TEMPORALLY GROUNDED AND EVIDENCE-LINKED CLINICAL ABSTRACTION FROM SCANNED LONGITUDINAL RECORDS IN HEART TRANSPLANT RECIPIENTS

Meng Ma, Austin Tao, Curtis Focht, Jing Huang, Jeffrey Teuteberg, Jeremy Kobulnik, Li Li
CareDx, Brisbane, CA, USA
Corresponding author: lli@caredx.com, jkobulnik@caredx.com

Clinical information relevant to heart-transplant surveillance is frequently dispersed across longitudinal, image-based PDF medical records and must be interpreted relative to a patient-specific assessment date. We developed an end-to-end integrated AI system that combines page-preserving document parsing, large language model (LLM)-based clinical information extraction and temporal interpretation, index-date–conditioned evidence selection, and source-page evidence linking. The fixed AI system was evaluated in 157 validation cases from 141 patients, comprising 5,033 pages from 52 U.S. transplant centers. Outputs for echocardiogram date, left ventricular function, patient-reported symptoms, and physical-examination signs were compared with EdiPhy (MedManagement, LLC), an external professional medical-review service (EPS), using masked and randomized physician adjudication under prespecified clinical and temporal criteria. In a secondary pooled summary across 628 paired case-domain comparisons, the AI system produced 626 correct abstractions (99.7%; 95% CI, 98.8%–99.9%), compared with 518 (82.5%; 95% CI, 79.3%–85.3%) from the EPS; the paired difference was 17.2 percentage points (95% CI, 14.2–20.2; exact McNemar P<.001). Temporal selection was correct in 155 of 157 echocardiogram cases (98.7%) and in all symptom and examination cases. Page-linked evidence was present in 390 outputs, comprising 1,035 page-reference occurrences; 388 of the 390 page-linked abstractions (99.5%) were clinically correct. The two AI errors were echocardiogram-date errors associated with incomplete or incorrect source-text parsing. These findings demonstrate strong internal validation performance for index-date–conditioned, provenance-preserving abstraction from scanned longitudinal records, while external validation, multi-adjudicator assessment, and prospective workflow evaluation remain necessary before clinical deployment.

Keywords**:** clinical natural language processing; large language models; temporal reasoning; document parsing; heart transplantation; longitudinal medical records

1. Introduction

Heart transplant recipients require lifelong follow-up to identify changes in allograft function, rejection risk, and post-transplant complications. Clinical assessment integrates complementary information from echocardiography, ventricular function, patient-reported symptoms, physical examination, laboratory findings, and longitudinal context. No single data element necessarily represents the patient's condition; rather, the interpretation depends on how evidence aligns with the clinical assessment under review [1-3].

The same longitudinal evidence is also central to medical necessity assessments for healthcare institutions, payers, and service providers. Assessment of medical necessity is complex, time-consuming, and resource intensive for health plans and healthcare provider organizations [19,20]. Determining whether a procedure, diagnostic test, medication, or surveillance activity is clinically justified requires reviewers to verify that relevant clinical facts were present, appropriately documented, and applicable to the assessment being reviewed. In heart-transplant care, this may

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company anddistributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

include confirming recent echocardiographic findings, left ventricular function, patient-reported symptoms, physical-examination signs, and other contextual evidence to properly adjudicate the clinical need for a service. Thus, medical necessity review depends not only on identifying relevant clinical facts, but also on determining whether those facts were applicable to the clinical context at the time of assessment.

This abstraction problem is therefore inherently temporal [4]. Longitudinal records may contain remote episodes of volume overload, historical ventricular dysfunction, chronic dyspnea, resolved edema, planned studies, and repeated statements copied forward across encounters. A finding may have different implications depending on whether it is current, new or worsening, chronic, resolved, or historical. The task is therefore not merely to recognize a clinical concept, but to determine whether a specific occurrence is relevant to a patient-specific assessment date of interest (hereafter, index date).

Fax-based exchange of external clinical information further compounds this challenge. In this study, we received image-based PDF record packages from multiple transplant centers, including clinic notes, progress notes, echocardiogram reports, laboratory documentation, procedure reports, inpatient notes, and discharge summaries. Formatting, page organization, image quality, and documentation conventions varied substantially. Therefore, abstraction from these records requires recovery of machine-readable text while preserving page boundaries and source provenance.

Clinical NLP has long addressed extraction from unstructured text [5], while LLMs have increasingly been studied across medicine and biomedicine [6,9], including recent efforts to convert report-level clinical narratives into structured variables through prompt engineering or hybrid rule-based–LLM pipelines [21,22]. These systems primarily evaluate extraction from bounded text documents, whereas longitudinal abstraction must determine which of multiple dated occurrences is applicable to a specified index date; recent work indicates that temporal reasoning across patient timelines remains challenging for LLMs [4,23]. Related work on verifiable longitudinal chart summarization has also emphasized linking generated content back to source notes for human review [24]. Accordingly, the contribution here is system-level rather than a new foundation model: page-preserving parsing, constrained LLM extraction and classification, deterministic index-date–conditioned temporal selection, and source-page evidence linking are integrated and fixed before validation for heterogeneous scanned record packages.

