et al., 2022) is an annotation layer for suicide attempts (SA) and suicidal ideation (SI) built over a subset of MIMIC-III, a de-identified clinical electronic health record (EHR) dataset from Beth Israel Deaconess Medical Center in Boston (Johnson et al., 2016). It contains 12,759 clinical notes across 697 hospital stays from 669 patients, with 19,690 span-level annotations. SA labels distinguish between positive, negative, unsure, and neutral cases based on inferred intent; SI labels capture presence or absence of ideation. We use ScAN as a case study to examine how EHR-based suicidality labels are shaped by the documentation, cohort, and annotation choices through which clinical text becomes machine-learning data.

Clinical notes often synthesize patient report, clinician observation, collateral information, prior records, institutional requirements, and risk-management practices. We use *documentation-mediated* to mean that the resulting NLP labels are labels over documented clinical evidence, not direct measurements of suicidality itself. Below, we trace three dimensions of this operationalization: documentation-mediated perspective, episodic temporal framing, and intent resolution.

2.1 Documentation-Mediated: Governance, Access, and Perspective

ScAN inherits the governance model of MIMIC-III. Although annotations and code are publicly released, the underlying clinical text requires credentialing through PhysioNet, completion of human subjects training, and agreement to a Data Use Agreement. Reproducibility is therefore conditional on institutional resources and compliance capacity.

The source text is clinician-authored documentation. This does not mean that patient perspectives are absent from the notes: patient reports, statements, and histories may be represented within clinician-authored text. However, direct patient-authored self-report, intake questionnaires, and patient-facing communications are outside the annotation scope. The dataset therefore represents suicidality as it appears in clinical documentation, rather than as it might appear across all possible sources of patient expression.

The annotation process further mediates the relationship between patient experience and dataset labels. Only notes from 24 pre-selected clinical section types were retained for annotation, based on expected relevance to suicidality. Annotation was performed by a single trained annotator supervised by a physician, who independently reviewed 330 of 12,759 notes (2.6%). Disagreements were resolved by updating labels to match the physician review. Thus, the labels reflect the ScAN annotation protocol as applied to clinician-authored documentation under partial physician review. They should be interpreted as structured judgments about documented evidence, not as neutral or exhaustive records of suicidality.

De-identification procedures, including date shifting and age masking for patients over 89, further mediate the relationship between the data and the clinical events they represent (Johnson et al., 2016).

2.2 Episodic: Cohort Selection and Temporal Flattening

ScAN's cohort is seeded using ICD codes associated with suicide and overdose, restricting the dataset to hospital stays where suicidality or self-harm was already suspected, documented, or coded. ICD codes capture only a fraction of suicidality-related events, approximately 3% of SI and 19% of SA by some estimates (Anderson et al., 2015), so patients whose suicidal behavior went uncoded are outside the dataset.

Labeling is performed at the hospital-stay level. This design is practical for constructing an EHR-based dataset, but it also frames suicidality as a bounded hospital episode. Temporal distinctions, such as whether an event occurred during the current admission, in a recent prior encounter, or years earlier, are not directly encoded in the label schema even when such distinctions appear in the clinical text. This matters because suicidality is often longitudinal and recurrent rather than confined to a single encounter. Many individuals who die by suicide sought care well before their deaths (Kessler et al., 2020), and ideation may recur across encounters.

2.3 Intent-Resolved: Annotation Design and Category Collapse

ScAN's annotation scheme uses CDC-based definitions that distinguish suicidal from non-suicidal self-harm, requiring intent to be inferred from clinical documentation. Non-suicidal self-injury (NSSI) is excluded by design.

Cases where intent is unclear are labeled "unsure," but this category is collapsed with "negative" in the downstream modeling pipeline used by the companion ScANER system. This merges qualitatively different situations: documentation that explicitly negates suicidal intent, documentation where intent is not specified, and documentation where the available evidence is genuinely ambiguous. As we show in Section 2.5, these categories have distinct linguistic profiles. The issue is not that the original annotation decision was unreasonable, but that downstream label collapse can erase clinically meaningful uncertainty.

