xCures®

Validated Clinical Checklists for Scalable LLM-Based EHR Data Extraction

Timothy J. Stuhlmiller, AJ Rabe, Kristi Lui, William Mahoney, Glenn K. Kramer, Mika Newton, Kenny K. Wong

Abstract

Healthcare organizations increasingly rely on granular, patient-level clinical data to support treatment decisions, eligibility verification, and clinical operations workflows. However, critical information remains distributed across heterogeneous medical records, within unstructured narrative text, and represented in embedded unstructured image files. xCures has developed validated configurable clinical Checklists, a scalable AI-enabled extraction framework that uses targeted LLM-based Q&A prompts integrated with a semantic document processing pipeline. Checklist items retrieve clinically meaningful data elements from both structured and unstructured sources, synthesize evidence across multiple documents, and return discrete, workflow-ready outputs. Each checklist item is designed for traceability, clinical interpretability, and measurable performance, supported by human validation and continuous automated quality control (QC). This approach extends traditional schema-based extraction by enabling fully customizable, domain-specific clinical data capture at scale.

Introduction

Clinical information extraction from unstructured electronic health records (EHRs) has been an active area of research for more than a decade. Systematic reviews of clinical natural language processing (NLP) highlight both the promise of automated extraction and the persistent challenges posed by heterogeneous documentation styles, variable terminology, and incomplete context (Fraile Navarro et al., 2023). Early benchmark efforts, such as the i2b2 medication extraction challenge, demonstrated that high accuracy is achievable but requires careful task definition and validation (Patrick & Li, 2010). More recent transformer-based architectures have further improved contextual understanding of clinical text (Chen et al., 2023). At the same time, large language models (LLMs) have emerged as powerful few-shot clinical information extractors when appropriately constrained and validated (Agrawal et al., 2022).

A central challenge in healthcare extraction is not merely that information is unstructured, but that it is often inconsistent across the record. A diagnosis date, biomarker result, or treatment decision may appear in multiple notes with conflicting values due to transcription errors, templated copy-forward behavior, delayed documentation, or clinical reassessment. This conflict creates a real-world problem: extraction systems must not only find information, but must identify the contextually preferred and clinically correct representation of that information.

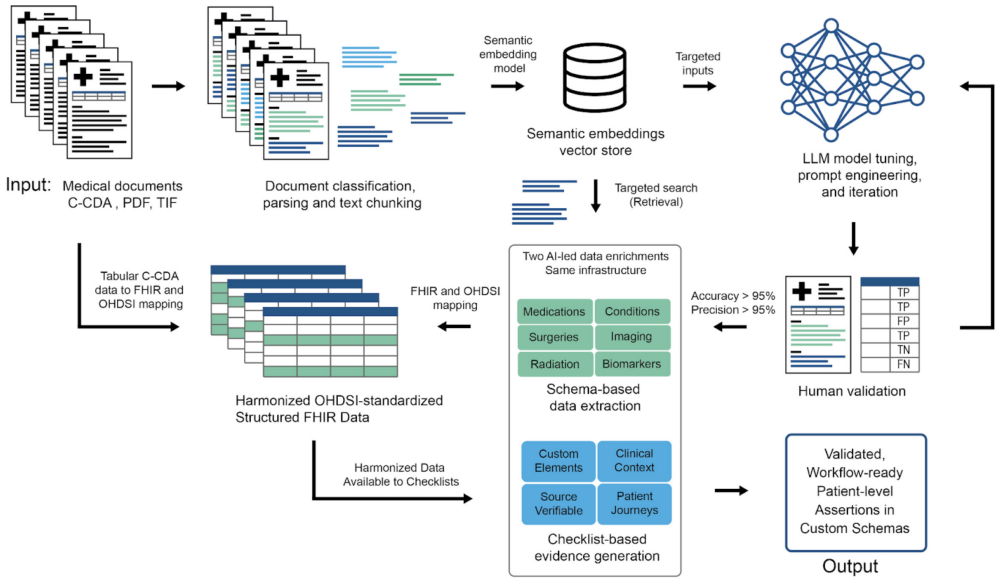
Despite advances in NLP, real-world clinical data processing remains error-prone. Meta-analyses of clinical research workflows show non-trivial error rates across manual, semi-automated, and automated abstraction methods, underscoring the need for validated and quality-controlled extraction systems (Garza et al., 2025). Generative systems also introduce hallucination risk, reinforcing the need for evidence-constrained extraction and traceable outputs (Lee et al., 2019; Ye et al., 2023).