For this project, we needed to determine whether patients had stable function of the transplanted heart on the index date, as defined by both clinical and echocardiographic parameters. To do this, we developed a four-component framework comprising (1) page-preserving document parsing and text extraction; (2) LLM-based clinical information extraction and temporal interpretation; (3) index-date–conditioned evidence selection; and (4) evidence-linked structured abstraction for human review. We evaluated the fixed AI system in an independent validation cohort and compared it with an external professional abstraction service using masked physician adjudication. The primary goal was to create an index-date–conditioned abstraction method that retains page-level provenance while operating on heterogeneous scanned longitudinal records.

2. Materials and Methods

2.1. Design and Cohort Construction

We conducted a retrospective internal validation study of an AI system for clinical abstraction from longitudinal medical records of adult heart transplant recipients. Records were associated with post-transplant assessment index dates between December 30, 2024, and January 30, 2026. The unit of analysis was a validation case consisting of a patient-specific medical-record package paired with its corresponding index date.

A development cohort of 40 cases was used for prompt development, workflow refinement, and specification of clinical and temporal rules. These cases were excluded from validation. After development, the prompts, classification criteria, and inference workflow were fixed. The independent validation cohort comprised 157 randomly selected validation cases from 141 patients; 14 patients contributed more than one validation case. No cases overlapped between the development and validation cohorts.

2.2. End-to-End Workflow

The AI system accepted a faxed PDF medical-record package and a patient-specific index date of interest. It produced structured abstractions for four prespecified domains—echocardiogram date, left ventricular function, heart-failure–related patient-reported symptoms, and heart-failure–related physical-examination signs—together with supporting text and source-page identifiers. The final report was designed to support retrospective administrative chart review through human verification rather than autonomous clinical decision-making. This study evaluated abstraction accuracy only; reviewer efficiency, downstream administrative decisions, and bedside clinical decision-making were outside the study scope. The complete workflow is summarized in Figure 1.

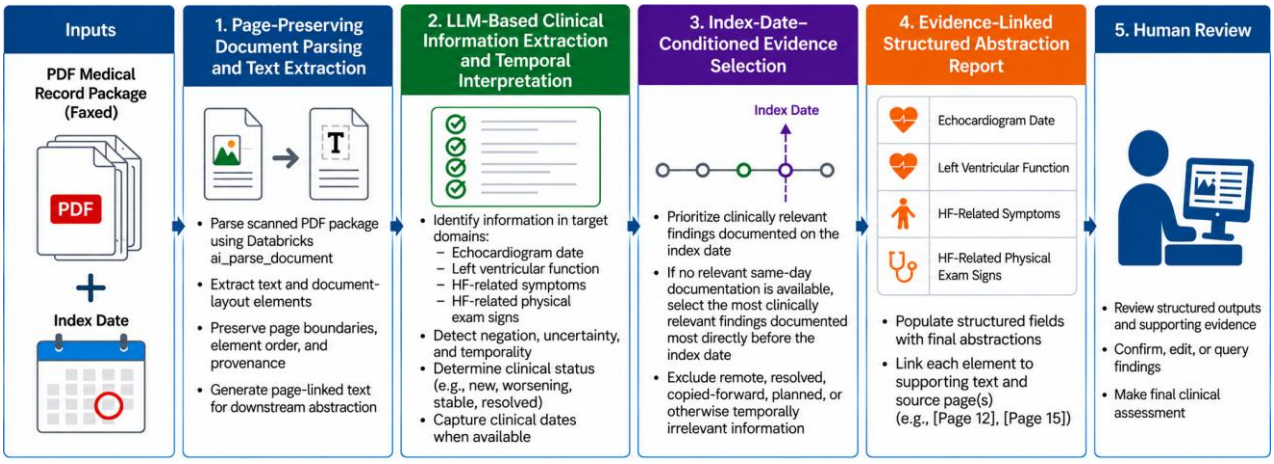


Figure 1. End-to-end workflow for temporally grounded and evidence-linked clinical abstraction. A faxed PDF medical-record package and index date were converted to page-linked text; the LLM extracted structured candidates and temporal metadata; deterministic logic and branch-specific prompts applied index-date–conditioned selection and classification; and outputs were linked to supporting text and source pages for human review.

2.3. Page-Preserving Document Parsing and Text Extraction

We processed all PDF packages using Databricks ai_parse_document [7], a document-parsing function that extracts text and layout elements from scanned and mixed-content documents. The function returned discrete elements, including body text, section headers, titles, tables, page headers and footers, page numbers, captions, figures, and footnotes. Each element retained its type and source-page information. We assembled the extracted elements in the original page order and normalized them into a page-linked textual representation for downstream abstraction. We assessed package-level parsing completeness for the target abstraction tasks rather than at the character or word level. We classified a package as incomplete when review of the original PDF and parsed representation identified omitted or mis-transcribed source content relevant to one or more target domains.

2.4. Prespecified Clinical Abstraction Definitions and Classification Criteria

Before prompt development and validation, we prespecified operational definitions and classification criteria for four clinical abstraction domains: echocardiogram studies, left ventricular function, patient-reported symptoms, and physical-examination signs. These specifications defined the eligible source information, allowable output categories, clinical thresholds, and precedence rules for abstraction. We applied the same criteria to the AI system, the external professional abstraction service, and physician adjudication to ensure that both methods were evaluated against a common reference framework.

