# Open Source Clinical Safety Signal Detection for Behavioral Health: Architecture, Taxonomy, and Compliance Framework

**Author:** Tanmaya Kumar **Affiliation:** Behavioral Health Open Source (bh-healthcare.org) **Date:** April 2026 **Version:** 1.0 **Project:** bh-sentinel **License:** Apache 2.0

---

## Table of Contents

1. The Problem
    - 1.1 The Behavioral Health Technology Gap
    - 1.2 What Exists Today
    - 1.3 What Does Not Exist
    - 1.4 Why Social Media Models Fail on Clinical Text
2. Design Philosophy
    - 2.1 What bh-sentinel Does
    - 2.2 What bh-sentinel Does Not Do
    - 2.3 Hard Constraints
3. Clinical Flag Taxonomy
    - 3.1 Domains
    - 3.2 Severity Levels
    - 3.3 Self-Harm and Suicidal Ideation
    - 3.4 Harm to Others
    - 3.5 Medication Non-Adherence
    - 3.6 Substance Use
    - 3.7 Clinical Deterioration
    - 3.8 Protective Factors

4. Detection Architecture
    - 4.1 Pipeline Overview
    - 4.2 Layer 1: Deterministic Pattern Matching
    - 4.3 Layer 2: Transformer Classification
    - 4.4 Layer 3: Emotion and Sentiment Analysis
    - 4.5 Layer 4: Configurable Rules Engine
5. Escalation Rules
6. Compound Risk Rules
7. Emotion Lexicon
8. FDA Non-Device Clinical Decision Support Compliance
9. Deployment Architecture
10. What bh-sentinel Is Not
11. Roadmap
12. References

---

## Abstract

Behavioral health organizations serve some of the most vulnerable patient populations in the US healthcare system, yet the technology infrastructure supporting clinical safety detection in this sector remains fragmented, proprietary, and largely inadequate. Fewer than 20% of behavioral health providers use measurement-based care, and no production-ready open source tool exists for multi-domain clinical safety signal detection in behavioral health text.

bh-sentinel is an open source, stateless, platform-agnostic text analysis system designed to detect clinical safety flags in unstructured behavioral health documentation. It operates as a pure “text in, flags out” processor: clinical text enters, structured risk flags with severity scores, evidence spans, confidence levels, and emotion signals exit. No text is stored. No PHI is persisted. The system is designed for deployment inside a healthcare organization’s own AWS environment with zero data egress.

The system detects 40 clinical safety flags across 6 domains (self-harm and suicidal ideation, harm to others, medication non-adherence, substance use, clinical deterioration, and protective factors) using a 4-layer parallel detection pipeline combining deterministic pattern matching, transformer-based classification, emotion and sentiment analysis, and a configurable compound risk rules engine.

bh-sentinel is a clinical decision support tool. It detects signals and prompts clinician review. It does not diagnose, prescribe, or take autonomous clinical action. Its architecture is designed to satisfy the four criteria for Non-Device Clinical Decision Support under Section 520(o)(1)(E) of the FD&C Act.

This paper describes the clinical flag taxonomy, detection architecture, compound risk logic, and FDA compliance framework in full technical detail.

---

## 1. The Problem

### 1.1 The Behavioral Health Technology Gap

Behavioral health care in the United States operates under conditions that would be unrecognizable in most other healthcare sectors. Clinical documentation is fragmented across EMRs that were designed for primary care, not group therapy. Intake processes run on paper forms, PDF uploads, and manual data entry. Clinical safety signals - suicidal ideation, medication non-adherence, substance use escalation, psychotic symptoms - are identified through clinician judgment alone, with no systematic detection layer between the patient’s words and the provider’s next-session review.

This is not a criticism of clinicians. It is a description of the infrastructure they are forced to work within.

The digital mental health market is projected to reach $82 billion by 2032, yet the foundational tooling for clinical safety detection remains either proprietary, narrowly scoped, or trained on data that does not reflect how behavioral health documentation actually reads.

### 1.2 What Exists Today

The open source and commercial landscape for behavioral health text analysis includes:

| Tool / Model | Type | Trained On | Limitations |
|---|---|---|---|
| MentalBERT | Pre-trained language model | Reddit mental health subreddits | Base model only. Requires fine-tuning. Not trained on clinical notes. |
| KevSun/mentalhealth_LM | BERT classifier (0-5 severity) | Psychiatric diagnoses | 78% accuracy. Generic severity, not flag-specific. Social media text. |
| DunnBC22/canine-c | Binary classifier | Kaggle mental health corpus | Binary only (concern / no concern). No flag granularity. |
| sentinet/suicidality | BERT suicidality classifier | Reddit, Twitter, Kaggle | Social media only. Binary classification. Not validated on clinical text. |
| Various LSTM/BiLSTM models | Suicidal ideation detectors | Reddit SuicideWatch, Twitter | Research-grade. Not packaged for production. Binary only. |
| BiomedBERT | Pre-trained language model | PubMed abstracts + PMC full text | Strong biomedical language understanding. Not mental-health-specific. |
| Amazon Comprehend Medical | PHI detection, entity extraction | General medical text | No behavioral health-specific flag detection. Sentiment is a blunt instrument for BH context. |

A January 2025 study by Vance et al. (BMC Medical Informatics and Decision Making) trained a transformer model on actual clinical notes from the NeuroBlu behavioral health database, achieving F1 scores of 0.89-0.99 for suicidal ideation and anhedonia detection. This demonstrates that transformer models can achieve high accuracy on clinical behavioral health notes when trained on appropriate data. However, the model is proprietary, covers only two clinical concepts, and is not available as open source.

### 1.3 What Does Not Exist

There is no production-ready, open source tool that provides:

- Multi-flag behavioral health detection across self-harm, harm to others, medication, substance use, clinical deterioration, and protective factors
- Validation against clinical intake notes rather than social media posts
- A deployable service with HIPAA-appropriate architecture and zero data egress
- An extensible, version-controlled flag taxonomy with compliance mappings
- A deterministic safety layer (pattern matching) that does not depend on ML model accuracy for CRITICAL flag detection
- Compound risk rules that model clinically significant flag interactions
- FDA Non-Device CDS compliance by design

bh-sentinel is built to fill this gap.

### 1.4 Why Social Media Models Fail on Clinical Text

Models trained on Reddit and Twitter posts perform poorly on clinical documentation for several reasons. Clinical notes use a different language register: abbreviated charting notation (`+SI`, `-HI`, `AH present`), structured negation patterns (`Denies SI/HI`), temporal qualifiers (`History of attempt in 2019`), and professional clinical vocabulary that does not appear in social media corpora.

A patient posting “I want to die” on Reddit and a clinician writing “Pt denies SI, reports passive death wish with no plan or intent” describe overlapping clinical realities in fundamentally different language. A model trained exclusively on the first cannot reliably parse the second.

bh-sentinel addresses this through a hybrid architecture that combines deterministic pattern matching (optimized for clinical language and charting shorthand) with transformer-based classification (for implicit signals and nuanced language), running in parallel.

---

## 2. Design Philosophy

### 2.1 What bh-sentinel Does

bh-sentinel is designed to do three things:

1. **Detect concern categories** across a clinically informed, C-SSRS-mapped flag taxonomy covering 40 flags in 6 domains.
2. **Return explicit evidence context and confidence** so the clinician understands the basis of each flag - what text triggered it, where it appears, what detection layer produced it, and what the system's confidence level is.
3. **Trigger workflow, never autonomous clinical action.** Outputs feed escalation tasks, provider notifications, and case queue entries. The system never makes a clinical decision.

