Schema-Constrained, Clinician-Verified LLM Abstraction to Improve Documentation Completeness for CNS Trial Eligibility Screening: A Real-World Methodology Pilot

Padder, T[1] · Nisar, A[1]

Real-world methodology pilot — 150 real, de-identified psychiatric encounters (75 outpatient + 75 inpatient). No simulated cases. Preliminary; not a clinical efficacy study.

150	75 / 75	+4.3	p<0.001	4.9 → 3.1	96%
Real psychiatric encounters	Outpatient / Inpatient	Mean completeness gain (of 24)	Paired t & Wilcoxon (d_z=1.97)	Trial-critical gaps per case (of 7)	Inpatient admissions trial-ineligible

1 · Background & Methodological Issue

Why CNS psychiatric trial screening is challenging

- Screen-fail rates remain high; intake documentation gaps can drive both avoidable exclusions and avoidable enrollments.
- Psychiatric notes are narrative and are not built for downstream eligibility extraction; completeness also varies by care setting.

Documentation deficits in routine practice

- Diagnostic specifiers, episode polarity, and remission status are often absent.
- Prior medication trial doses, durations, and adequacy are rarely structured.
- Suicide risk: narrative statements such as “denies SI” are common; validated tools such as the C-SSRS are seldom recorded.
- Mania/hypomania, psychosis, and substance-use details are often implicit.

Gap addressed by this pilot

A reproducible, trial-specific, **schema-constrained abstraction layer operating under mandatory clinician verification** converts existing narrative documentation into trial-relevant structured fields, flags missing items rather than imputing them, and surfaces protocol-relevant exclusion concerns before formal screening. LLM-based extraction already exists; **the contribution here is the constrained, clinician-verified workflow, not AI itself.**

2 · Methods & Workflow

Design — real-only, two settings

- **Outpatient:** 75 de-identified encounters (specialty practice + community clinic).
- **Inpatient:** 75 de-identified acute admission psychiatric evaluations.
- All 150 cases are **real and de-identified. No simulated cases; nothing invented.**
- Primary comparison axis: **outpatient vs inpatient documentation.**
- Retrospective real-world pilot; each case served as **its own control** (original documentation vs schema-constrained, clinician-verified abstraction).
- Encounter-level unit of analysis; one documentation event per case.

De-identification

- Identifiers removed: names, MRNs, exact dates of birth, addresses, contact information, employer/school details.
- Ages grouped in 5-year bands; medication names generic; substance-use information qualitative.

Schema-constrained structured abstraction

- **Twelve pre-specified domains** scored 0 = absent, 1 = partial, 2 = complete (**24-point score**), using an operationalized rubric with explicit anchor definitions per level to minimize rater subjectivity.
- This Documentation Completeness Score is an **investigator-developed rubric** constructed *a priori* from trial eligibility requirements for methodological comparison — **not a psychometrically validated instrument**.
- **Sixteen pre-specified** trial-relevant safety/eligibility flags applied per case.
- Structures available content, identifies implicit content, and flags missing items — **without imputing clinical facts**.
- Fixed extraction schema with **mandatory clinician review of every output**; the model extracts or flags-missing only and is barred from clinical inference. Field-level extraction fidelity (structured fields vs source text) was clinician-confirmed.
- Feasibility / effect-size design: single trained rater; a powered two-reviewer reliability sub-study (κ) is the **pre-registered next step, not an omission**.

Statistical analysis

- **Within-case:** paired *t*-test (primary), Wilcoxon signed-rank (sensitivity); Cohen’s d_z; 95% CIs (*t* and bootstrap).
- **Domain level:** McNemar exact test, by setting and overall.
- **Outpatient vs inpatient:** Welch *t*-test and Mann–Whitney test. Proportions: 95% Wilson CIs.
- Cohen’s κ was pre-specified but was not computable because no second-reviewer data were available.

Table 1. Case characteristics (real, N = 150)

Variable	Outpt (75)	Inpt (75)	Total (150)
Sex F : M	43:32	45:30	**88:62**
Pediatric (12–17)	17	21	**38**
Adult (≥18)	58	54	**112**
≥1 psychiatric comorbidity	54	55	**109**
≥1 medical comorbidity	34	73	**107**
Validated severity scale (complete)	0	0	**0**

Fig. 1. Abstraction workflow (every output clinician-verified)

1. De-identified source note →
2. Fixed 12-domain schema applied →
3. Extract present / flag missing *(no imputation)* →
4. Mandatory clinician verification →
5. Structured fields + 16 eligibility flags

The model may extract or mark-as-missing only; clinical inference is barred and every field is clinician-confirmed against source text before scoring.