2.4 Dataset Composition

ScAN's single-site origin produces a demographically narrow dataset: 70% White, 92% English-speaking, and concentrated in the 25–54 age range, which includes approximately 70% of patients. No non-White racial or ethnic group exceeds 22 admissions, making subgroup analysis infeasible for most categories. The structured demographic data also contains some ambiguities; full details and tables are in Appendix E. We therefore treat demographic composition primarily as a description of dataset scope, rather than as the basis for subgroup claims.

2.5 Linguistic Framing Analysis

If labels flatten temporal, epistemic, and intentional distinctions present in the clinical text, then spans sharing the same label should exhibit heterogeneous linguistic profiles. We test this expectation through a focused framing analysis.

We analyze 15,585 annotated spans after excluding those with missing metadata. For each span, we identify negation, historical reference, and uncertainty using lexical indicators and MedSpaCy's ConText algorithm (Harkema et al., 2009). We classify a span as *unmodified* when both methods agree on the absence of all three modifier types. This is a descriptive linguistic category, not a clinical judgment: "unmodified" means that our detectors did not find negation, historical reference, or uncertainty governing the span. It does not imply clinically active suicidal intent. Details and inter-method agreement rates are in Appendices A and B.

Within-label heterogeneity. Spans sharing the same label exhibit substantial variation in framing (Table 1). Two-thirds of SA spans are unmodified, but only about half of present-SI spans are. Of particular note, 27.8% of present-SI spans contain historical markers, indicating that ideation labeled as present may still be documented in relation to prior psychiatric history rather than as a purely present-tense event.

Ambiguous categories and the cost of collapse. The categories that encode uncertainty show the clearest effect (Table 2). "Unsure" spans are frequently unmodified but contain elevated uncertainty markers (17.1%), reflecting epistemic

Table 1: Framing profiles by span type. "Unmodified" means no detected negation, historical reference, or uncertainty marker.

Span type	n	**% Unmod.**	**% Hist.**	**% Neg.**
All SA	14,048	66.8	13.2	5.8
Present-SI	897	52.2	27.8	7.0
Absent-SI	554	13.5	16.1	65.2

Table 2: Framing profiles of selected SA categories.

Category	n	**% Unmod.**	**% Hist.**	**% Neg.**	**% Unc.**
Intentional (X71–X83)	1,271	83.3	12.1	1.1	1.7
Unsure	2,237	61.6	19.3	6.8	17.1
Unspecified (T14.91)	770	19.0	67.8	3.4	5.1

ambiguity in the documentation. Unspecified-intent spans (T14.91) are predominantly historical (67.8%), indicating that lack of intent specification often co-occurs with temporal displacement. When these categories are merged with "negative" labels in downstream pipelines, epistemic uncertainty and temporal displacement are collapsed into a single class.

Variation across documentation contexts. Framing varies substantially by clinical note section. Discharge summaries are 83.6% unmodified while history of present illness (HPI) sections are only 49.9% unmodified (Appendix D). The same label therefore carries different linguistic signals depending on the documentation context in which it appears.

Summary. Across all three framing dimensions, labels that appear uniform encode meaningfully different clinical situations. Temporal context may be flattened, uncertainty may be resolved or collapsed, and intent may be treated as more determinate than the surrounding documentation supports. These findings do not show that the labels are wrong. Rather, they show that EHR-based suicidality labels are operationalizations of documented clinical evidence, and that downstream NLP systems should be evaluated with those operationalizations in view.

3 Discussion

The issues we identify are not unique to ScAN. Any EHR-based suicidality dataset will reflect some set of choices about whose perspective is encoded, how episodes are bounded, and how ambiguity is resolved. The question is whether these choices are examined. Below, we trace the downstream consequences of the operationalization we have described and outline steps towards making these choices more visible.

3.1 Possible Downstream Effects

If training labels subsume heterogeneous linguistic framings, classifiers trained on those labels would be expected to exhibit systematic and predictable errors. The historical-marker rate among present-SI spans (27.8%) predicts that models would learn to associate retrospective language with active ideation, potentially overflagging historical mentions in new clinical text. The gap in unmodified rates between discharge summaries and HPI sections (83.6% vs. 49.9%) predicts that classifier performance would vary substantially across note section types, even within a single hospital stay. The collapse of unsure with negative SA labels predicts that models would be less calibrated on cases involving genuine epistemic uncertainty—precisely the cases where clinical decision support is most needed. We do not test these predictions here, but each is falsifiable through standard error analysis on classifiers trained with and without label distinctions preserved.