### 2.2 What bh-sentinel Does Not Do

- It does not diagnose any condition.
- It does not prescribe or recommend specific treatments.
- It does not take autonomous clinical action.
- It does not store, log, cache, or transmit clinical text outside the processing boundary.
- It does not replace validated clinical instruments (C-SSRS, PHQ-9, GAD-7). It prompts their administration.
- It is not FDA-cleared or approved.

### 2.3 Hard Constraints

| Constraint | Requirement | Rationale |
|---|---|---|
| Zero data egress | All PHI processing runs inside the deploying organization's AWS account and VPC. No clinical text | HIPAA compliance. PHI never leaves the organization's control boundary. |

| Constraint | Requirement | Rationale |
|---|---|---|
| | is ever sent to a third-party system. | |
| Sub-second latency | Full pipeline (text in, flags out) completes in under 1 second on warm invocations. | Clinical workflows cannot tolerate multi-second delays on safety-critical operations. |
| HIPAA-safe by design | No PHI is persisted, logged, cached, or transmitted outside the processing boundary. Only flag metadata (IDs, severity, confidence) is ever stored or logged. | Architectural PHI containment eliminates an entire class of compliance risk. |
| Clinician-in-the-loop | bh-sentinel flags possible signals. Clinicians assess using validated instruments. The system never makes autonomous clinical decisions. | Clinical safety decisions require human judgment. Technology assists, it does not decide. |

## 3. Clinical Flag Taxonomy

The bh-sentinel flag taxonomy defines 40 clinical safety flags organized into 6 domains. The taxonomy is informed by the Columbia Suicide Severity Rating Scale (C-SSRS), standard behavioral health screening instruments (PHQ-9, GAD-7), and common PHP/IOP clinical workflows.

The canonical machine-readable taxonomy is maintained in the project repository as `config/flag_taxonomy.json` with semantic versioning. All API responses include a `taxonomy_version` field so consuming systems can detect version drift.

### 3.1 Domains

| Domain | Flag Range | Count | Primary Clinical Scope |
|---|---|---|---|
| Self-Harm and Suicidal Ideation | SH-001 through SH-008 | 8 | Suicide spectrum from passive death wish through preparatory behavior, plus NSSI and attempt history |
| Harm to Others | HO-001 through HO-006 | 6 | Homicidal ideation spectrum, violent urges, abuse disclosure (perpetrating and experiencing) |
| Medication Non-Adherence | MED-001 through MED-005 | 5 | Stated and implied non-adherence, misuse, adverse reactions, desire to discontinue |
| Substance Use | SU-001 through SU-005 | 5 | Active use, relapse, escalation, withdrawal, risky use patterns |

| Domain | Flag Range | Count | Primary Clinical Scope |
|---|---|---|---|
| Clinical Deterioration | CD-001 through CD-008 (CD-005a-d) | 11 | Hopelessness, isolation, sleep/appetite disruption, psychosis (4 sub-flags), dissociation, acute anxiety, mania |
| Protective Factors | PF-001 through PF-005 | 5 | Treatment engagement, social support, future orientation, coping skills, medication adherence |

### 3.2 Severity Levels

| Level | Meaning | Expected Response |
|---|---|---|
| CRITICAL | Imminent risk to self or others | Immediate provider notification |
| HIGH | Significant clinical concern | Same-day provider review |
| MEDIUM | Clinically relevant, not urgent | Flag for next session review |
| LOW | Informational | Log for trend analysis |
| POSITIVE | Protective factor detected | Log for treatment progress |

Severity levels shown in the taxonomy tables below are defaults. The compound rules engine (Section 6) can escalate severity at analysis time based on co-occurring flags. For example, a passive death wish (SH-001, default HIGH) escalates to CRITICAL when it co-occurs with hopelessness (CD-001).

### 3.3 Domain: Self-Harm and Suicidal Ideation

The self-harm taxonomy is mapped to the Columbia Suicide Severity Rating Scale (C-SSRS), the most widely used standardized instrument for suicide risk assessment in the United States. C-SSRS mapping is shown for each flag where applicable.

| Flag | Name | Default Severity | Definition | Clinical Rationale | C-SSRS Mapping |
|---|---|---|---|---|---|
| SH-001 | Passive death wish | HIGH | Statements reflecting a wish to be dead, not wake up, or disappear without describing self-harm action. | Passive death wish is often the earliest detectable suicide-spectrum marker and raises concern even when intent is denied. | Wish to be Dead |
| SH-002 | Active suicidal ideation, nonspecific | CRITICAL | General suicidal thoughts without a stated method, plan, or preparatory step. | Active suicidal thinking marks a meaningful rise above passive ideation and warrants | Non-Specific Active Suicidal Thoughts |

| Flag | Name | Default Severity | Definition | Clinical Rationale | C-SSRS Mapping |
|---|---|---|---|---|---|
| | | | | urgent clinical review. | |
| SH-003 | Active suicidal ideation, with method | CRITICAL | Suicidal thinking that includes a described method, such as overdose, hanging, or firearm use, without a complete plan. | Access to or rehearsal of method increases feasibility and shortens the path from ideation to action. | Active Suicidal Ideation with Any Methods (Not Plan) without Intent to Act |
| SH-004 | Active suicidal ideation, with some intent | CRITICAL | Suicidal thinking accompanied by some expressed intent or expectation of acting, without a specific plan for when, where, or how. | Intent to act, even without a formed plan, marks a significant escalation in near-term risk and typically warrants same-day safety assessment. | Active Suicidal Ideation with Some Intent to Act, without Specific Plan |
| SH-005 | Active suicidal ideation, with plan and intent | CRITICAL | Suicidal thinking that includes both a specific plan and stated intent, expectation of acting, or inability to stay safe. | The combination of a specific plan with expressed intent represents the highest ideation-level severity on the C-SSRS and requires immediate intervention. | Active Suicidal Ideation with Specific Plan and Intent |
| SH-006 | Preparatory behavior | CRITICAL | Behaviors such as gathering pills, writing notes, giving away belongings, visiting a location, or rehearsing steps toward suicide. | Preparatory acts demonstrate transition from thought to behavior and are treated as near-attempt level acuity. | Preparatory Acts or Behavior |
| SH-007 | Self-injury without suicidal intent | HIGH | Deliberate self-injury such as cutting, burning, or hitting oneself without expressed desire to die. | Non-suicidal self-injury predicts emotional dysregulation and elevates future suicide risk even when intent is absent. | No direct C-SSRS equivalent |
| SH-008 | History of attempt referenced | HIGH | Mention of a prior suicide attempt, aborted attempt, or | Prior attempt history is among the strongest predictors of future suicidal | Actual Attempt (historical) |

| Flag | Name | Default Severity | Definition | Clinical Rationale | C-SSRS Mapping |
|---|---|---|---|---|---|
| | | | serious self-harm event in the past. | behavior and increases baseline risk. | |