3 · Primary Completeness Outcome

Fig. 2 Documentation Completeness: Original vs Verified Abstraction (paired within-case; p<0.001 all groups)

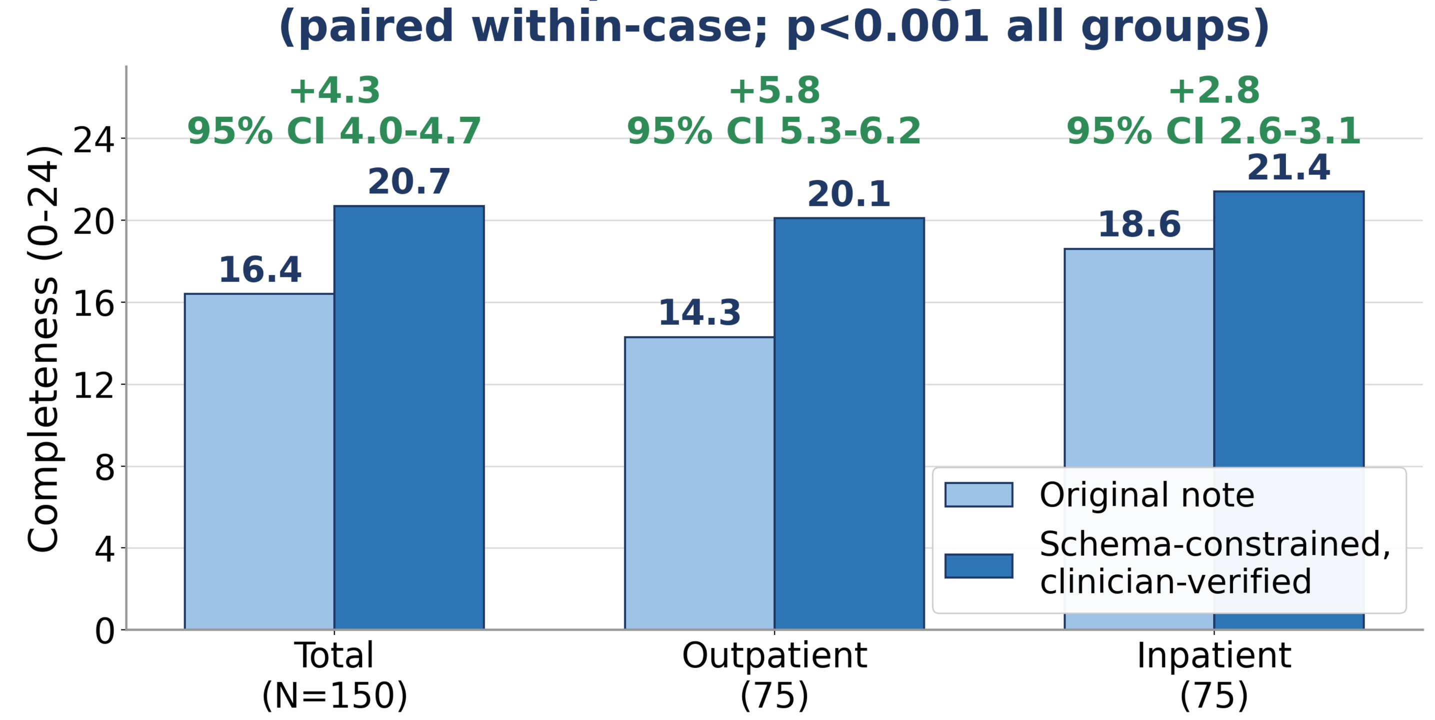

+4.3
95% CI 4.0-4.7
+5.8
95% CI 5.3-6.2
+2.8
95% CI 2.6-3.1
16.4
20.7
14.3
20.1
18.6
21.4
Completeness (0-24)
Total (N=150)
Outpatient (75)
Inpatient (75)
Original note
Schema-constrained, clinician-verified

Fig. 2. Original vs clinician-verified structured completeness; paired within-case, p<0.001 for all groups.