2.4.1. Echocardiogram Studies

Qualifying studies included resting transthoracic echocardiograms, limited studies, transesophageal echocardiograms and stress echocardiograms. The system extracted study date, study type, procedural status, temporal relationship to the index date, supporting text, and source page. Completed studies were distinguished from ordered, scheduled, planned, administrative, and comparison-reference dates.

2.4.2. Left Ventricular Function

Left ventricular function was classified as Normal, Abnormal, or Unknown using the selected echocardiogram. Normal was assigned for LVEF ≥50% or a clear qualitative description of normal or preserved systolic function. Abnormal was assigned for LVEF <50% or a clear qualitative description of reduced or dysfunctional systolic function. Numeric LVEF took priority when numeric and qualitative findings differed; Unknown was used when interpretable information was unavailable.

2.4.3. Patient-Reported Symptoms

Symptoms were categorized as Asymptomatic, Symptoms of heart failure, or Other symptoms. Cases with no documented patient-reported symptoms or only explicit symptom denials in the selected clinical documentation were categorized as Asymptomatic. Prespecified heart-failure symptoms included breathlessness, paroxysmal nocturnal dyspnea, swelling, and orthopnea. These

symptoms were classified as Symptoms of heart failure when explicitly documented as new, worsening, progressive, or increasing, or when associated with declining functional status. Stable, persistent, ongoing, baseline, or otherwise nonqualifying symptoms were classified as Other symptoms.

2.4.4. Physical-Examination Signs

Examination findings were categorized as Normal examination, Signs of heart failure, or Nonspecific physical-examination abnormalities. Cases with no documented abnormal physical-examination findings in the selected clinical documentation were categorized as Normal examination. Prespecified qualifying signs included jugular venous distention, adventitious breath sounds, ascites, qualifying peripheral edema, and an explicitly documented S3 gallop.

2.5. Temporal Interpretation and Index-Date Anchoring

Clinical information was interpreted relative to the patient-specific index date. For echocardiography, an LLM first extracted candidate echo mentions with date precision, date type, temporal relationship to the index date, supporting text, and source-page information; a deterministic Python step then performed final date selection. A qualifying completed study on the index date was prioritized. Otherwise, the most recent eligible pre-index procedure-date or comparison-reference candidate was selected, with day-precision evidence preferred over month-only evidence and procedure dates prioritized when dates were tied. Planned or administrative records, non-echo dates, and post-index candidates were excluded. For symptoms and signs, same-day clinical documentation was prioritized; when unavailable, the most recent qualifying pre-index assessment and clinically meaningful intervening documentation through the index date were used.

2.6. Prompt Development and Workflow Finalization

We developed prompts iteratively using the 40-case development cohort; no validation cases were used for prompt development or workflow refinement. Recurring ambiguities were resolved by clarifying general decision rules rather than encoding case-specific answers. The finalized system used a hybrid deterministic–LLM architecture. A clinical-note selection prompt returned structured JSON fields Same_Day, Note_Date, and Description that governed same-day versus pre-index symptom and examination pathways. Echocardiogram extraction produced structured Echo_Candidates entries with JSON fields Date_Raw, Date_Normalized, Date_Precision, Date_Type, Echo_Type, Relative_to_Index, Page_Number, and Evidence_Quotes; final echo-date selection was performed deterministically in Python using the temporal-ranking rules described in Section 2.5.

LVF used dated or undated prompt pathways according to whether an echo date had been selected; the dated pathway was constrained to the selected echocardiogram except for the prespecified echo-to-echo comparison fallback. Representative LVF JSON fields included LVF_Status, Echo_Date, LVF_Value, Evidence_Quotes, and Reason. Symptoms and physical-examination signs used separate same-day and no-same-day prompt pathways controlled by the

upstream clinical-note result, with structured Classification, Items, and Evidence outputs. When no eligible clinical-note date was identified, the prespecified Asymptomatic and Normal examination outputs were assigned deterministically for the symptom and examination domains, respectively, without an additional LLM call. All prompts specified eligible temporal windows, allowable outputs, clinical thresholds and precedence rules, negation and temporal-modifier handling, evidence requirements, and verbatim/page-linked supporting evidence. After development, all prompt templates, branching logic, classification criteria, and output schemas were frozen before validation, and no changes were made in response to validation results.

2.7. LLM-Based Clinical Abstraction and Evidence Linking

Final validation inference was performed in July 2026 using the Databricks model identifier system.ai.claude-opus-4-8, accessed through the databricks-claude-opus-4-8 serving endpoint. No custom generation parameters were specified in the validation notebooks; Databricks service defaults were used. Inference was performed within the company-approved Databricks environment configured for processing protected health information [8]. For each case, the model received page-linked clinical text, the patient-specific index date, and relevant upstream workflow variables such as the selected echocardiogram date or same-day documentation status. Prompts constrained the model to record content, prohibited unsupported inference, required machine-readable JSON, and requested supporting quotations and page identifiers. Structured outputs were passed to report generation without modifying the frozen clinical or temporal rules.