### 3.4 Domain: Harm to Others

| Flag | Name | Default Severity | Definition | Clinical Rationale |
|---|---|---|---|---|
| HO-001 | Homicidal ideation, nonspecific | CRITICAL | General thoughts of killing, assaulting, or seriously harming another person without naming a target. | Even nonspecific homicidal ideation can escalate quickly and requires immediate violence-risk assessment. |
| HO-002 | Homicidal ideation, specific target | CRITICAL | Homicidal or violent thoughts directed toward an identifiable person or group. | A specific target materially increases plausibility, duty-to-protect considerations, and urgency of intervention. |
| HO-003 | Homicidal ideation, with plan | CRITICAL | Violent or homicidal thoughts that include a plan, method, timing, or preparation. | Planning substantially elevates imminent risk and may require emergency action and protective steps. |
| HO-004 | Violent urges | HIGH | Strong urges to hit, attack, or lose control physically, even without articulated intent. | Intense violent urges can precede impulsive assaultive behavior, especially under agitation or intoxication. |
| HO-005 | Abuse disclosure, perpetrating | HIGH | Disclosure of abusing, assaulting, coercing, or otherwise perpetrating harm toward another person. | Perpetration signals active risk to others and may trigger mandated reporting or safeguarding workflows. |
| HO-006 | Abuse disclosure, experiencing | HIGH | Disclosure of currently experiencing physical, sexual, emotional, or domestic abuse. | Active victimization is a major safety concern and can increase risk of suicide, retaliation, or acute destabilization. |

### 3.5 Domain: Medication Non-Adherence

| Flag | Name | Default Severity | Definition | Clinical Rationale |
|---|---|---|---|---|
| MED-001 | Non-adherence, stated | MEDIUM | Explicit statement that prescribed medication is not being taken as directed. | Stated non-adherence can directly worsen symptoms, destabilize treatment response, |

| Flag | Name | Default Severity | Definition | Clinical Rationale |
|---|---|---|---|---|
| | | | | and increase relapse risk. |
| MED-002 | Non-adherence, implied | MEDIUM | Indirect evidence of non-adherence, such as running out, cost barriers, avoidance, or self-stopping. | Implied adherence problems are common early indicators of treatment drift and merit follow-up before decompensation. |
| MED-003 | Medication misuse | HIGH | Taking more than prescribed, hoarding doses, mixing unsafely, or otherwise misusing medication. | Misuse raises overdose risk, toxicity risk, and in some cases creates suicide means concern. |
| MED-004 | Adverse reaction reported | MEDIUM | Reports of side effects, adverse reactions, or feeling harmed by medication. | Unaddressed adverse effects are a common driver of discontinuation and worsening engagement. |
| MED-005 | Desire to discontinue | MEDIUM | Expressed desire or plan to stop medication despite an active prescription. | Desire to discontinue may signal poor insight, emerging side effects, or impending non-adherence. |

## 3.6 Domain: Substance Use

| Flag | Name | Default Severity | Definition | Clinical Rationale |
|---|---|---|---|---|
| SU-001 | Active substance use disclosed | MEDIUM | Current use of alcohol, cannabis, illicit drugs, or misuse of prescribed substances. | Active use can impair judgment, worsen mood instability, and amplify both suicide and violence risk. |
| SU-002 | Relapse indicator | HIGH | Return to substance use after a period of sobriety or recovery. | Relapse often coincides with demoralization, reduced inhibition, and increased psychiatric instability. |
| SU-003 | Substance use escalation | HIGH | Increasing quantity, frequency, potency, or loss of control over use. | Escalation suggests worsening disorder severity and a higher likelihood of medical or behavioral consequences. |
| SU-004 | Withdrawal symptoms | HIGH | Symptoms of withdrawal such as tremor, sweats, nausea, agitation, or seizure risk. | Withdrawal can create acute medical danger and severe distress that compounds psychiatric risk. |

| Flag | Name | Default Severity | Definition | Clinical Rationale |
|---|---|---|---|---|
| SU-005 | Risky substance behavior | HIGH | High-risk patterns such as mixing substances, using alone, IV use, or intoxicated hazardous behavior. | Risky use patterns increase overdose, accidental death, and impulsive self-harm potential. |

## 3.7 Domain: Clinical Deterioration

### 3.7.1 General Clinical Deterioration Flags

| Flag | Name | Default Severity | Definition | Clinical Rationale |
|---|---|---|---|---|
| CD-001 | Hopelessness | HIGH | Pervasive belief that nothing will improve or that the future is irredeemably bleak. | Hopelessness is a well-established proximal correlate of suicide risk and severe depressive burden. |
| CD-002 | Severe isolation | MEDIUM | Marked withdrawal from family, friends, treatment, or normal social contact. | Severe isolation reduces buffering support and can allow risk to escalate unnoticed. |
| CD-003 | Sleep disruption, severe | MEDIUM | Major insomnia, near-total sleep loss, or extreme hypersomnia. | Severe sleep disruption worsens mood regulation, impulse control, psychosis risk, and suicide vulnerability. |
| CD-004 | Appetite disruption, severe | MEDIUM | Major reduction in intake, bingeing, purging, or other severe appetite disturbance. | Significant appetite disruption can reflect worsening depression, eating pathology, or medical compromise. |
| CD-006 | Dissociation | MEDIUM | Depersonalization, derealization, or feeling detached from self, body, or surroundings. | Dissociation can impair self-protection, increase distress, and complicate suicide risk assessment reliability. |
| CD-007 | Panic / acute anxiety | MEDIUM | Acute panic, extreme anxiety, or physiologic overwhelm causing loss of function. | Severe anxiety can rapidly intensify suicidal crises, impulsivity, and psychotic distress. |
| CD-008 | Mania indicators | HIGH | Indicators of mania such as decreased need for sleep, racing thoughts, grandiosity, pressured speech, or risky behavior. | Mania elevates impulsivity, impaired judgment, and exposure to high-risk behaviors requiring urgent treatment adjustment. |

#### 3.7.2 Psychosis Flags (CD-005a through CD-005d)

The original single “psychotic symptoms” concept has been decomposed into four granular sub-flags. This decomposition improves detection precision, enables more specific clinical routing, and allows the compound rules engine to model interactions between psychosis subtypes and other clinical signals.

| Flag | Name | Default Severity | Definition | Clinical Rationale |
|---|---|---|---|---|
| CD-005a | Auditory hallucinations | HIGH | Hearing voices, auditory command hallucinations, running commentary. | Auditory hallucinations, especially command content, can directly elevate risk of self-harm or aggression. |
| CD-005b | Visual hallucinations | HIGH | Seeing things others cannot see, visual distortions perceived as real. | Visual hallucinations indicate psychotic decompensation or substance/medical causes requiring urgent evaluation. |
| CD-005c | Paranoid ideation | HIGH | Belief of being watched, followed, plotted against, or targeted without evidence. | Paranoid ideation can drive defensive violence, elopement, or rapid disengagement from care. |
| CD-005d | Delusional thinking | HIGH | Fixed false beliefs resistant to evidence, including persecutory, grandiose, somatic, or referential delusions. | Delusional thinking signals significant loss of reality testing and can impair judgment around safety. |

**Anxiety-to-Psychosis Escalation:** When any psychosis flag (CD-005a through CD-005d) co-occurs with acute anxiety (CD-007), compound rule COMP-002 triggers. All detected psychosis flags escalate to CRITICAL severity, the session is marked for immediate review, and the recommended action is to evaluate for psychotic decompensation with anxiety prodrome. This captures the clinically significant pattern where anxiety is prodromal to, or reactive to, emerging psychotic experiences - requiring immediate psychiatric assessment rather than standard anxiety-track routing. See Section 6 for the full compound rules specification.