Setting Contrast & Domain-Level Detail

Fig. 3 Inpatient notes start more complete; structuring adds more to outpatient documentation

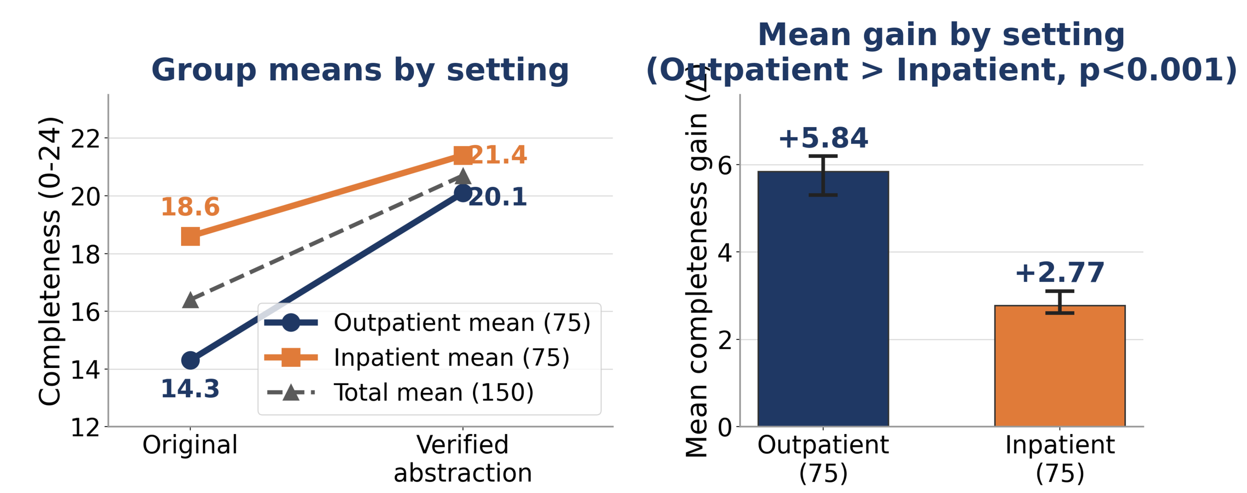
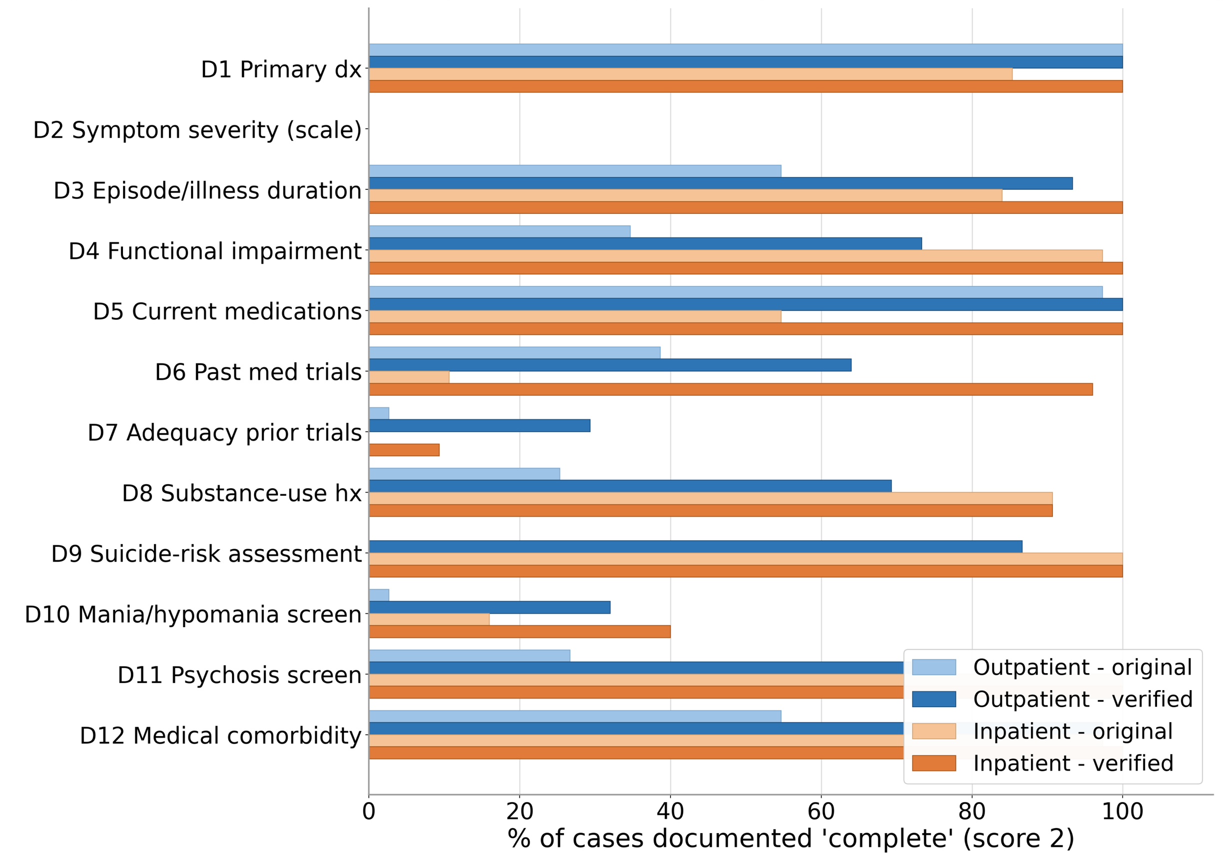