3.2 Alternative Design Choices and Recommendations

Each dimension of the operationalization we describe suggests both an alternative dataset design and a corresponding evaluation practice that can surface hidden assumptions even when the dataset itself cannot be rebuilt.

For temporal framing, labeling across encounters rather than within a single hospital stay would allow models to represent suicidality as a recurring process. Where cross-encounter labeling is not feasible, reporting model performance separately by clinical note section type would reveal whether classifiers generalize across the variation in linguistic framing we observed—for example, between discharge summaries and HPI notes.

For intent resolution, retaining "unsure" as a distinct modeling category rather than merging it with "negative" would preserve clinically meaningful uncertainty in the label space. This matters not only for model calibration but also for clinicians who may encounter decision-support tools trained on these labels: a system that treats genuinely ambiguous cases as resolved negatives risks understating uncertainty in precisely the situations where clinical judgment is most needed. At minimum, pipelines that collapse ambiguous categories should evaluate performance on each constituent category before and after the merger.

For the documentation-mediated dimension, two related limitations apply: what text is available for annotation, and how consistently that text is interpreted. Clinical notes are authored by clinicians and reflect their documentation practices, even when they incorporate patient-reported information. Incorporating patient-authored sources—such as self-report instruments, intake questionnaires, or patient-facing communications—alongside clinician-authored notes would allow datasets to represent suicidality across multiple perspectives rather than through clinical documentation alone. Independently, multi-annotator labeling with reported inter-annotator agreement would make transparent how consistently annotators interpret the clinical text that is available.

Each of these alternatives introduces practical trade-offs: longitudinal labeling requires cross-encounter linking infrastructure, preserving ambiguous categories adds complexity to model training and evaluation, and multi-annotator schemes increase annotation effort. These costs are real, but they are the costs of making the operationalization an explicit design choice rather than an invisible default. More broadly, dataset documentation should state which operationalization choices a pipeline inherits—how cohorts are seeded, at what level labels are aggregated, whether ambiguous categories are collapsed, and how many annotators contributed—so that downstream users, including clinicians evaluating decision-support systems, can assess what the labels represent before acting on model outputs.

4 Conclusion

We have argued that EHR-based suicidality datasets should be read as operationalizations shaped by governance, cohort design, and annotation practice, rather than as neutral ground truth. Using ScAN as a case study, we showed that these choices produce labels that flatten temporality, resolve ambiguity, and encode clinician-documented judgments. Our linguistic analysis confirmed that identical labels subsume spans with different temporal, negation, and uncertainty profiles.

None of these observations are criticisms of ScAN's creators, who made reasonable decisions under real constraints. The point is that these decisions constitute an operationalization, and that operationalization should be visible to downstream users. When labels are treated as ground truth without reference to the assumptions that produced them, the gap between what the dataset represents and what models are trained to predict becomes difficult to assess. Making these assumptions visible is a necessary first step toward clinical NLP systems that are transparent about what they predict. Toward that end, we have outlined testable predictions about downstream model behavior and practical recommendations that can help move this conversation forward.

5 Limitations

Our case study examines one dataset from one academic medical center. The specific distributions we report (e.g., the rate of historical markers in present-SI spans) may differ in datasets with other governance models, coding systems, or annotation protocols. Our argument is structural: any pipeline that seeds cohorts from billing codes, labels at the hospital-stay level, and resolves intent from clinical text will encode analogous assumptions. We do not train or evaluate classifiers on ScAN, so we cannot say whether the within-label heterogeneity we identify produces systematic model errors; establishing that link is a natural next step.

Our linguistic analysis covers negation, temporality, and uncertainty using lexical indicators and MedSpaCy's ConText algorithm, both of which operate at the span level without syntactic parsing or coreference resolution. This means we do not capture attribution, cross-note temporal reasoning, or scope ambiguities where a modifier may or may not govern the suicidality mention. Because our detection methods are conservative, the heterogeneity we report is likely a lower bound.