### 3.8 Domain: Protective Factors

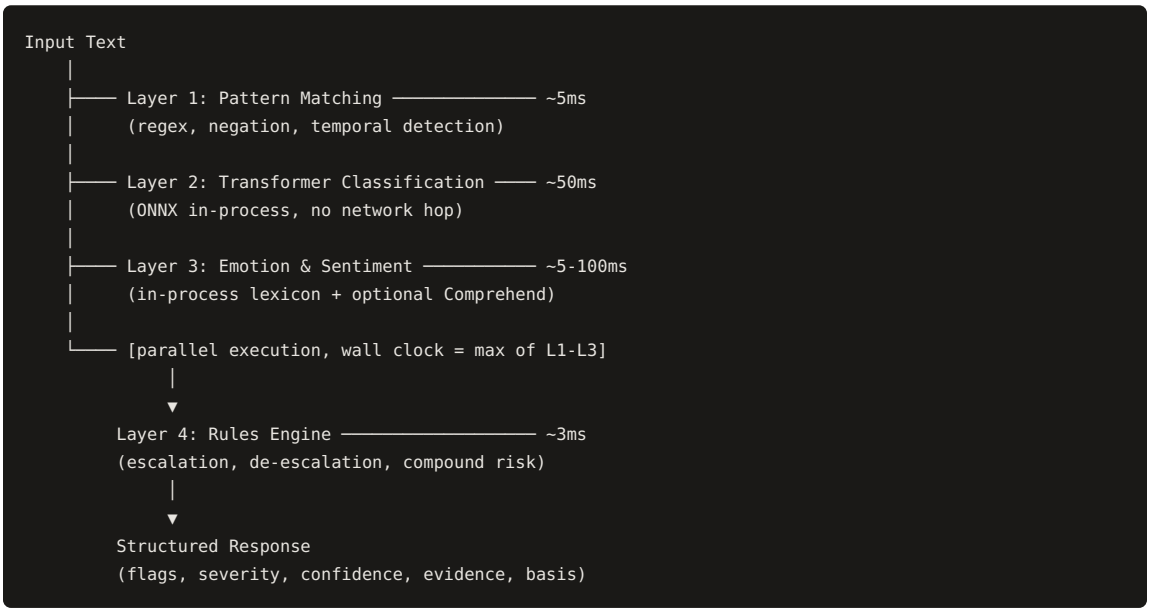
Protective factors are not absence-of-risk markers. They are positive clinical signals that indicate resilience, engagement, and support structures that buffer against acute decompensation. Detecting protective factors alongside risk factors gives clinicians a more complete picture and supports treatment progress documentation.

| Flag | Name | Default Severity | Definition | Clinical Rationale |
|---|---|---|---|---|
| PF-001 | Treatment engagement | POSITIVE | Clear commitment to therapy, psychiatry, safety planning, or follow-up care. | Engagement improves monitoring, responsiveness to intervention, and willingness to use crisis supports. |
| PF-002 | Social support | POSITIVE | Presence of supportive family, friends, partner, community, or other reliable interpersonal support. | Social support can reduce isolation, improve means restriction, and strengthen crisis containment. |
| PF-003 | Future orientation | POSITIVE | Statements about future plans, goals, responsibilities, or reasons to keep going. | Future orientation is protective because it counters entrapment and supports reasons for living. |
| PF-004 | Coping skills usage | POSITIVE | Active use of coping skills such as grounding, distraction, journaling, reaching out, or distress-tolerance tools. | Demonstrated coping capacity increases the likelihood of surviving acute distress without acting impulsively. |
| PF-005 | Medication adherence | POSITIVE | Consistent medication adherence with prescribed psychiatric treatment. | Reliable adherence supports symptom stabilization and lowers relapse and decompensation risk. |

## 4. Detection Architecture

### 4.1 Pipeline Overview

bh-sentinel uses a 4-layer parallel detection pipeline. Layers 1 through 3 execute concurrently via `asyncio.gather`. Layer 4 runs after all three complete, combining their outputs through the configurable rules engine.




Total wall clock time is the maximum of Layers 1-3 (not the sum), plus Layer 4. On warm invocations: approximately 60-120ms with Comprehend, approximately 55-60ms without.

### 4.2 Layer 1: Deterministic Pattern Matching

Layer 1 is the safety net. It uses compiled regex patterns from a YAML configuration file to detect explicit clinical language. Every CRITICAL-severity flag has pattern coverage in Layer 1. This means that even if the transformer model (Layer 2) fails, is miscalibrated, or produces a false negative, explicit safety-critical language is still caught deterministically.

**Pattern library scope:** Approximately 330 regex patterns across all 40 flags and 6 domains, organized into two registers:

- **Patient/narrative language** - patterns targeting how patients describe their experiences and how clinicians write narrative notes. Example: `"want(?:s|ed|ing)? to (?:kill|end) (?:myself|my life|it all)"`
- **Clinical shorthand** - patterns targeting charting abbreviations and structured documentation notation. Example: `"(?:\\+|positive) (?:for )?SI"` matching the common clinical shorthand `+SI`

These are maintained as separate pattern lists per flag because the language registers are structurally different and need independent review and testing.

**Confidence scoring:** Each flag in the pattern library has a default confidence score (0.0-1.0) reflecting the specificity of its pattern set. Highly specific, low-ambiguity patterns (e.g., SH-005 “active SI with plan and intent”) carry confidence of 0.96-0.97. Broader patterns covering fuzzier language (e.g., PF-002 “social support”) carry confidence of 0.78-0.80.

#### 4.2.1 Negation Detection

Clinical text is dense with negation. A single intake note might contain “Denies SI/HI. Reports feeling hopeless. No psychotic symptoms.” The negation detector must suppress SI and HI flags, suppress psychosis flags, and still detect the hopelessness flag - all from one paragraph.

bh-sentinel uses clause-bounded forward-window negation with scope termination. The algorithm:

1. When a pattern match is found, the detector scans backward (configurable window, default 5 tokens) for negation cues: `denies`, `no`, `not`, `never`, `negative`, `-` (shorthand), and flag-specific negation phrases.

2. If a negation cue is found, the detector checks whether a scope terminator appears between the negation cue and the match. Scope terminators are: period (`.`), comma (`,`), semicolon (`;`), `but`, `however`, `although`, `though`, `except`. If a terminator is present, the negation does not carry through.
3. Forward-window post-negation is also supported: `denied`, `absent`, `negative` appearing after the match within a configurable forward window suppress the match.
4. Conjunctive list connectors (`/`, `and`, `or`) are NOT scope terminators. `Denies SI/HI` suppresses both. `Denies SI and HI` suppresses both. This reflects clinical charting convention where slash-delimited and conjunction-linked lists share the negation.

**Pseudo-negation handling:** The following constructions look like negations but are not, and are explicitly excluded from suppression:

| Construction | Example | Why It Is Not a Negation |
|---|---|---|
| “no longer denies” | “Patient no longer denies suicidal ideation.” | Double negation. The patient IS endorsing SI. |
| “unable to” + symptom | “Unable to sleep for three days.” | Negates ability, not the symptom. The sleep disruption IS present. |
| “cannot stop” | “Cannot stop thinking about suicide.” | The thinking IS happening. |
| “no improvement in” | “No improvement in suicidal ideation.” | Negates progress, not the finding. SI IS present. |
| “no reason to live” | “Reports no reason to live.” | “No reason to live” is an affirmative passive death wish pattern (SH-001). The “no” is part of the clinical expression, not a negation cue. |

#### 4.2.2 Temporal Detection

Clinical notes routinely mix historical and current findings. “History of suicide attempt in 2019, currently denies SI” contains a past-tense SH-008 and a negated present-tense SH-002. The temporal detector classifies each pattern match as `present` or `past` using keyword proximity.

