

Human-Validated Real-World Data: The Role of Expert Clinical Review in Ensuring Research-Grade Quality of Technology-Derived EHR Elements for Alzheimer’s Disease Studies

Daniel Gautieri (SiteRx, Vice President of Strategy & Data Services); Daniel Barulli (SiteRx, Director of Clinical Data Science)

A technology screen surfaced candidate patients at scale across ~40,000 records and cut screening volume by ~95-99%; a recall-oriented ML (TF-IDF) signal, not an eligibility predictor, provided efficiency while trained clinicians adjudicated every chart against protocol. Approval rose with the strength of the cognitive impairment phenotypic signal, but only expert review resolves the contextual calls that decide eligibility. RWE scale comes from technology; research-grade validity comes from expert clinical review.

INTRODUCTION	METHODS (cont.)	RESULTS	CONCLUSIONS	REFERENCES
Real-world evidence in Alzheimer's disease (AD) increasingly relies on technology-enabled EHR extraction. At network scale, only technology can surface candidates across millions of records. But methods leaning on structured fields alone (ICD codes, medications) have limited sensitivity and positive predictive value¹ for cognitive-impairment detection, and many cases go undetected in routine clinical care². Natural Language Processing (NLP) and AI improve on structured-only methods³, yet AD documentation poses problems they cannot resolve on their own: • Cognitive complaints overlap with psychiatric and neurological comorbidities • Key context often exists only in unstructured narrative • Differential diagnoses are often coded identically to confirmed conditions Without systematic human adjudication, technology-derived datasets risk embedding misclassification, especially in de-identified observational designs where no investigator verifies eligibility.	A two-layer review framework 1. EHR ingestion (agnostic). Continuous de-identified records; up to an 18-month look-back per patient. 2. Pre-qualification (technology). Two functions: it rank-orders candidates by phenotypic signal for cognitive impairment (CI) above a ≥ 0.5 threshold (a recall-oriented signal, not an eligibility predictor, cutting review burden ~95-99%), and it removes patients who clearly fall outside the study inclusion/exclusion criteria using structured features (medications, ICD codes). 3. HARE adjudication (human). Trained reviewers adjudicate technology outputs for each chart in a controlled de-identified environment: every gating question answered, source text highlighted, and traced to a single reviewer. 4. Validated dataset (output). Auditable, versioned eligibility decisions with lineage to the source encounter. Four gating questions per chart: • Was there a cognitive impairment complaint on the visit date? • Does the patient have a prior alternative diagnosis? • Have they received a prior AD-class medication? • Do they have a prior amyloid/tau biomarker test?	Pre-qualification - signaling. A broad-recall CI screen rank-orders patients by phenotypic signal and reduces review volume by ~half (36,961 → 19,959, -46%). The technology ranks candidates; it does not confirm eligibility. Pre-qualification - filtering (Figure 1). Structured features (medications, ICD conditions) automatically remove patients clearly outside the study I/E criteria: a cumulative -7,498 (-38%), with no human review required. CI-eligible after meds after ICD 19,959 17,299 12,461 -2,660 (13%) -4,838 (28%) Figure 1. Structured exclusions (medications, ICD). Technology removed a cumulative 7,498 (38%) of CI-eligible reviews automatically (n = 19,959). Meds = AD-related medications (e.g., acetylcholinesterase inhibitors, NMDA receptor antagonists). ICD = G30 and codes for other etiologies that can explain CI. HARE adjudication (Figure 2). The technology screen is a broad, representative recall of CI; human review is the essential second layer, resolving the unstructured-narrative complexity and ambiguity only expert adjudication can settle. The narrative CI-documentation call drives ~72% of non-qualification (each structured exclusion $\leq 10\%$). No qualifying CI Brain lesion / alt dx Prior AD drug Prior biomarker test Prior referral Serious illness 72% 10% 6% 4% 3% 1% Figure 2. Reason disqualified. Share of non-qualifiers (n ≈ 6,800); criteria may co-occur. The contextual CI-documentation judgment, not a coded exclusion, is the leading gate. HARE adjudication - concordance (Figure 3). Higher-signal bands are approved at higher rates (0.9-1.0: 53% vs 0.5-0.6: 12%) - the phenotypic signal rank-orders eligibility, though only review confirms it. 0.9-1.0 0.8-0.9 0.7-0.8 0.6-0.7 0.5-0.6 53% 37% 24% 18% 12% Figure 3. Approval by human review, banded by phenotypic signal. Higher CI signal → higher approval (review-level, n ≈ 20K).	Generating research-grade real-world evidence for Alzheimer's disease calls for a multi-tiered process, not a single tool: technology filtering for structured, rule-based exclusions; ML/model-driven signaling for cognitive impairment phenotypes; and expert human review for the contextual judgments that decide eligibility. Different inclusion and exclusion criteria demand different modalities, and matching each criterion to the appropriate layer is what makes RWE generation both scalable and accurate. EHR data is among the richest data substrates, but no single layer suffices to utilize it effectively. Structured fields do not capture cognitive impairment, which lives almost entirely in unstructured narrative; structured-only screening would retain large numbers of patients who do not qualify, biasing the cohort and undermining the power and validity of downstream analyses. Yet relying on narrative without human adjudication drives excessive disqualifications which lack full clinical context. Only an integrated, layered approach yields validated RWD that reflects clinical reality, critical for AD studies where data fitness determines evidence credibility. Abbreviations. CI = cognitive impairment; RWD = real-world data; AD = Alzheimer's disease; HARE = Human Assisted Review Engine. Data & methods. SiteRx analytics snapshot Jun 2026; n = 9,485 completed reviews; phenotypic signal ≥ 0.5; 18-month look-back. Approval-by-band is review-level (n = 20,390).	1. Bhattacharyya, J., Barnes, L. L., Chen, Y., Gianattasio, K. Z., Grodstein, F., James, B. D., ... & Power, M. C. (2025). Evaluating linked ICD-10 Medicare claims data as a method of dementia case ascertainment in research settings. <i>Alzheimer's & Dementia</i>, 21(5), e70200. 2. Lang, L., Clifford, A., Wei, L., Zhang, D., Leung, D., Augustine, G., ... & Chen, R. (2017). Prevalence and determinants of undetected dementia in the community: a systematic literature review and a meta-analysis. <i>BMJ open</i>, 7(2), e011146. 3. Shankar, R., Bundele, A., & Mukhopadhyay, A. (2025). Natural language processing of electronic health records for early detection of cognitive decline: a systematic review. <i>NPJ digital medicine</i>, 8(1), 133.
METHODS	METHODS (cont.)			CONTACT
SiteRx operationalizes real-world evidence studies on de-identified longitudinal EHR from a large U.S. community provider network. Records are de-identified at the source and analyzed centrally, so studies run without any investigator contacting or re-identifying patients. Within this environment, technology pre-qualifies candidates at scale and HARE (Human Assisted Review Engine) supplies the human adjudication, with full audit trails throughout.	Results reflect patients with cognitive impairment and no prior AD diagnosis, alternative etiology for the impairment, or AD-related treatment on record, screened above a ≥ 0.5 propensity threshold on CI filtering.			 Daniel Gautieri, SiteRx daniel@siterx.com 101 6th Ave, Floor 10, New York, NY 10013  About SiteRx SiteRx is on a mission to democratize access to clinical research as a treatment alternative, advancing healthcare and health outcomes for patients and their loved ones today. Learn more by scanning the QR code.