Group means by setting
Completeness (0-24)
18.6
21.4
20.1
14.3
Original
Verified abstraction
Outpatient mean (75)
Inpatient mean (75)
Total mean (150)
Mean gain by setting
(Outpatient > Inpatient, p<0.001)
Mean completeness gain (Δ)
+5.84
+2.77
Outpatient (75)
Inpatient (75)

Fig. 3. Inpatient notes start more complete; structuring adds more to outpatient documentation (Δ +5.8 vs +2.8, p<0.001).

Table 2. Completeness outcomes (N = 150)

Outcome	Outpt	Inpt	Total
Mean original (/24)	14.3	18.6	**16.4**
Mean verified abstraction (/24)	20.1	21.4	**20.7**
Mean Δ	**+5.8**	**+2.8**	**+4.3**
95% CI of Δ	5.3–6.2	2.6–3.1	**4.0–4.7**
D9 Suicide-risk complete	0 → 65	75 → 75	**75 → 140**
D8 Substance-use complete	19 → 52	68 → 68	**87 → 120**
Trial-critical gaps (before → after)	—	—	**4.9 → 3.1**
Validated suicide-risk tool used	0%	0%	0%

Fig. 4 Domain-Level Completeness by Setting, Before vs After Structuring (N=150; McNemar)


D1 Primary dx
D2 Symptom severity (scale)
D3 Episode/illness duration
D4 Functional impairment
D5 Current medications
D6 Past med trials
D7 Adequacy prior trials
D8 Substance-use hx
D9 Suicide-risk assessment
D10 Mania/hypomania screen
D11 Psychosis screen
D12 Medical comorbidity
Outpatient - original
Outpatient - verified
Inpatient - original
Inpatient - verified
% of cases documented 'complete' (score 2)

D2 (validated severity scale) absent in all 150 notes - not manufactured. D9 suicide risk: outpatient 0% → 87% vs inpatient already 100%.

Fig. 4. Domain completeness by setting, before vs after structuring (McNemar).

Table 3. Safety & eligibility flags by setting (n per 75)

Flag	Out — present	Out — unclear	In — present	In — unclear
Bipolar-spectrum	11	1	12	0
Psychosis	6	0	32	43
Active substance	16	0	20	0
Benzo/sedative	8	0	18	0
Stimulant/controlled	42	0	8	0
Antipsychotic/mood-stab	19	0	46	0
Recent med change	20	38	0	75
Inadequate med-trial documentation	41	33	10	65
Acute SI	5	17	57	18
Recent self-harm	1	6	52	6
Unstable medical	12	0	18	0
Pregnancy/repro	3	17	0	29
Cognitive/consent	2	15	0	61
Nonadherence	12	1	12	0
Prior trial pattern	9	37	8	0
Other protocol	26	0	21	0

4 · Eligibility, Safety & Flags

Fig. 5 Preliminary Eligibility After Structured Screening - Outpatient vs Inpatient