References

Heather D Anderson, Wilson D Pace, Elias Brandt, Rodney D Nielsen, Richard R Allen, Anne M Libby, David R West, and Robert J Valuck. 2015. Monitoring suicidal patients in primary care using electronic health records. *The Journal of the American Board of Family Medicine*, 28(1):65–71.

Bradley E Belsher, Derek J Smolenski, Larry D Pruitt, Nigel E Bush, Erin H Beech, Don E Workman, Rebecca L Morgan, Daniel P Evatt, Jennifer Tucker, and Nancy A Skopp. 2019. Prediction models for suicide attempts and deaths: a systematic review and simulation. *JAMA psychiatry*, 76(6):642–651.

Stevie Chancellor, Michael L Birnbaum, Eric D Caine, Vincent MB Silenzio, and Munmun De Choudhury. 2019. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 79–88.

Sindhu Kiranmai Ernala, Michael L Birnbaum, Kristin A Candan, Asra F Rizvi, William A Sterling, John M Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–16.

Joseph C. Franklin, Jessica D. Ribeiro, Kathryn R. Fox, Kate H. Bentley, Evan M. Kleiman, Xieyining Huang, Katherine M. Musacchio, Adam C. Jaroszewski, Bernard P. Chang, and Matthew K. Nock. 2017. Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological Bulletin*, 143(2):187–232.

Henk Harkema, John N Dowling, Tyler Thornblade, and Wendy W Chapman. 2009. Context: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of biomedical informatics*, 42(5):839–851.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2021. On the state of social media data for mental health research. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 15–24.

Alistair Johnson, Tom Pollard, and Roger Mark. 2016. Mimic-iii clinical database (version 1.4). *PhysioNet*, 10(C2XW26):2.

Ronald C Kessler, Robert M Bossarte, Alex Luedtke, Alan M Zaslavsky, and Jose R Zubizarreta. 2020. Suicide prediction models: a critical review of recent research with recommendations for the way forward. *Molecular psychiatry*, 25(1):168–179.

Matthew Large, Muthusamy Kaneson, Nicholas Myles, Hannah Myles, Pramudie Gunaratne, and Christopher Ryan. 2016. Meta-analysis of longitudinal cohort studies of suicide risk assessment among psychiatric patients: Heterogeneity in results and lack of improvement over time. *PLOS ONE*, 11(6):e0156322.

Rose McCabe, Imren Sterno, Stefan Priebe, Rebecca Barnes, and Richard Byng. 2017. How do healthcare professionals interview patients to assess suicide risk? *BMC Psychiatry*, 17(1):122.

Bhanu Pratap Singh Rawat, Samuel Kovaly, Hong Yu, and Wilfred Pigeon. 2022. Scan: suicide attempt and ideation events dataset. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1029–1040.

Jared F. Roush, Sarah L. Brown, Danielle R. Jahn, Sean M. Mitchell, Nathanael J. Taylor, Paul Quinnett, and Richard Ries. 2018. Mental Health Professionals' Suicide Risk Assessment and Management Practices: The Impact of Fear of Suicide-Related Outcomes and Comfort Working With Suicidal Individuals. *Crisis*, 39(1):55–64.

Robert C Schwartz. 2014. Racial disparities in psychotic disorder diagnosis: A review of empirical literature. *World Journal of Psychiatry*, 4(4):133.

Michael Sun, Tomasz Oliwa, Monica E Peek, and Elizabeth L Tung. 2022. Negative patient descriptors: Documenting racial bias in the electronic health record: Study examines racial bias in the patient descriptors used in the electronic health record. *Health Affairs*, 41(2):203–211.

A Framing Analysis: Definitions

Lexical keyword lists. The following case-insensitive terms are matched within each three-sentence context window:

- **Negation:** *no, not, denies, denied, without, never, none*
- **Historical:** *history of, h/o, hx of, previous, prior, past, ago, years ago, months ago*
- **Uncertainty:** *may, might, possibly, unclear, uncertain, equivocal, question of, rule out, r/o*

Lexical matching does not resolve scope: a keyword may modify a concept unrelated to suicidality. We therefore run MedSpaCy's ConText algorithm (Harkema et al., 2009) in parallel with default `ConTextComponent` settings, retaining only modifiers whose scope overlaps with the annotated span.