To address these requirements, xCures developed xCures Checklists, designed to operationalize semantic retrieval, targeted LLM extraction, and rigorous patient-level validation into a scalable production platform.

xCures Medical Record Processing Overview

The xCures platform ingests diverse patient medical records (HIE/QHIN, C-CDA, PDF/TIFF), processes them via OCR, parsing, and text chunking, and embeds them into a semantic vector space. These document embeddings enable cost-effective, precise retrieval-augmented generation (RAG) by semantically searching for the most relevant documents before LLM extraction, consistent with modern architectures. Structured data (conditions, medications, etc.) are parsed from C-CDA and enhanced by Schema-based LLMs, which extract domain-specific details (e.g., cancer stage, genomic results) from free-text and OCR'd clinical narratives. This enriched data is then harmonized, standardized using OHDSI vocabularies (SNOMED, LOINC, etc.), and unified in a single data model (Stuhlmiller et al. 2025).

xCures Platform - Technical Components



xCures®

What is a Checklist?

A Checklist is a series of pre-configured LLM prompts that answer a set of questions for an individual reviewing medical records, automatically identifying the relevant documents and presenting the results with document provenance. Each constituent prompt in this sequence is termed a Checklist Item, and may incorporate one or more governing Properties, determined by the specific requirements for resolving the underlying clinical questions.

As an example, a checklist that seeks to define a cancer patient's disease may include Checklist Items to extract the Initial Diagnosis Details, the Current Diagnosis Details, and the Tumor Characteristics, with each item dictating a series of properties to extract (Initial Diagnosis Date, Initial Diagnosis, Initial Histology, Initial Stage, Initial T, N, and M stage, and so on). The end user will call the checklist endpoint in the UI or via a dedicated API, and typically within 30 seconds will receive the full extracted results of each checklist item property in a pre-defined JSON schema with provenance to the documents referenced.

Checklist Design and Behavior

A defining characteristic of the checklist framework is the use of detailed, prescriptive instructions detailing the extraction or summarization activity, conflict handling rules, guidance on document types to include or exclude, relevant date ranges, and a listing of property definitions and data types. Checklist instructions enforce objective, evidence-based criteria for determinations. Conditions may only be marked as present when explicit documentation or quantitative thresholds are met, improving consistency and reducing ambiguity. This approach mitigates the risk of hallucinated or unsupported outputs, which have been documented across neural generation systems and LLMs (Lee et al., 2019; Ye et al., 2023). The platform supports multiple model classes, enabling the selection of efficient models for high-throughput extraction tasks and more advanced reasoning-oriented models for complex longitudinal synthesis.

Source-verifiable Extraction

Each checklist item retrieves specific clinical facts or determinations, capable of interrogating both structured clinical data and unstructured document text. Checklists efficiently synthesize evidence across multiple documents, producing patient-level determinations supported by longitudinal documentation rather than interpreting a single isolated document. Outputs are returned as discrete data elements, such as dates, numeric values, categorical classifications, or boolean indicators, designed for immediate use in downstream workflows. Conventional generative AI is also configurable to output endpoints as natural language summaries.
Every checklist output includes links to source documentation, enabling rapid validation and audit. Each checklist item also includes a concise, natural-language justification, typically two to three sentences, summarizing the supporting evidence and identifying the type of source document (e.g., "Surgical Pathology Report" or "Cardiology Consultation Note"). This design supports explainability while maintaining standardized outputs appropriate for large-scale processes.

Checklist items are not generic schema fields; they are designed to answer specific clinical questions supporting defined workflows. Their flexibility is a core design principle: checklist items can be customized for each client, each project, and each medical domain without requiring changes to the underlying data model.

Table 1: Checklist Features

Customizable	Queries any element within structured data or free text of a document
Intelligent search	Relevant documents are automatically identified by semantic content
Time relevancy	Enforceable document date limit (last 6 months, e.g.)
Document selection	Limit to only specific documents (oncology notes, cardiology visits, e.g.)
Behavior	Q&A, Verbatim extraction, Inference, Reasoning, Summarization
Provenance	Links to source documents
Justification	Natural language summary of evidence and document types referenced
Validation	Extraction is optimized and validated on a patient level
Run Cadence	Can be run at any time by the client
Client Data Access	Dedicated tab in UI and API

Comprehensive Patient-Level Data Extraction

Checklist items successfully extract comprehensive patient-level clinical information across clinical domains, supporting diverse use cases in pre-operative surgical risk stratification, oncology molecular diagnostics, women's health screening and genetic testing, transplant evaluation, evaluation of medical necessity, and compliance with healthcare quality measures such as HEDIS.