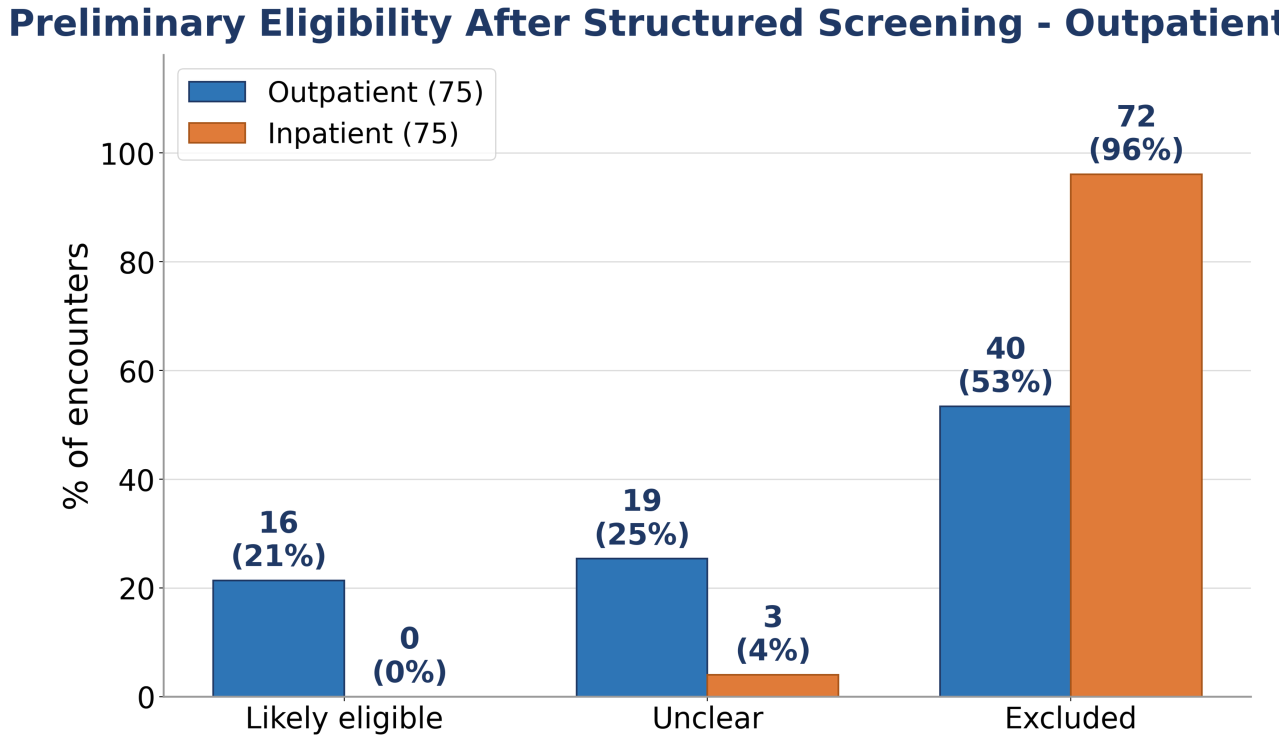

Outpatient (75)
Inpatient (75)
% of encounters
16 (21%)
0 (0%)
19 (25%)
3 (4%)
40 (53%)
72 (96%)
Likely eligible
Unclear
Excluded

Fig. 5. Preliminary eligibility after structured screening — outpatient vs inpatient.

Fig. 6 Protocol-Relevant Safety & Eligibility Flags by Setting (N=150)