Composite measure. A span is framed as *unmodified* when both methods agree on the absence of negation, historical, and uncertainty modifiers. This conservative definition means some modified spans may carry modifiers that do not pertain to the suicidality judgment.

B Signal Agreement Rates

Table 3 reports span-level agreement between the lexical and ConText detectors ($n = 15{,}585$).

Table 3: Base rates and agreement between lexical and ConText detectors. The uncertainty divergence ($9.4\times$) reflects how often terms like *may* modify non-suicidality concepts.

Signal	Lex	Ctx	Agree	Ratio
Negation	6.3%	7.6%	92.1%	0.8×
Historical	19.6%	8.4%	86.6%	2.3×
Uncertainty	9.4%	1.0%	91.2%	9.4×

C Category-Level Framing

Tables 4 and 5 extend the summaries in Section 2.5 to all annotation categories. Lex-H / Ctx-H = lexical / ConText historical; Lex-N / Ctx-N = negation; Lex-U = lexical uncertainty (ConText uncertainty $\leq 1\%$ everywhere, omitted).

D Section-Level Framing

Table 6 reports framing rates by clinical note section. Sections with $n < 40$ are included for completeness. % SI = share of SI spans in section.

E Demographic Analysis

Analyses use 486 admissions with structured data from MIMIC-III's `ADMISSIONS` and `PATIENTS` tables joined to ScAN annotations. We report Pearson's χ^2 tests with uncorrected p-values; given small subgroups and the exploratory scope, we do not apply multiple-comparison corrections.

Subgroup sizes. Insurance: Private 181, Medicaid 122, Medicare 103, Government 56, Self Pay 22, missing 2. Gender: Female 257, Male 227, missing 2. Race/ethnicity (collapsed): White 339, Unknown/Not Specified[1] 76, Black/African American 22, Hispanic/Latino 18, Other 22, Asian 9. No non-White group exceeds 22 admissions; race/ethnicity comparisons are not reported.

Tables 8, 9, 10, 11, 12, 13 and 14 show the respective distributions of gender, age, ethnicity, language spoken, insurance status, marital status, and religion of patients in the dataset. Additionally, the distribution of living vs. deceased patients is reported in Table 15.

Ambiguities. Ambiguities were noted in the gender, language, and ethnicity data. ScAN included one male-to-female transgender individual. While the MIMIC-III documentation defines gender as "the genotypical sex of the patient" (Johnson et al., 2016), the transgender patient's gender was listed as Female, which is not their genotypical sex. Further, MIMIC-III, and by extension ScAN, uses non-standard 4 letter language codes. While some codes are easy to infer, such as *SPAN* for Spanish, other codes such as *CAPE* and *PTUN* are not intuitive and have no documented definition. Finally, ethnicity data is provided with varying levels of specificity. For example, an individual could be listed as *White*, *White – Brazilian*, *White – Russian*, or *White – Other European*. However, we acknowledge that it may not always be possible to obtain a patient's specific ethnicity.

[1] Includes UNKNOWN/NOT SPECIFIED (n=54), UNABLE TO OBTAIN (n=17), PATIENT DECLINED (n=3), missing (n=2).

Table 4: Framing profiles for SA categories (n = 14,048), ordered by % Unmodified.

Cat.	Description	n	% Unmod.	Lex-H	Ctx-H	Lex-N	Ctx-N	Lex-U
T71	Asphyxiation	440	86.6	9.6	2.1	0.5	0.7	3.6
X71–X83	Intentional self-harm	1,271	83.3	12.1	5.0	1.1	1.6	1.7
T51–T65	Toxic subst. (non-pharma)	531	74.0	16.2	6.8	2.8	5.1	4.7
T36–T50	Pharma. poisoning	8,635	69.5	16.3	6.5	2.5	4.7	10.1
Unsure	Annotator unsure	2,237	61.6	19.3	2.3	6.8	6.2	17.1
T14.91	Unspecified intent	770	19.0	67.8	50.4	3.4	6.6	5.1
N/A	No ICD code	164	17.1	37.8	23.8	50.0	56.1	9.2

Table 5: Framing profiles for SI spans (n = 1,537) by status label.