Checklists efficiently return binary determinations of complex comorbidities (e.g., severe frailty, chronic congestive heart failure, uncontrolled Type 1 diabetes, or untreated obstructive sleep apnea), extracting the relevant supporting quantitative evidence such as BMI or hemoglobin A1c values. Checklists extract detailed surgical oncology pathology results (e.g., T/N/M stage, tumor size, specimen details), and key molecular and hormone receptor markers (e.g., ER/PR/HER2, IDH, MGMT status). Checklists verify clinical actions (e.g., counseling, chemotherapy discussions, clinical reasoning for test ordering) for compliance and quality. They support value-based care by automating HEDIS measure assertions (qualification and measure satisfaction), extracting diagnoses, dates, and procedure results with source documentation. Automated checks can be scheduled monthly or quarterly to identify care gaps.

Checklist items are executed in parallel, enabling dozens of operations to run simultaneously. Results are automatically generated and formatted according to a precise, preconfigured custom schema, ready for immediate ingestion into client workflows. Checklist properties can be arrayed, such that multiple events or clinical endpoints and metadata can be resulted in series. A single checklist item typically returns in <10 seconds (for boolean assertions or simple extraction procedures), but can take up to 1 minute for complex arrayed extraction tasks (extracting all medications from a recent visit as medication name, dose, formulation, route of administration, patient instructions, start date, end date) or reasoning-based workflows (HEDIS measures that rely on hospitalization follow-up reporting).

Patient medical histories typically comprise hundreds of documents. To address a single clinical checklist item, the system first narrows this down to dozens of documents based on broad semantic relevance. This selection is further refined based on relevance to the specific answer, resulting in a median of only four cited documents per checklist item. This process effectively streamlines the overall medical record by automatically identifying the most relevant documents and extracting the required information, moving from hundreds to dozens to fewer than 10 documents to inform a single checklist item.

Table 2: Examples of Checklist Items and Properties

Domain	Checklist Item	Item Summary	Properties
Comorbidity Assessment	Active anemia	Two most recent Hgb measurements < 10 g/dL.	Present/Absent (boolean), Justification
	Chest pain in last 3 months	Documented, contemporaneous, cardiac-attributed chest pain/anginal symptoms within the prior 3 months	
	Mild pulmonary hypertension	Documented mild pulmonary hypertension (e.g., mean PAP 21-30 mmHg on RHC, or estimated PASP 36-45 mmHg on echo).	
Oncology	Initial Cancer Diagnosis	Extract structured data elements corresponding to the initial, complete, and active malignant cancer diagnosis, prioritizing pathology and excluding historical or irrelevant cancers.	Cancer Type, Initial Diagnosis, Histology, Date, Stage, Grade, T, N, M, Evidence, Justification
	Cancer Pathology Results	Exhaustively extract all cancer-related results from surgical pathology reports, returning an array of properties for each specimen, strictly avoiding information from other clinical sections.	Diagnosis, Histology, Specimen Collection Date, Specimen ID, Anatomic Site, Facility Name, Justification
	Molecular Testing	Extract and classify all cancer-related molecular testing information (IHC, molecular assays, NGS, e.g.) and summarize key, actionable, or pathogenic biomarker results.	Any molecular testing, Limited-panel testing, Tissue-based CGP, Liquid biopsy NGS, Molecular results summary, Justification