2.8. Blinded Comparative Validation and Statistical Analysis

For each validation case, the AI system and EdiPhy (MedManagement, LLC), an external professional medical-review service (EPS), independently processed the same PDF medical-record package and patient-specific index date. Before abstraction, EdiPhy received the same prespecified clinical definitions, allowable classification categories, and temporal-selection rules described in Sections 2.4 and 2.5 and used by the AI system and physician adjudicator. Both methods generated outputs for the four target domains. An internal physician then reviewed the original source PDF and two standardized outputs labeled Output A and Output B. The order of the two outputs was randomized, and the source identity of each output was masked so that the physician did not know whether a given output came from the AI system or EdiPhy. Each output was judged as correct or incorrect against the source documentation and the prespecified clinical abstraction criteria. Paired outcomes were categorized as both correct, AI only correct, EPS only correct, or both incorrect.

Single-method accuracy was summarized with Wilson 95% confidence intervals [16]. Paired differences were defined as (AI-only correct minus EPS-only correct)/N; confidence intervals used the observed within-pair discordance variance. Exact two-sided McNemar tests [17] compared methods. Each validation case contributed one paired comparison in each of the four clinical domains, yielding four paired case-domain comparisons per case. Domain-specific analyses were primary. The pooled analysis across 628 case-domain comparisons was a secondary summary because the four domain outcomes contributed by each validation case were not independent [18].

As sensitivity analyses, we defined an all-four-domains-correct outcome for each validation case and estimated paired-difference 95% confidence intervals for both the case-level and pooled case-domain analyses using patient-clustered robust (CR1) standard errors. All outcomes and repeated validation cases from the same patient were assigned to a single cluster, and t-based confidence intervals used G−1 degrees of freedom.

2.9. Privacy, Ethics, and Data Governance

All protected health information was processed within company-approved systems with access restricted to authorized personnel. AI-generated outputs were used solely for retrospective evaluation and were not used to guide patient care.

3. Results

3.1. Validation Cohort

The independent validation cohort comprised 157 validation cases from 141 adult heart transplant recipients, including records from 52 U.S. transplant centers. Fourteen patients contributed more than one validation case. The 157 record packages contained 5,033 pages and varied substantially in length, formatting, and documentation conventions. The median package length was 17 pages (IQR, 9–29; range, 2–198), and all packages consisted entirely of scanned or image-based pages. Both methods generated outputs for four domains, yielding 157 paired comparisons per domain and 628 paired case-domain comparisons overall.

Table 1. Validation Cohort and Medical-Record Package Characteristics

Characteristic	Value
Development cohort, cases	40
Validation cases	157
Unique validation patients	141
Patients with >1 validation case	14
Contributing transplant centers	52
External medical-record packages	157
Total record pages	5,033
Pages per package, median (IQR; range)	17 (9–29; 2–198)
Paired case-domain comparisons	628

3.2. Page-Linked Parsed Output Was Complete in 153 of 157 Packages

Page-linked text and document-layout elements were generated for all 157 packages and were available for downstream abstraction. Manual review identified incomplete recovery of target-relevant source content in 4 packages (2.5%); the remaining 153 packages (97.5%) had complete parsed output under the study definition. The median number of extracted elements was 443 per package (IQR, 231–773; range, 31–8,520) and 26 per page (IQR, 21–34; range, 6–76). These results should not be interpreted as character-level OCR accuracy.

3.3. Temporal Selection Was 98.7% Accurate for Echocardiograms and 100% Accurate for Clinical Assessments

A qualifying echocardiogram performed on the index date was available in 39 of 157 cases (24.8%). For the remaining cases, the AI system evaluated pre-index echocardiographic evidence and selected the most recent qualifying echocardiogram date under the prespecified deterministic ranking rules when one was available. When no qualifying echocardiogram date was supported on or before the index date, the correct output was no study selected. Multiple candidate echocardiograms or related dates were present in 94 cases; the correct candidate was selected in 92 of these cases (97.9%). Among cases with a qualifying pre-index study, the median interval between the selected study and the index date was 82 days (IQR, 21–169; range, 1–1,334). Overall, the AI system produced the correct temporally grounded echocardiogram-date output in 155 of 157 cases (98.7%; 95% CI, 95.5%–99.6%), including both correct study selection and correct identification of cases with no qualifying study.

For clinical assessments, same-day documentation was available in 67 of 157 cases (42.7%) for each domain. When same-day documentation was unavailable, the system evaluated the most relevant pre-index evidence under the prespecified temporal rules. Under the prespecified task definitions, cases with no documented patient-reported symptoms were classified as Asymptomatic, whereas documented symptoms that did not meet the prespecified new/worsening heart-failure criteria were classified as Other symptoms. Temporal selection and handling of absent or negative evidence were correct in all 157 cases (100%) for both domains. Planned or scheduled studies, post-index studies, and historical, resolved, or copied-forward findings were excluded from the selected evidence.