**Past markers:** `history of`, `hx of`, `PMH`, `prior`, `previous`, `past`, `former`, `remote`, `years ago`, `months ago`, `in [year]`, `as a child/teenager/adolescent`, `used to`, `had been`, `resolved`, `no longer`

**Present markers:** `currently`, `now`, `today`, `this week`, `last night`, `recently`, `right now`, `actively`, `presently`, `ongoing`, `again`, `started`, `resumed`, `returned`, `came back`

**Resolution rules:** - If both past and present markers are found for the same match, present wins. This is the safe default for a safety system: when in doubt, treat the signal as current. - If no temporal markers are found, the match defaults to `present`. Again, the safe default. - When the same flag is matched multiple times in a text (e.g., SH-007 matched in both a historical and current context), each match carries its own temporal classification. If any match is classified `present`, the flag is emitted as present.

**De-escalation:** Historical findings are de-escalated by one severity level via rule DE-001 in the rules engine. A historical suicide attempt (SH-008, default HIGH) would emit as MEDIUM. This preserves the signal for clinical context while reflecting reduced immediacy.

#### 4.2.3 Evidence Span Generation

Every pattern match generates an `evidence_span` - character offset metadata (start position, end position) identifying the exact text that triggered the flag. The integrating application resolves these offsets back to highlighted text in the source document, enabling the clinician to see precisely what the system detected and in what context.

This is not a usability feature. It is a compliance requirement. FDA Non-Device CDS Criterion 4 requires that the healthcare professional can independently review the basis of each recommendation. Evidence spans are what make that possible. See Section 8 for the full FDA analysis.

#### 4.2.4 Basis Description Generation

Each flag match also generates a `basis_description` - a human-readable explanation of what was detected and which detection layer produced it. Example: “Pattern match on patient language indicating active suicidal ideation without specific method or plan. Detected by Layer 1 (deterministic pattern matching) with 0.95 confidence.”

The `basis_description`, `evidence_span`, and `detection_layer` fields together satisfy the Criterion 4 transparency requirement: the clinician knows what was found, where it was found, and how it was found.

### 4.3 Layer 2: Transformer Classification

Layer 2 handles what deterministic pattern matching cannot: implicit signals, indirect language, euphemism, and clinical presentations that do not match any predefined phrase but carry clinical significance.

**Phase A (v0.1-v0.2): Zero-shot classification** using a quantized DistilBART-MNLI model exported to ONNX format and running entirely in-process. No network hop. The model receives the input text and a set of hypothesis templates derived from the flag taxonomy (e.g., “This text describes suicidal ideation” for flags SH-001 through SH-005) and returns per-hypothesis confidence scores.

Zero-shot classification provides useful signal on implicit and euphemistic language without requiring labeled training data. Its primary limitation is calibration: confidence scores from zero-shot models are not well-calibrated to clinical risk without additional tuning. During Phase A, a conservative discount factor is applied to Layer 2 confidence scores, and Layer 1 carries the primary detection burden for all CRITICAL flags.

**Phase B (v0.3+): Fine-tuned multi-label classification** using domain-adapted models trained on public behavioral health datasets (C-SSRS Reddit Suicide Dataset, UMD Reddit Suicidality Dataset v2, Kaggle Mental Health Sentiment, SMHD) and, where available, de-

identified clinical notes from deploying organizations. Fine-tuned models will replace or supplement the zero-shot layer, with A/B evaluation comparing per-flag precision and recall.

**Model specifications:** - Architecture: DistilBERT (66M parameters) or DistilBART-MNLI (139M parameters) - Quantization: INT8 via ONNX Runtime - Model size: approximately 65MB (DistilBERT INT8) - Inference time: approximately 50ms per request (in-process, no network) - Container overhead: total Lambda image approximately 333MB including model, well within the 10GB limit

#### 4.4 Layer 3: Emotion and Sentiment Analysis

Layer 3 provides emotional context that supports clinical interpretation of detected flags. It is explicitly designed as a **supporting signal**, not a primary safety detector.

**In-process emotion lexicon:** A curated behavioral health emotion lexicon containing approximately 250 terms mapped to 11 categories: hopelessness, agitation, anxiety, anger, sadness, guilt, shame, mania, dissociation, positive valence, and negative valence. Terms can belong to multiple categories (e.g., “trapped” maps to hopelessness, anxiety, and negative valence). Scoring produces a 0.0-1.0 density score per category.

The lexicon is curated from standard behavioral health clinical vocabulary, DSM-5 descriptive language, common patient-reported emotional terms, and clinical assessment instrument language. No proprietary datasets were used. It is a project-owned seed lexicon intended for clinical team review and extension.

**Optional AWS Comprehend integration:** Amazon Comprehend provides general sentiment analysis (positive/negative/neutral/mixed) and entity extraction. Comprehend Medical adds PHI detection and medical entity extraction. Both are HIPAA-eligible, coverable under an AWS BAA, and accessible via VPC endpoint so traffic never leaves the VPC.

**Why Comprehend sentiment is insufficient alone:** A patient describing a suicide attempt in clinical, detached language could score as "neutral" sentiment. A patient expressing intense anger about a parking ticket could score as strongly "negative." Sentiment polarity does not map reliably to clinical risk in behavioral health contexts. This is why bh-sentinel uses the domain-specific emotion lexicon as its primary Layer 3 signal and treats Comprehend as an optional supplement.

**Degradation strategy:** Comprehend is accessed via VPC endpoint with a 200ms timeout. If Comprehend is unavailable, slow, or not configured, Layer 3 falls back to the in-process emotion lexicon alone. The pipeline never fails due to Comprehend unavailability.

### 4.5 Layer 4: Configurable Rules Engine

Layer 4 is where individual signals become clinical risk assessments. It takes the outputs of Layers 1-3 and applies a configurable set of rules to escalate severity, de-escalate historical findings, identify compound risk patterns, and generate recommended clinical actions.

Rules are defined in a JSON configuration file (`config/rules.json`) with semantic versioning and a dependency on the taxonomy version. The rules engine evaluates rules in a defined order: escalation rules first, then de-escalation rules, then compound rules, then action rules.

The rules engine supports the following condition types:

| Condition Type | Description |
|---|---|
| `flag_present` | A specific flag was detected, with optional minimum confidence threshold. |
| `any_flag_present` | Any of a listed set of flags was detected. |
| `domain_present` | Any flag in a specified domain was detected. |
| `domain_severity` | Any flag in a specified domain reached at least a given severity (evaluated after escalation). |
| `temporal_context` | The temporal qualifier of the matched text (past or present). |
| `emotion_above` | A Layer 3 emotion score exceeds a given threshold. Used sparingly; flag co-occurrence is preferred. |
| `domain_flag_count` | Number of flags in a domain meets or exceeds a minimum count, with option to include sub-threshold candidates. |

## 5. Escalation Rules

Escalation rules model clinically significant interactions between flags. When two or more signals co-occur, the combined risk profile can be materially higher than either signal alone. The rules below are informed by clinical risk assessment literature and reviewed with behavioral health clinical leadership.