xCures®

Domain	Checklist Item	Item Summary	Properties
Telehealth	Family History	Extract every family history condition (diseases, disorders, mutations, traits) and affected family members in an array, preserving sibling distinctions.	Condition, Relationship, Justification
	Recent Hospitalizations	Extract an array of structured inpatient admission data (within 2 years), strictly excluding observation/outpatient encounters.	Admission/Discharge Dates, Reason, Discharge Disposition, Hospital name, Justification
	SDOH	Extract a complete, structured, and justified summary of Social Determinants of Health (SDOH) from the electronic medical record, prioritizing explicit documentation and standardized screening scores.	Marital Status, Living Situation, Education Level, Employment Status, Functional Barriers, Health Literacy, Transportation Access, Financial Insecurity, Housing Insecurity, Food Insecurity, Utility Insecurity, Insurance Status, Mental Health, Justification
HEDIS	Breast Cancer Screening (BCS-E)	Assess patient eligibility and exclusion for the BCS-E HEDIS measure. If qualified, evaluate measure satisfaction based on a mammogram between 2024-10-01 and 2026-12-31.	Patient Qualifies, Measure Satisfied, Justification
	Blood Pressure Control for Patients with Hypertension (BPC-E)	Assess BPC-E HEDIS qualification (adults 18-85 with hypertension by 12/31/2026; excluding ESRD, pregnancy, hospice, death) and measure satisfaction based on most recent 2026 BP reading <140/90 mmHg.	
	Kidney Health Evaluation for Patients With Diabetes (KED)	Assess KED HEDIS qualification (age 18–75, type 1 or 2 diabetes diagnosis in 2025–2026, and exclusions) and measure satisfaction based on documentation of a numeric eGFR (or serum creatinine) and a numeric uACR (or urine albumin/creatinine ratio) with 2026 collection dates.	

Validation Framework

Clinical-grade extraction requires measurable performance and ongoing monitoring. Validation is therefore a core design element of xCures Checklists, consistent with schema-based LLM extraction methodologies described recently (Stuhlmiller et al., 2025).
Patient Sampling and Initial Extraction
A randomized set of patient records is selected. The configured checklist prompt is applied to these patients, yielding structured outputs with provenance links and justifications for each extracted item.

Clinician Review of Extracted Evidence (TP/FP Identification)
Human clinical reviewers evaluate each checklist response. Reviewers first examine the cited documents to confirm the extracted value is supported by evidence, classifying outputs as true positives (TP) or false positives (FP) based on the cited documentation.

Independent Semantic Search Across the Entire Record (FP/FN Confirmation)
To ensure clinical correctness at the patient level, reviewers perform an independent semantic search across the entire record, independent of the checklist results. This addresses potential conflicts or outdated information in the documentation. Patient-level validation ensures the output reflects the clinically preferred and most reliable evidence, as dictated by the prompt's prioritization rules.

If the reviewer finds more authoritative evidence contradicting the output, it is a false positive (FP). If relevant evidence was missed, it is a false negative (FN). If no evidence exists and the checklist correctly extracts a null field, it is a true negative (TN). This ensures checklist performance aligns with real healthcare use cases, focusing on accurate clinical interpretation across the entire record.

Edge Case Capture and Document Attribution
When errors occur, reviewers record the accurate result and the document(s) containing the correct information. This feedback is used to refine semantic search parameters, prompt logic, and/or document prioritization rules, systematically capturing edge cases and improving robustness.

Prompt Iteration and Successive Validation Rounds
Prompts are iteratively refined based on reviewer feedback through manual and LLM-assisted optimization. Successive validation rounds are performed until performance meets deployment thresholds, typically ≥95% PPV (precision) and ≥95% sensitivity (recall).

Checklist Report Cards and Metric Reporting
Validation results are compiled into standardized performance summaries ("checklist report cards") that quantify performance at the property, item, and checklist level using Accuracy, PPV (Precision), Sensitivity (Recall), and F1 Score. Articulated limitations or strict rules built into the prompts are included to inform end consumers of necessary considerations for appropriate implementation, providing transparent performance evidence.

xCures®

Ongoing QC and Random Sampling

Post-deployment, the xCures platform continuously monitors checklist performance using automated quality control and random sampling with clinical re-review of client data. This ongoing surveillance identifies performance drift and ensures the checklist maintains high PPV and sensitivity over time.

Table 3. Performance metrics for select checklists

Checklist	Comorbidity Assessment (boolean for discrete conditions)	Cancer Diagnosis and Tumor Characteristics	Specific Biomarker Status (wild-type, mutant, unknown)	Clinical Rationale for Molecular Testing	2026 HEDIS Measures
Total elements reviewed	6399	1881	400	157	7199
True Positive	3889	1724	393	108	6739
True Negative	2494	123	0	40	0
False Positive	14	23	5	5	356
False Negative	2	11	2	4	104
Accuracy	0.997	0.982	0.983	0.943	0.936
Precision (PPV)	0.996	0.987	0.987	0.956	0.950
Recall (Sensitivity)	0.999	0.994	0.995	0.964	0.985
F1	0.998	0.990	0.991	0.960	0.967
Median total documents	413	209.5	273.5	298	51
IQR total documents	189-630.5	106-425	125-503	161-601	20-108
Median documents referenced by checklist	4	4	4	4	5

True Positive (TP): Extracted value is supported by cited documentation and is the preferred/most reliable evidence from the entire patient record.