3.4. Source-Page Evidence Was Present for 390 Clinical Abstractions

Across the four domains, 390 AI-generated abstractions contained at least one standardized source-page identifier, with 1,035 page-reference occurrences in total. These comprised 351 references across 99 symptom outputs, 186 across 120 echocardiogram-date outputs, 393 across 131 left ventricular function outputs, and 105 across 40 physical-examination outputs. Because one abstraction could be supported by multiple locations, the median was 2 page references per page-linked abstraction (IQR, 1–4; range, 1–9). Among the 390 page-linked abstractions, 388 (99.5%) were clinically correct. Of 1,035 page-reference occurrences, 1,033 were attached to clinically correct abstractions. Although page-linked abstractions were assessed for clinical correctness, we did not separately evaluate whether each quoted excerpt was fully faithful to the source text or whether each page reference pointed to the exact supporting location.

The presence of source-page evidence varied according to the clinical classification. Specific echocardiogram dates, determinate left ventricular function, reported symptoms, and abnormal examination findings were generally accompanied by source-page evidence, whereas evidence output fields were generally blank for outputs indicating no qualifying finding, including unavailable echocardiogram information, an Asymptomatic classification, or a Normal examination.

Table 2. Source-Page Evidence Linking by Domain

Domain	Page-Linked Abstractions, n	Page-Reference Occurrences, n
Echocardiogram date	120	186
Left ventricular function	131	393
Patient-reported symptoms	99	351
Physical-examination signs	40	105
Overall	390	1,035

3.5. Comparative Clinical Abstraction Performance and Error Analysis

3.5.1. Secondary Pooled and Case-Level Sensitivity Analyses

Across 628 paired case-domain comparisons, the AI system produced 626 correct abstractions (99.7%; 95% CI, 98.8%–99.9%), compared with 518 correct abstractions for the EPS (82.5%; 95% CI, 79.3%–85.3%). The paired difference was 17.2 percentage points (95% CI, 14.2–20.2; exact McNemar P<.001). Both methods were correct in 516 comparisons; the AI system alone was correct in 110; and the EPS alone was correct in 2. There were no comparisons in which both methods were incorrect. The pooled result is a secondary summary across case-domain comparisons because four domain outcomes were nested within each case and because correctness was adjudicated under the study's prespecified internal reference framework. In the case-level sensitivity analysis, all four domains were correct within the same validation case in 155/157 AI cases (98.7%) and 87/157 EPS cases (55.4%), corresponding to a paired difference of 43.3 percentage points (patient-clustered 95% CI, 35.2–51.4). Re-estimation of the pooled paired difference across case-domain comparisons using patient-clustered robust standard errors yielded 17.2 percentage points (95% CI, 13.5–20.9), indicating that accounting for within-patient dependence did not materially change the estimated difference.

3.5.2. Primary Domain-Specific Paired Accuracy

The AI system was correct in 155 of 157 echocardiogram-date comparisons (98.7%) and all 157 comparisons in each of the other three domains. EPS accuracy was 137 of 157 (87.3%) for echocardiogram date, 147 of 157 (93.6%) for left ventricular function, 109 of 157 (69.4%) for patient-reported symptoms, and 125 of 157 (79.6%) for physical-examination signs. Under the shared prespecified abstraction framework, the fixed AI system had higher paired accuracy than the external professional service in every domain evaluated in this study, with the largest difference for patient-reported symptoms (Figure 2).

Clinically positive reference classes were uncommon: abnormal left ventricular function was present in 11 of 157 cases (7.0%), heart-failure symptoms in 10 of 157 (6.4%), and heart-failure signs in 6 of 157 (3.8%). The AI system correctly classified all positive cases (11/11, 10/10, and 6/6, respectively), with no observed false-positive classifications in these categories; corresponding EPS outputs were adjudicated correct in 10/11, 1/10, and 1/6 cases.

Table 3. Paired Accuracy by Clinical Domain

Domain	AI Correct, n/N (%)	EPS Correct, n/N (%)	EPS Missed	AI Missed	Exact McNemar P
Echocardiogram date	155/157 (98.7)	137/157 (87.3)	20	2	<.001
Left ventricular function	157/157 (100)	147/157 (93.6)	10	0	0.002
Patient-reported symptoms	157/157 (100)	109/157 (69.4)	48	0	<.001
Physical-examination signs	157/157 (100)	125/157 (79.6)	32	0	<.001
Overall	626/628 (99.7)	518/628 (82.5)	110	2	<.001

Note. EPS = external professional abstraction service. EPS Missed indicates paired comparisons in which the AI system was correct and the EPS was incorrect. AI Missed indicates paired comparisons in which the EPS was correct and the AI system was incorrect. The Overall row is a secondary pooled summary across case-domain comparisons.