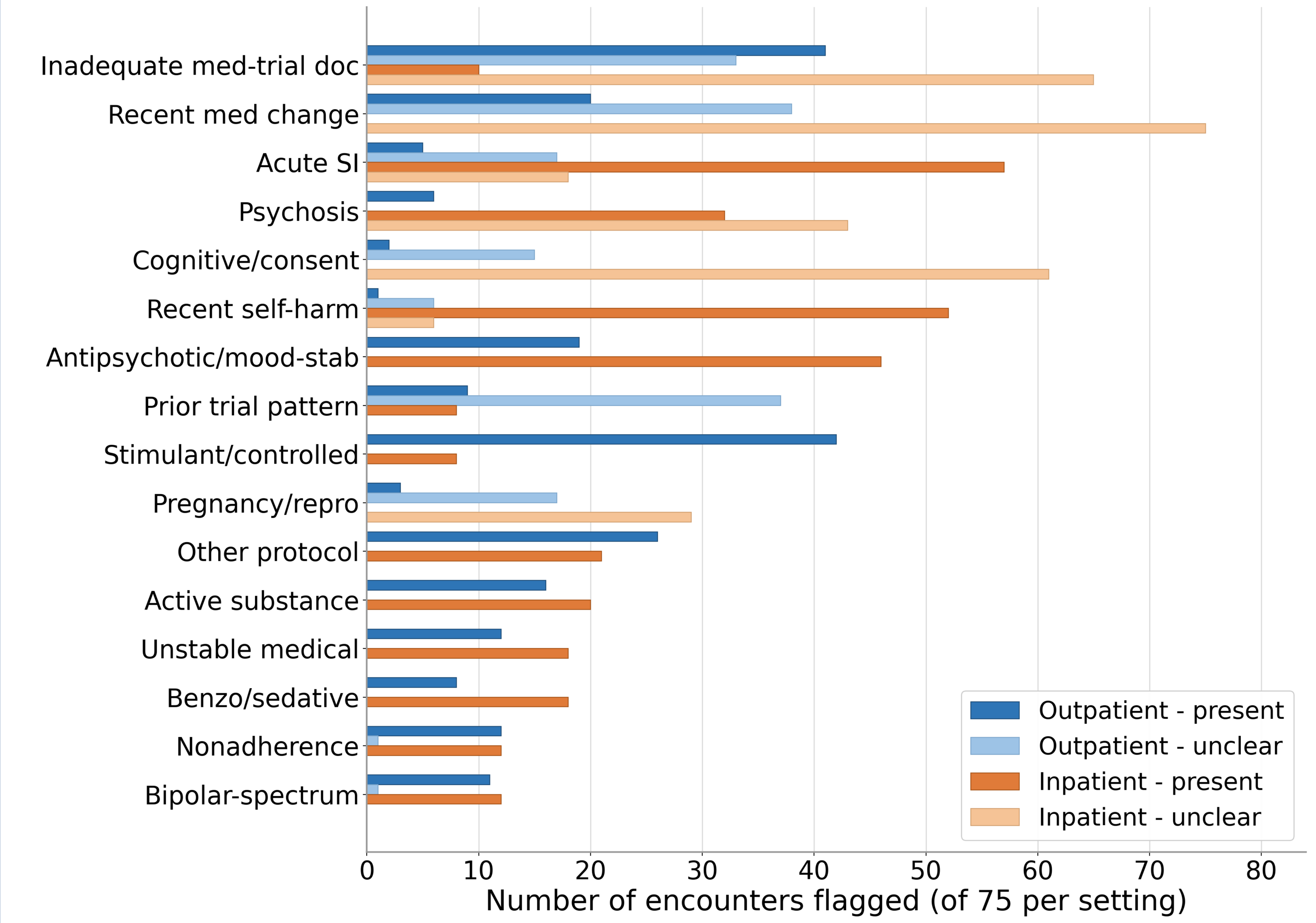

Inadequate med-trial doc
Recent med change
Acute SI
Psychosis
Cognitive/consent
Recent self-harm
Antipsychotic/mood-stab
Prior trial pattern
Stimulant/controlled
Pregnancy/repro
Other protocol
Active substance
Unstable medical
Benzo/sedative
Nonadherence
Bipolar-spectrum
Outpatient - present
Outpatient - unclear
Inpatient - present
Inpatient - unclear
Number of encounters flagged (of 75 per setting)

Fig. 6. Protocol-relevant flags surfaced by structured abstraction.

5 · Interpretation & Conclusions

Interpretation

- Structuring increased mean completeness by **+4.3 / 24** across 150 real encounters (p<0.001), **without imputing facts**.
- Inpatient admission notes were more complete at baseline (18.6 vs 14.3), so structuring added more value in outpatient notes (**+5.8 vs +2.8, p<0.001**), where documentation was sparsest.
- Suicide-risk documentation showed the clearest setting contrast: complete in **0% of outpatient** vs **100% of inpatient** notes at baseline; structuring raised outpatient completeness to **87%**.
- No notes used a validated suicide-risk tool, and none included a complete validated severity scale; structuring cannot manufacture an unmeasured scale (D2 unchanged).
- Acute inpatient admissions were **96% trial-ineligible** (active SI, psychosis, substance use, or pediatric age) vs a more mixed outpatient sample. The workflow helps de-risk screen failures and flags items requiring review.

Limitations

- Single-rater scoring (feasibility / effect-size design; reliability κ is the pre-registered next step, not an omission); investigator-estimated time metric; no prospective validated scales or toxicology; inpatient templates may differ across sites.

Conclusions

- Schema-constrained, clinician-verified abstraction **significantly improved documentation completeness** across 150 real outpatient and inpatient encounters.
- The method **structures and flags existing documentation; it does not invent clinical content.**
- Planned next steps: multi-site expansion, two-reviewer adjudication for κ, prospective C-SSRS and symptom-scale capture, and formal time-and-motion measurement.

Disclosures. No conflicts of interest to disclose. This is a methodology pilot, not a clinical efficacy evaluation. This structured-abstraction workflow is not a diagnostic device and does not replace clinician judgment. All data were de-identified prior to analysis.

Health Synergy Clinical Research Institute — West Palm Beach, Florida, USA

ISCTM Annual Meeting — Poster Session

Contact: Padder, T · Nisar, A