True Negative (TN): Checklist cozrrectly extracted a null field (e.g., 'Unknown', 'Not available') because the desired data element is absent from the record.

False Positive (FP): Extracted value lacks cited support or is contradicted by more authoritative evidence in the record.

False Negative (FN): Checklist failed to capture existing, relevant evidence, even if a null field cited plausible documentation.

xCures®

Conclusion

xCures Checklists provide a scalable, domain-specific solution for extracting clinically meaningful patient-level data from heterogeneous medical records. By combining semantic retrieval, prescriptive LLM-based extraction, rigorous patient-level validation, and full provenance, the checklist framework enables healthcare and life-science organizations to operationalize AI-driven extraction with confidence.

This approach directly addresses the realities of healthcare documentation, where critical information is unstructured, inconsistently recorded, and often contradictory, by ensuring that checklist outputs are not only traceable but also clinically correct in a full longitudinal context. Supported by iterative validation, measurable performance reporting, and automated QC monitoring, xCures checklists deliver customizable clinical insights at production scale.

References

1. Fraile Navarro D, Ijaz K, Rezazadegan D, Rahimi-Ardabili H, Dras M, Coiera E, Berkovsky S. Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review. Int J Med Inform. 2023;177:105122. doi:10.1016/j.ijmedinf.2023.105122.
2. Patrick J, Li M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. J Am Med Inform Assoc. 2010;17(5):524–527. doi:10.1136/jamia.2010.003939.
3. Chen A, Yu Z, Yang X, Guo Y, Bian J, Wu Y. Contextualized medication information extraction using Transformer-based deep learning architectures. J Biomed Inform. 2023;142:104370. doi:10.1016/j.jbi.2023.104370.
4. HL7 International. HL7 FHIR Release 4 Resource List. Accessed February 24, 2025. http://hl7.org/fhir/R4/resourcelist.html
5. OpenAI. New embedding models and API updates. Accessed February 21, 2025. https://openai.com/index/new-embedding-models-and-api-updates/
6. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large language models are few-shot clinical information extractors. In: Proceedings of EMNLP 2022. 2022:1998–2022. doi:10.18653/v1/2022.emnlp-main.130.
7. Lee K, Firat O, Agarwal A, Fannjiang C, Sussillo D. Hallucinations in Neural Machine Translation. ICLR. 2019.
8. Ye H, Liu T, Zhang A, Hua W, Ji W. Cognitive Mirage: A Review of Hallucinations in Large Language Models. arXiv. 2023. doi:10.48550/arXiv.2309.06794.
9. HL7 International / Clinical Interoperability Council. Minimal Common Oncology Data Elements (mCODE) Implementation Guide. Accessed February 24, 2025. https://build.fhir.org/ig/HL7/fhir-mCODE-ig/
10. Observational Health Data Sciences and Informatics (OHDSI). Standardized Vocabularies. Accessed February 24, 2025. https://github.com/OHDSI/Vocabulary-v5.0/wiki/Standardized-Vocabularies
11. IMO Health. Healthcare Data Standardization | IMO Precision Normalize. Accessed February 24, 2025. https://www.imohealth.com/imo-precision-normalize/
12. Garza MY, Williams T, Ounpraseuth S, Hu Z, Lee J, Snowden J, Walden AC, Simon AE, Devlin LA, Young LW, Zozus MN. Error rates of data processing methods in clinical research: A systematic review and meta-analysis. Int J Med Inform. 2025;195:105749. doi:10.1016/j.ijmedinf.2024.105749.
13. Stuhlmiller TJ, Rabe AJ, Rapp J, Manasco P, Awawda A, Kouser H, Salamon H, Chuyka D, Mahoney W, Wong KK, Kramer GA, Shapiro MA. A scalable method for validated data extraction from electronic health records with large language models. medRxiv. 2025. doi:10.1101/2025.02.25.25322898.