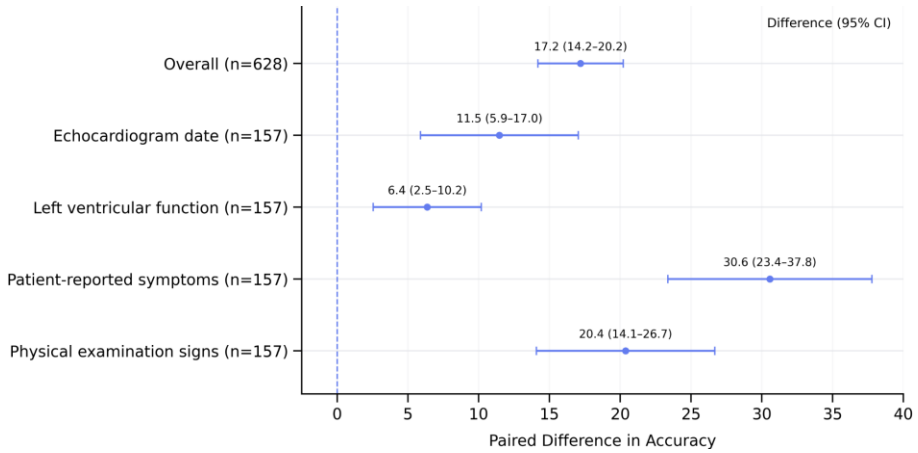


Figure 2. Paired differences in abstraction accuracy between the AI system and EPS. Points represent paired differences in percentage points, and horizontal lines represent 95% confidence intervals. Positive values favor the AI system. Domain-specific analyses included 157 paired comparisons per domain; the overall pooled summary included 628 case-domain comparisons across the four domains and is secondary because four domain outcomes were nested within each case.

3.5.3. Error Analysis Characterized Sources of Discordant Paired Comparisons

Among the 112 discordant paired comparisons, 110 favored the AI system and 2 favored the EPS. EPS-only errors occurred across all four domains, including 20 echocardiogram-date outputs, 10 left ventricular function outputs, 48 patient-reported symptom outputs, and 32 physical-examination outputs (Table 3). The largest concentrations of discordance were therefore in symptom and physical-examination abstraction. Of the 20 EPS echocardiogram-date errors, 10 involved selection of an older rather than the most recent qualifying echocardiogram, 6 involved a blank date despite a qualifying study being present, and 4 involved citation of a date without a corresponding qualifying study. Of the 10 EPS left ventricular function errors, 3 involved classification of normal function as abnormal, 6 involved leaving normal function blank, and 1 involved leaving abnormal function blank. EPS symptom errors included misclassification of documented symptoms as Asymptomatic or Other symptoms, as well as blank outputs despite documentable symptom information. EPS physical-examination errors most commonly involved classifying nonspecific abnormalities as normal (14 cases), with 2 cases under-classifying signs of heart failure as nonspecific abnormalities, and the remaining 16 leaving the classification blank despite documented abnormal findings.

The two AI errors both occurred in echocardiogram-date abstraction and were attributable to upstream document parsing and incomplete text recognition. In one case, the source date "05/14/2025" was recognized as "06/14/2025"; in the other, the relevant source text was not parsed completely. No AI errors were observed in left ventricular function, symptom, or physical-examination classification.

EPS discordances showed a different profile. Echocardiogram-date errors included selection of older or non-qualifying studies rather than the completed study most relevant to the index date. Symptom discordances frequently reflected failure to capture documented symptoms or to distinguish asymptomatic status, stable or other symptoms, and explicitly new or worsening heart-failure symptoms. Physical examination discordances similarly included missed abnormal findings and classification of nonspecific abnormalities as normal examinations. Because both methods were evaluated against the same prespecified classification definitions and temporal rules, these findings indicate differences in observed temporal selection, information capture, and clinical classification under the study framework. In this validation cohort, AI failures were localized to source-text recovery, whereas EPS errors occurred across all four domains and multiple stages of abstraction.

4. Discussion

This internal validation study evaluated an end-to-end AI system that combines page-preserving document parsing, LLM-based clinical information extraction and temporal interpretation, index-date-conditioned evidence selection, and source-page evidence linking (Figure 1). Domain-specific accuracy was 98.7% for echocardiogram date and 100% for left ventricular function, symptoms, and physical-examination signs. In the secondary pooled summary across 628 case-domain comparisons, the AI system achieved 99.7% accuracy compared with 82.5% for the EPS. These results demonstrate high performance under the prespecified study definitions. Sensitivity analyses were consistent with the primary findings: all four domains were correct within the same validation case in 155/157 AI cases (98.7%) versus 87/157 EPS cases (55.4%), and patient-clustered analysis of the pooled case-domain comparisons yielded a paired difference of 17.2 percentage points (95% CI, 13.5–20.9).

The principal methodological contribution is system-level: the workflow combines page-preserving representation, constrained LLM-based candidate extraction and clinical classification, deterministic index-date–conditioned temporal selection, and evidence provenance in a single fixed pipeline for scanned longitudinal records. This differs from prior document-level extraction [21,22] by requiring selection among competing dated evidence relative to an index date, while sharing with recent temporal-reasoning [23] and source-grounded chart-review work [24] an emphasis on longitudinal context and source traceability. Only 39 of 157 cases had a qualifying echocardiogram on the index date; the remaining 118 required evaluation of pre-index echocardiographic evidence, including selection of the most relevant prior date when supported or determination that no qualifying date was available. Multiple candidate dates were present in 94 cases. The observed temporal-selection performance therefore reflects selection of the clinically relevant occurrence rather than simple concept recognition, consistent with prior work emphasizing the importance of temporal reasoning in clinical natural language processing [4].