Each rule includes an ID, the clinical scenario it addresses, its trigger conditions, the resulting action, and the clinical rationale for the escalation.

#### ESC-001: Passive SI + Hopelessness

**Scenario:** Passive suicidal ideation (SH-001) co-occurs with pervasive hopelessness (CD-001).

**Action:** Escalate SH-001 to CRITICAL. Recommend C-SSRS administration.

**Rationale:** Hopelessness is a well-established proximal correlate of suicide. When combined with expressed death wish, the risk profile exceeds what either signal indicates alone. C-SSRS administration can differentiate passive from active ideation and identify intent that the patient may not have spontaneously disclosed.

#### ESC-002: Self-Injury + Substance Use

**Scenario:** Non-suicidal self-injury (SH-007) co-occurs with any substance use flag.

**Action:** Escalate SH-007 to CRITICAL. Recommend assessment for suicidal intent in context of active substance use and self-injury.

**Rationale:** Substances lower behavioral inhibition and increase the probability that NSSI escalates to suicidal behavior. The transition from self-injury to suicide attempt is not always preceded by expressed intent, particularly under intoxication.

#### ESC-003: Active SI + Command Auditory Hallucinations

**Scenario:** Any active suicidal ideation flag (SH-002 through SH-005) co-occurs with auditory hallucinations (CD-005a).

**Action:** Escalate all co-occurring SI and AH flags to CRITICAL. Mark session for immediate review. Recommend emergency evaluation.

**Rationale:** Command auditory hallucinations directing self-harm represent an acute emergency. The patient may act on command content with little warning or volitional control.

#### ESC-004: Homicidal Ideation + Substance Use

**Scenario:** Any harm-to-others flag co-occurs with any substance use flag.

**Action:** Escalate all HO flags to CRITICAL. Mark session for immediate review. Recommend violence risk assessment.

**Rationale:** Intoxication amplifies violence risk and impulsivity. Homicidal ideation under the influence of substances requires immediate provider notification and may trigger duty-to-protect obligations.

#### ESC-005: Mania + Risky Substance Behavior

**Scenario:** Mania indicators (CD-008) co-occur with risky substance behavior (SU-005).

**Action:** Escalate CD-008 to CRITICAL. Mark session for immediate review. Recommend acute psychiatric evaluation.

**Rationale:** Substance use during manic episodes is a common and dangerous destabilization pattern in bipolar disorder. Impaired judgment compounds with substance-induced disinhibition.

#### ESC-006: Prior Attempt + Current SI

**Scenario:** History of suicide attempt (SH-008) co-occurs with any current suicidal ideation flag (SH-001 through SH-005).

**Action:** Escalate all current SI flags to CRITICAL. Recommend C-SSRS administration.

**Rationale:** Prior attempt history is among the strongest predictors of future suicidal behavior. Current ideation in a patient with attempt history is a high-risk combination that warrants escalated response regardless of expressed intent level.

#### ESC-007: Hopelessness + Severe Isolation

**Scenario:** Hopelessness (CD-001) co-occurs with severe isolation (CD-002).

**Action:** Escalate both to CRITICAL. Mark session for immediate review. Recommend assessment for acute decompensation.

**Rationale:** Isolation removes the protective buffering of social support and allows risk to escalate unmonitored. Hopelessness combined with isolation eliminates both internal (hope) and external (people) protective factors simultaneously.

---

## 6. Compound Risk Rules

Compound rules detect clinically significant patterns that do not necessarily change individual flag severity but require specific clinical attention or workflow routing.

#### COMP-001: Substance Use + Self-Harm Co-occurrence

**Trigger:** Any substance use flag co-occurs with any self-harm flag.

**Action:** Mark session for immediate review. Recommend co-occurring risk assessment.

**Rationale:** The intersection of active substance use and suicidal ideation or behavior is one of the highest-risk clinical combinations. Substances lower inhibition, impair judgment, and increase lethality of self-harm behavior.

#### COMP-002: Anxiety-to-Psychosis Escalation

**Trigger:** Acute anxiety (CD-007) co-occurs with any psychosis flag (CD-005a through CD-005d).

**Action:** Escalate all psychosis flags to CRITICAL. Mark session for immediate review. Recommend evaluation for psychotic decompensation with anxiety prodrome.

**Rationale:** Anxiety is frequently prodromal to, or reactive to, emerging psychotic experiences. This compound signal suggests the patient may be in active psychotic decompensation rather than experiencing isolated anxiety, requiring immediate psychiatric assessment rather than standard anxiety-track routing.

#### COMP-003: Self-Harm Signal Density

**Trigger:** Three or more self-harm domain flags detected (including sub-threshold candidates that passed negation and temporal detection but fell below the confidence threshold).

**Action:** Mark session for immediate review. Recommend full-note review for accumulated risk signals.

**Rationale:** Multiple self-harm indicators scattered across a clinical note, even if individually below confidence thresholds, suggest accumulated preparatory or risk behavior that may not be apparent from any single signal. The density of sub-threshold signals is itself a clinical signal.

#### COMP-004: Harm-to-Others Signal Density

**Trigger:** Three or more harm-to-others domain flags detected (including sub-threshold candidates).

**Action:** Mark session for immediate review. Recommend full-note review for accumulated violence risk.

**Rationale:** Same density logic as COMP-003, applied to the harm-to-others domain.

#### COMP-005: Medication Misuse + Self-Harm

**Trigger:** Medication misuse (MED-003) co-occurs with any self-harm flag.

**Action:** Mark session for immediate review. Recommend assessment for preparatory behavior.

**Rationale:** Medication stockpiling or misuse in a patient expressing self-harm signals may indicate means acquisition. This pattern warrants specific assessment for preparatory behavior (SH-006) even when the patient has not explicitly described preparatory acts.

#### COMP-006: Homicidal Ideation + Paranoid Ideation

**Trigger:** Any harm-to-others flag co-occurs with paranoid ideation (CD-005c).

**Action:** Mark session for immediate review. Recommend violence risk assessment with focus on paranoia-driven threat perception.

**Rationale:** Paranoia can drive defensive violence by causing the patient to perceive threats that do not exist, increasing the plausibility that homicidal ideation will be acted upon. Evaluation of reality testing is essential.

#### COMP-007: Self-Injury + Passive Death Wish

**Trigger:** Non-suicidal self-injury (SH-007) co-occurs with passive death wish (SH-001).

**Action:** Mark session for immediate review. Recommend C-SSRS administration.

**Rationale:** Active self-injury combined with expressed desire to be dead suggests escalating risk even when neither flag alone reaches CRITICAL severity. The combination indicates the patient is both behaviorally harming themselves and ideating about death.

#### De-escalation: Historical References (DE-001)

**Trigger:** A flag's temporal context is classified as `past` by the temporal detector.

**Action:** Reduce severity by one level (CRITICAL becomes HIGH, HIGH becomes MEDIUM, MEDIUM becomes LOW).

**Rationale:** Historical findings carry clinical relevance (particularly attempt history, which is a predictor regardless of recency) but do not represent the same immediacy as current findings. De-escalation preserves the signal while reducing alert urgency.

#### Action Rule: C-SSRS Prompt (ACT-001)

**Trigger:** Any self-harm domain flag reaches CRITICAL severity (after escalation rules have been applied).

**Action:** Recommend C-SSRS administration per organization protocol.

**Rationale:** CRITICAL self-harm flags indicate a level of risk that warrants structured suicide risk assessment using a validated instrument. bh-sentinel prompts the administration of C-SSRS; it does not administer, score, or interpret the instrument.