Status	n	% Unmod.	Lex-H	Ctx-H	Lex-N	Ctx-N	Lex-U
Present	897	52.2	27.8	12.5	7.0	11.8	7.1
Absent	554	13.5	16.1	9.2	65.2	57.2	3.1
N/A	86	23.3	14.0	8.1	52.3	37.2	11.6

Table 6: Framing by clinical note section, ordered by % Unmodified.

Section	n	% Unmod.	Lex-H	Ctx-H	Lex-N	Ctx-N	% SI
Discharge summary	3,358	83.6	9.3	2.4	1.5	1.7	4.0
Assessment / plan	4,262	66.1	16.5	6.0	6.3	7.7	12.4
Medical history	743	61.5	28.1	15.3	3.4	4.7	10.5
Other	31	61.3	32.3	6.5	3.2	3.2	6.5
Chief complaint	2,567	56.7	23.3	9.5	6.3	12.5	7.8
Impression	840	55.4	21.6	11.6	8.6	6.8	10.2
Social history	2,270	51.7	23.6	12.3	9.1	11.2	12.5
HPI	1,474	49.9	33.4	16.2	12.8	9.2	14.9
Psychiatric history	40	45.0	35.0	10.0	12.5	2.5	15.0

Table 7: Demographic association tests at or near significance. All other comparisons returned $p > 0.10$.

Comparison	Outcome	p	
Medicaid vs. Private 38/122 (31.1%) vs. 33/181 (18.2%)	≥1 unsure SA	0.014	*
Female vs. Male 221/257 (86.0%) vs. 181/227 (79.7%)	≥1 historical	0.087	†
Female vs. Male 159/257 (61.9%) vs. 120/227 (52.9%)	≥1 uncertainty	0.056	†

* $p < 0.05$; † $p < 0.10$ (marginal)

Gender	%
Female	50.9
Male	49.1

Table 8: Gender distribution of ScAN patients.

Age Group	%
18–24	13.5
25–34	21.6
35–44	22.8
45–54	25.5
55–64	9.7
65–74	5.1
75+	1.9

Table 9: Age distribution of ScAN patients.

Ethnicity	%
White	70.0
Unknown / Other / Declined	16.4
Black / African American	5.9
Hispanic or Latino	3.6
Asian	1.7
European	0.7
Black / Cape Verdean	0.6
Multi-race	0.4
South American	0.3
American Indian / Alaska Native	0.1
Native Hawaiian / Pacific Islander	0.1

Table 10: Ethnicity distribution of ScAN patients. 'European' includes the categories 'Portugese', 'White – Russian', and 'White – Other European'. 'Asian' includes the categories 'Asian', 'Asian – Japanese', 'Asian – Chinese', 'Asian – Vietnamese'. South American includes the categories 'South American' and 'White – Brazilian'. 'Hispanic or Latino' includes the categories 'Hispanic or Latino', 'Hispanic/Latino – Puerto Rican', and 'Hispanic/Latino – Dominican'.

Language	**%**
English (ENGL)	92.3
Spanish (SPAN)	2.4
Unknown (PTUN)	1.8
Russian (RUSS)	1.0
Cantonese (CANT)	0.4
Portugese (PORT)	0.4
Unknown (CAPE)	0.4
Other	1.3

Table 11: Language distribution of ScAN patients. 'Other' includes Japanese (JAPA), Arabic (ARAB), Mandarin (MAND), American Sign Language (AMER), French (FREN), and Haitian (HAIT). Each of these languages have 0.2% speakers in ScAN.

Insurance Type	**%**
Private	34.9
Medicaid	26.2
Medicare	24.1
Government	10.8
Self-pay	4.0

Table 12: Insurance distribution of ScAN patients.

Marital Status	**%**
Single	62.1
Married	25.3
Divorced	7.6
Widowed	2.8
Separated	1.2
Unknown	1.0

Table 13: Marital status distribution of ScAN patients.

Religion	**%**
Catholic	57.1
Other	13.7
Protestant / Quaker	12.7
Jewish	11.4
Christian – Other	3.8
Buddhist	1.3

Table 14: Religion distribution of ScAN patients among the 45.4% who elected to provide their information. The remaining 54.6% either did not specify their religion or this information was unobtainable.

Mortality	**%**
Alive	84.9
Deceased (post-discharge)	10.2
Deceased (during visit)	4.9

Table 15: Mortality distribution of ScAN patients.