The paired analysis showed higher accuracy for the fixed AI system than for the EPS as operationally delivered in each prespecified domain under the shared abstraction framework (Figure 2). The largest difference occurred in symptom abstraction, where classifying findings as Asymptomatic, Symptoms of heart failure, or Other symptoms—and determining whether documented symptoms were new or worsening—depended on temporal modifiers and clinical context. Physical-examination classification similarly required separation of qualifying heart-failure signs from nonspecific abnormalities. Because the comparator was an external professional service rather than individual clinicians or human abstractors evaluated under experimentally controlled staffing and workflow conditions, these results should not be interpreted as evidence that the AI system is superior to physicians, clinicians, or professional abstractors generally [4,10].

Evidence linking added reviewability rather than an independent guarantee of correctness. Page identifiers were present in 390 abstractions and occurred 1,035 times, frequently linking one output to multiple record locations. Among the 390 page-linked abstractions, 388 were clinically correct. However, the study did not independently adjudicate each quotation, page location, or evidence set for faithfulness and completeness. Future work should evaluate abstraction correctness, quotation faithfulness, page-location accuracy, evidence completeness, and appropriate evidence absence as separate endpoints [10,12,14].

The two observed AI errors were localized to source-text recovery for echocardiogram dates, highlighting the importance of instituting quality control upstream of the LLM. Such measures may include document-quality assessment, targeted verification of dates and numeric values, reprocessing of low-confidence pages, and escalation when relevant text is incomplete or internally inconsistent [13,14]. Because only two AI errors occurred, the present cohort cannot characterize uncommon failure modes or establish their relative frequency.

Several design features strengthen the evaluation: development and validation cohorts were separated; prompts, rules, and workflow were fixed before validation; both methods processed the same records and index dates; EdiPhy received the same prespecified clinical definitions, allowable classifications, and temporal-selection rules used by the AI system and physician adjudicator; and outputs were masked and randomized for physician review. In addition, the inclusion of 5,033 scanned pages from 52 U.S. transplant centers, encompassing substantial variation in record length, formatting, and documentation conventions, strengthens the real-world relevance of the evaluation and supports performance across the heterogeneous document conditions represented in this cohort [10,13].

Important limitations remain. Correctness in the present study was adjudicated by a single internal physician using the original source PDF, prespecified clinical definitions, and masked, randomized presentation of AI and EPS outputs. Although masking reduced the potential for source-identification bias during adjudication, reliance on a single adjudicator precluded assessment of inter-rater agreement and may reflect adjudicator-specific interpretation. Accordingly, the near-ceiling accuracy observed in this study should be interpreted as internal validation under the prespecified reference framework rather than as a definitive estimate of broadly generalizable clinical accuracy. A larger prospective validation study using two independent adjudicators is currently underway to assess reproducibility of the reference standard and agreement between reviewers.

Additional limitations relate to the external comparator and study design. Although EdiPhy received the same prespecified clinical definitions and temporal-selection rules, its medical-review process was externally managed, and the investigators did not have visibility into the identities of individual abstractors or reviewers, internal reviewer assignment, staff training procedures, time per case, or quality-control processes. The observed results should therefore be interpreted as a comparison with the external professional service as operationally delivered rather than as a controlled comparison between identically trained individual human abstractors and the AI system. The study was retrospective and internal; 14 patients contributed more than one validation case. Case-level and patient-clustered sensitivity analyses addressed the resulting dependence structure, but the limited number of repeated cases restricted characterization of within-patient variability across repeated validation cases. Clinically positive classes were uncommon, limiting the precision of class-specific performance estimates despite the absence of observed AI classification errors in these categories. The study used one parser, one LLM configuration, and one fixed prompt set; no component-ablation analysis, simpler computational baseline, alternative parser, or alternative LLM configuration was evaluated. Accordingly, the present study estimates performance of the fixed end-to-end system and does not isolate the incremental contribution of individual pipeline components. The study also did not measure review time, workload, usability, downstream decisions, automation bias, character-level OCR accuracy, or independent citation-level correctness. The structured JSON outputs used study-specific schemas, and standards-based interoperability, including mapping to FHIR, LOINC, or SNOMED CT, was not evaluated. These limitations constrain generalizability and underscore the need for transparent reporting and staged clinical evaluation [10,13-15].

Before clinical deployment, the method requires external and prospective validation, multiple independent adjudicators, larger cohorts for more precise class-specific performance estimates, and direct evaluation of evidence faithfulness and completeness [10,14,15]. Prospective workflow studies should evaluate review time, usability, downstream decision effects, and automation bias [11,15], while future integration work should evaluate standards-based mappings where appropriate. The present approach is intended to complement, rather than replace, interoperable structured data exchange when such data are available. The present results support feasibility of an index-date-conditioned, provenance-preserving abstraction framework, not autonomous clinical decision-making.