---

## 7. Emotion Lexicon

### 7.1 Design

The bh-sentinel emotion lexicon is a curated vocabulary of approximately 250 terms organized into 11 emotion categories relevant to behavioral health clinical assessment:

| Category | Description | Example Terms |
|---|---|---|
| Hopelessness | Language expressing futility, entrapment, or absence of future | hopeless, trapped, futile, no way out, nothing matters |
| Agitation | Language expressing restlessness, irritability, or inability to be still | agitated, pacing, wound up, crawling out of my skin |
| Anxiety | Language expressing worry, fear, panic, or physiologic overwhelm | anxious, terrified, panicking, heart racing, can't breathe |
| Anger | Language expressing hostility, rage, or aggressive feeling | furious, enraged, seething, hostile, seeing red |
| Sadness | Language expressing grief, loss, emptiness, or depressed mood | crying, miserable, empty, grieving, heartbroken, anhedonia |
| Guilt | Language expressing self-blame, regret, or responsibility for harm | guilty, my fault, regret, let everyone down, I caused this |
| Shame | Language expressing worthlessness, defectiveness, or self-disgust | worthless, pathetic, defective, burden, waste of space |
| Mania | Language expressing grandiosity, invincibility, or elevated energy | invincible, unstoppable, euphoric, limitless, don't need sleep |
| Dissociation | Language expressing detachment, numbness, or unreality | detached, numb, unreal, watching myself, losing time |
| Positive valence | Language expressing wellbeing, progress, or stability | hopeful, grateful, improving, resilient, safe, calm |
| Negative valence | Language expressing distress, suffering, or deterioration | terrible, suffering, falling apart, unbearable, shattered |

Terms can belong to multiple categories. "Trapped" maps to hopelessness, anxiety, and negative valence simultaneously, reflecting the clinical reality that emotional states are not discrete categories.

### 7.2 Scoring

For each category, the lexicon counts the number of input words that map to that category and divides by total word count, producing a 0.0-1.0 density score. Categories are not mutually exclusive. A text can score high on both hopelessness and anxiety.

### 7.3 Clinical Use

Emotion scores are used in the rules engine as secondary signals. The `emotion_above` condition type allows rules to incorporate emotional context, but flag co-occurrence is preferred over raw emotion scores for clinical decisions. The primary role of the emotion layer is to provide clinicians with an emotional profile of the text that contextualizes the flags detected by Layers 1 and 2.

---

## 8. FDA Non-Device Clinical Decision Support Compliance

### 8.1 Regulatory Framework

bh-sentinel is designed to satisfy the four criteria for Non-Device Clinical Decision Support under Section 520(o)(1)(E) of the FD&C Act (21st Century Cures Act). Software that meets all four criteria is excluded from the definition of a medical device and is not subject to FDA premarket review.

This analysis reflects the FDA's January 2026 updated guidance on Clinical Decision Support Software. Organizations deploying bh-sentinel should review this analysis with their own legal counsel.

### 8.2 Criterion 1: Does Not Acquire, Process, or Analyze a Medical Image or Signal

bh-sentinel operates on already-generated clinical documentation and structured medical information. It does not ingest radiology images, waveform data, or physiologic signal streams for diagnostic interpretation. The product's core function is to identify clinical safety signals within text and record-derived artifacts.

To maintain compliance with this criterion: - Inputs are limited to textual records, coded data, and other human-interpretable clinical information. - Any upstream conversion of signals or images into narrative reports must occur outside bh-sentinel. - Product messaging must not imply that bh-sentinel interprets primary diagnostic data.

If a future implementation introduces analysis of DICOM images, bedside monitoring streams, or algorithmic feature extraction from physiologic data, this criterion would require re-evaluation.

### 8.3 Criterion 2: Displays, Analyzes, or Prints Medical Information

bh-sentinel organizes and analyzes existing medical information supplied in the clinical record. Its outputs are tied to identifiable source material and are intended to highlight what is present, missing, inconsistent, or potentially relevant for downstream human review.

This criterion is satisfied because the software displays and analyzes medical information that the healthcare professional can already review in the patient's record. In practice: - Each flag maintains clear linkage to underlying source content via `evidence_span`. - Output phrasing frames results as record-based findings or prompts for review. - User experience patterns emphasize inspection of existing information rather than passive acceptance of a conclusion.

### 8.4 Criterion 3: Supports or Provides Recommendations to a Healthcare Professional

bh-sentinel supports professional decision-making by surfacing clinical safety signals and recommending follow-up actions (e.g., "Administer C-SSRS per organization protocol"). This is consistent with the CDS exclusion so long as the product remains assistive and is directed to healthcare professionals rather than autonomous patient-facing decision-making.

The compliance position is strongest when: - Flags and recommendations are presented as decision support for qualified users. - Clinical judgment remains necessary before any diagnosis, treatment, or coding action is taken. - Language avoids suggesting that output is final, determinative, or self-executing.

### 8.5 Criterion 4: Enables Independent Review of the Basis for the Recommendation

Criterion 4 is the most important and typically the highest-risk element of the Non-Device CDS analysis. bh-sentinel addresses it through three fields in every flag response:

| Field | Purpose | Content |
|---|---|---|
| `basis_description` | Human-readable explanation of what was detected and why. | Example: “Pattern match on patient language indicating active suicidal ideation without specific method or plan.” |
| `evidence_span` | Character offset metadata identifying the exact text that triggered the flag. | Start position, end position, matched text reference. The integrating interface resolves these offsets to highlighted text in the source document. |
| `detection_layer` | Which layer of the detection pipeline produced the flag. | “pattern_match”, “transformer_classification”, or “compound_rule” |

Together, these fields enable the healthcare professional to answer: - What record content caused this flag? - Where in the chart or source material can that content be reviewed? - Is the output a direct extraction, rule-based inference, or model-assisted pattern match? - What uncertainty, limitation, or ambiguity should be understood before acting?

Independent review is weakened if the system presents opaque risk scores, unexplained classifications, or recommendations that cannot be traced back to source material. Future product changes must preserve explainability as a release-blocking requirement, not a secondary usability feature.

### 8.6 Risk Assessment

The primary regulatory risk is not that bh-sentinel performs support functions, but that users or deployers may treat it as an authoritative clinical decision engine. The highest-risk scenarios are:

- Outputs that appear conclusive without sufficient source attribution.
- User workflows that allow bulk acceptance of recommendations without chart review.
- Marketing or training materials that overstate accuracy or imply autonomous clinical judgment.
- Expansion of inputs into image, signal, or other device-like data sources.

- Product evolution toward closed-model scoring without transparent rationale.

Operational risk also exists if false positives create alert fatigue or if false negatives cause teams to overestimate record completeness. These concerns do not automatically change regulatory status, but they affect whether real-world use remains consistent with the assumptions supporting the Non-Device CDS position.

#### 8.7 Deployment Governance

Organizations deploying bh-sentinel should align product configuration, clinical workflow, and governance with the above criteria. Recommended controls:

- Limit use to qualified healthcare professionals or trained clinical documentation personnel under professional supervision.
- Preserve display of `basis_description`, `evidence_span`, and related rationale fields in the production interface. Render span-based evidence in visible source context rather than as offsets alone.
- Require human review before documentation, coding, diagnosis, or treatment changes are finalized.
- Maintain audit logs showing what recommendation was presented, what evidence was shown, and what action the user took.
- Train users that bh-sentinel is a decision-support aid, not a substitute for clinical judgment.
- Review new features for regulatory impact before release, especially features involving automation, summarization, prioritization, or new data modalities.