5. Conclusion

In this retrospective internal validation study, the fixed AI system achieved 98.7% accuracy for echocardiogram-date abstraction and 100% accuracy for left ventricular function, patient-reported symptoms, and physical-examination signs under a prespecified internal reference framework. Compared with EdiPhy as operationally delivered using the same clinical definitions and temporal-selection rules, the fixed AI system produced higher paired accuracy in all four evaluated abstraction domains. These findings support the feasibility of accurate, human-review–oriented abstraction from complex longitudinal records with source-page evidence linking. External validation, multi-adjudicator assessment, and prospective workflow evaluation remain necessary before clinical deployment or broader conclusions about generalizable clinical accuracy.

References

1. A. Velleca, M. A. Shullo, K. Dhital, et al., The International Society for Heart and Lung Transplantation (ISHLT) guidelines for the care of heart transplant recipients, J. Heart Lung Transplant. 42, e1 (2023).
2. D. Bacich, C. Tessari, G. Ciccarelli, et al., A comprehensive excursus of the roles of echocardiography in heart transplantation follow-up, J. Clin. Med. 13, 3205 (2024).
3. H. Alansari and J. R. Gorthi, Rejection surveillance: integrating cell-free DNA and gene expression panels to decrease invasiveness in routine monitoring of heart transplant recipients, Methodist Debakey Cardiovasc. J. 21, 51 (2025).
4. L. Zhou and G. Hripcsak, Temporal reasoning with medical data-A review with emphasis on medical natural language processing, J. Biomed. Inform. 40, 183 (2007).
5. Y. Wang, L. Wang, M. Rastegar-Mojarad, et al., Clinical information extraction applications: a literature review, J. Biomed. Inform. 77, 34 (2018).
6. A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, et al., Large language models in medicine, Nat. Med. 29, 1930 (2023).
7. Databricks, ai_parse_document function, Databricks on AWS Documentation (2026); available at https://docs.databricks.com/aws/en/sql/language-manual/functions/ai_parse_document (accessed July 2026).
8. Databricks, Query with the Anthropic Messages API, Databricks on AWS Documentation (2026); available at https://docs.databricks.com/aws/en/machine-learning/model-serving/query-anthropic-messages (accessed July 2026).
9. Z. Lu, Y. Peng, T. Cohen, M. Ghassemi, C. Weng, and S. Tian, Large language models in biomedicine and health: current research landscape and future directions, J. Am. Med. Inform. Assoc. 31, 1801 (2024).
10. J. Gallifant, M. Afshar, S. Ameen, et al., The TRIPOD-LLM reporting guideline for studies using large language models, Nat. Med. 31, 60 (2025).
11. K. Goddard, A. Roudsari, and J. C. Wyatt, Automation bias: a systematic review of frequency, effect mediators, and mitigators, J. Am. Med. Inform. Assoc. 19, 121 (2012).
12. S. Basu and B. Huynh, Mitigating hallucinations in healthcare AI: a systematic review of evidence-based strategies, BMC Health Serv. Res. 26, 1115 (2026); doi:10.1186/s12913-026-14851-1.
13. J. Wiens, S. Saria, M. Sendak, et al., Do no harm: a roadmap for responsible machine learning for health care, Nat. Med. 25, 1337 (2019).
14. K. Lekadir, A. Feragen, A. J. Fofanah, et al., FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare, BMJ 388, e081554 (2025); doi:10.1136/bmj-2024-081554.

15. B. Vasey, M. Nagendran, B. Campbell, et al., Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI, BMJ 377, e070904 (2022); doi:10.1136/bmj-2022-070904.
16. E. B. Wilson, Probable inference, the law of succession, and statistical inference, J. Am. Stat. Assoc. 22, 209 (1927).
17. Q. McNemar, Note on the sampling error of the difference between correlated proportions or percentages, Psychometrika 12, 153 (1947).
18. K.-Y. Liang and S. L. Zeger, Longitudinal data analysis using generalized linear models, Biometrika 73, 13 (1986).
19. N. R. Sahni, B. Istvan, C. Stafford, and D. M. Cutler, Perceptions of prior authorization burden and solutions, Health Aff. Sch. 2, qxae096 (2024); doi:10.1093/haschl/qxae096.
20. W. C. Chen, C. Carpenter, B. Sidiqi, et al., Integrating prior authorization into clinical workflows for care access and practitioner experience, JAMA Netw. Open 8, e2549093 (2025); doi:10.1001/jamanetworkopen.2025.49093.
21. J. Huang, D. M. Yang, R. Rong, et al., A critical assessment of using ChatGPT for extracting structured data from clinical notes, npj Digit. Med. 7, 106 (2024).
22. M. M. Dagli, Y. Ghenbot, H. S. Ahmad, et al., Development and validation of a novel AI framework using NLP with LLM integration for relevant clinical data extraction through automated chart review, Sci. Rep. 14, 26783 (2024); doi:10.1038/s41598-024-77535-y.
23. H. Cui, A. Unell, B. Chen, et al., TIMER: temporal instruction modeling and evaluation for longitudinal clinical records, npj Digit. Med. 8, 577 (2025); doi:10.1038/s41746-025-01965-9.
24. R. Verma, E. Alsentzer, Z. Strasser, et al., Verifiable summarization of electronic health records using large language models to support chart review, medRxiv (2025); doi:10.1101/2025.06.02.25328807.