The Non-Device CDS rationale should be treated as a living compliance position. Material product changes, new integrations, or new marketing claims should trigger renewed legal and regulatory review.

---

## 9. Deployment Architecture

### 9.1 Reference Architecture

bh-sentinel is designed for deployment as an AWS Lambda function inside a private VPC subnet with no internet access. All AWS service access (Comprehend, S3, DynamoDB) is via VPC endpoints so traffic never traverses the public internet.

**Key properties:** - Zero data egress: Lambda runs in a private subnet. Security groups allow no outbound internet traffic. VPC Flow Logs verify compliance. - Sub-second latency: approximately 60-120ms on warm invocations. Provisioned concurrency (configurable, default 3 instances) eliminates cold start latency for production traffic. - No PHI persistence: text is processed in-memory and discarded. Only flag metadata (IDs, severity, confidence) is ever stored or logged. - Cost-effective: approximately $50-115/month for the full stack, making clinical safety detection accessible to small and mid-size behavioral health organizations that cannot afford dedicated infrastructure teams.

### 9.2 Infrastructure Components

| Component | Service | Purpose |
|---|---|---|
| Compute | AWS Lambda | Stateless text analysis with ONNX model inference |
| Container Registry | Amazon ECR | Lambda container image with model artifacts |
| API | API Gateway | HTTPS endpoint with rate limiting and authentication |
| Networking | VPC + Private Subnets | Zero-egress boundary for PHI containment |
| Service Access | VPC Endpoints | Private connectivity to Comprehend, S3, DynamoDB |
| Monitoring | CloudWatch | Request volume, latency percentiles, per-layer breakdown, CRITICAL flag volume, error rates |
| Audit | bh-audit-logger | Compliance-grade audit trail for analysis events |
| Infrastructure as Code | Terraform | Reproducible deployment modules |

### 9.3 Why This Architecture Matters

Most behavioral health organizations are small. They do not have DevOps teams, security engineers, or infrastructure architects. If clinical safety detection requires standing up Kubernetes clusters, managing GPU instances, or configuring complex networking, it will not be adopted.

bh-sentinel's deployment is designed to be a single Terraform apply. The infrastructure modules handle VPC configuration, security groups, Lambda deployment, API Gateway setup, and monitoring. The cost model ($50-115/month) is within reach of PHP/IOP programs that operate on thin margins.

This is a deliberate architectural decision. Accessibility is not a secondary concern - it is a design requirement. A clinical safety tool that only large health systems can deploy does not solve the problem.

---

## 10. What bh-sentinel Is Not

This section exists because clarity about limitations is as important as clarity about capabilities, particularly for a tool operating in clinical safety.

**bh-sentinel is not a diagnostic tool.** It does not diagnose any medical or psychiatric condition. It detects signals in text that may indicate clinical concern. Diagnosis requires clinical assessment by a qualified provider.

**bh-sentinel is not a substitute for clinical judgment.** Every flag, severity level, and recommended action is intended to support the clinician's independent assessment. The system does not make clinical decisions, and no clinical decision should be based solely on its output.

**bh-sentinel is not FDA-cleared or approved.** The architecture is designed to satisfy the criteria for Non-Device CDS exclusion from device regulation. This is an analytical position, not a regulatory determination. Organizations deploying the system should review the analysis with their own legal counsel.

**bh-sentinel is not validated on your patient population.** The pattern library and transformer models reflect general behavioral health clinical language. Clinical validation against your organization's documentation practices, patient population, and clinical workflows is essential before production use.

**bh-sentinel is not a replacement for validated clinical instruments.** It prompts the administration of instruments like the C-SSRS, PHQ-9, and GAD-7. It does not administer, score, or interpret them.

**bh-sentinel does not store, log, or transmit clinical text.** The system processes text ephemerally. Only flag metadata is ever persisted. This is an architectural guarantee, not a configuration option.

---

## 11. Roadmap

#### v0.1: Core Pattern Engine (In Progress)

Core package scaffolding. Pydantic request/response models. Flag taxonomy v1.0 with version tracking. Text preprocessor with sentence splitting and character offset tracking. Negation detector, temporal detector, pattern matcher. Emotion lexicon integration. Rules engine with full escalation, de-escalation, compound risk, and action rule support. Pipeline orchestrator with parallel execution and graceful degradation. Pattern library covering all 40 flags across all 6 domains. Configuration validation CLI. Pattern test runner with clinician-readable output. Comprehensive test fixtures. PHI containment tests. Publication to PyPI as bh-sentinel-core.

#### v0.2: ML Layer + Reference Deployment

ONNX transformer inference layer (in-process). Zero-shot classification with DistilBART-MNLI. Model export tooling. Transformer evidence span and basis description generation. Confidence calibration. Flag deduplication and layer merge. AWS Comprehend integration (optional, degradable). AWS Lambda reference deployment with Terraform modules. CloudWatch monitoring and alerting. Load testing and latency validation.

#### v0.3: Fine-Tuning, Evaluation, and Spanish Language Support

Public dataset preparation scripts. Domain adaptation pre-training. Fine-tuning pipeline with multi-label classification head. A/B evaluation framework (zero-shot vs fine-tuned). Clinician feedback integration. Alert fatigue monitoring. Spanish language pattern library across all 40 flags. Multilingual transformer support. Spanish negation and temporal marker lists. Spanish-language test suite.

#### Future

EMR webhook integration examples. Flag trend analysis utilities. FHIR-compatible output format. Group session note analysis patterns. Model retraining pipeline. Clinician feedback dashboard.

---

## 12. References


1. Columbia Suicide Severity Rating Scale (C-SSRS). The Columbia Lighthouse Project. Columbia University.
2. FDA Guidance: Clinical Decision Support Software (Updated January 2026). US Food and Drug Administration.
3. 21st Century Cures Act, Section 520(o)(1)(E) of the FD&C Act.
4. Vance MC, et al. “Detection of suicidal ideation and anhedonia in clinical notes using transformer-based NLP.” BMC Medical Informatics and Decision Making. January 2025.
5. PHQ-9 (Patient Health Questionnaire-9). Kroenke K, Spitzer RL, Williams JB.

6. GAD-7 (Generalized Anxiety Disorder-7). Spitzer RL, Kroenke K, Williams JB, Löwe B.
7. HIPAA Security Rule, 45 CFR Part 164 Subpart C.
8. 42 CFR Part 2: Confidentiality of Substance Use Disorder Patient Records.

---

*This document should be reviewed with qualified behavioral health clinical leadership before any production deployment. The flag taxonomy requires clinical validation and may require modifications based on program-specific needs.*

*FDA CDS compliance analysis reflects the January 2026 updated guidance. Deploying organizations should review this analysis with their own legal counsel, particularly if extending the tool to patient-facing applications.*

*bh-sentinel is clinical decision support software. It is not a diagnostic tool, not a substitute for clinical judgment, and not FDA-cleared or approved. Organizations deploying this software in clinical settings are responsible for their own clinical validation, regulatory compliance, and patient safety protocols.*

*bh-sentinel is published under the Apache 2.0 license by the bh-healthcare open source organization.*