Configurable Clinical Information Extraction with Agentic RAG: What Works, What Breaks, and Why

Osman Alperen Çinar-Koraş[1,2*] **Marie Bauer**[1] **Sameh Khattab**[1,2] **Merlin Engelke**[1] **Moon Kim**[1]
Stephan Settelmeier[6] **Shigeyasu Sugawara**[1,5] **Fabian Freisleben**[1] **Felix Nensa**[1] **Jens Kleesiek**[1,2,3,4]

[1]Institute for Artificial Intelligence in Medicine (IKIM), University Medicine Essen, Essen, Germany
[2]Faculty of Computer Science, University of Duisburg-Essen, Essen, Germany
[3]Department of Physics, TU Dortmund University, Dortmund, Germany
[4]Lamarr Institute for Machine Learning and Artificial Intelligence, TU Dortmund University, Germany
[5]Advanced Clinical Research Center, Fukushima Medical University, Fukushima, Japan
[6]Department of Cardiology and Vascular Medicine, University Hospital Essen, Essen, Germany

Abstract

Patient contexts span hundreds of heterogeneous documents and thousands of structured data points, yet the document-level metadata that AI systems need for retrieval and triage is absent or incomplete. Standard retrieval-augmented generation fails on this data, mishandling temporal reasoning, cross-document dependencies, and missing metadata. We deploy ACIE (Agentic Clinical Information Extraction) at University Medicine Essen: an on-premise agentic RAG pipeline that reasons over complete patient contexts and grounds every answer in source passages for clinician verification. We quantify the metadata gap, trace the architectural decisions it shaped, and evaluate extraction alongside an independent retrospective lymphoma registry study, in which nuclear-medicine physicians verify every extracted value against its cited sources. Across 7,326 judgments, clinicians accepted 96.5% of extractions, with per-type acceptance ranging from 80% to 99%.

1 Introduction

Clinical workflows routinely require structured data compiled from patient records spanning thousands of documents and tens of thousands of structured data points across multiple hospital systems. Enrolling a single lymphoma patient in a clinical study, for example, requires reconstructing the treatment history and locating diagnostic markers across years of documents that may be duplicated, misdated, or buried among unrelated records. Clinicians perform this compilation by hand, and information is routinely missed (Moon et al., 2022).

Clinical information extraction (IE) has long aimed to alleviate this burden, yet even recent deployed systems require developer effort to adapt to new workflows (§2). Large language models can perform extraction without task-specific training (Singhal et al., 2023; Agrawal et al., 2022), but LLM-based clinical IE remains largely confined to research evaluations (Artsi et al., 2025). Two barriers explain the gap. First, transmitting patient data to external servers raises privacy and regulatory risks prompting for on-premise deployment (Dennstädt et al., 2025). Second, real patient records pose retrieval challenges that standard RAG is not designed for, because the metadata it depends on is unreliable, documents are interdependent, and conflicting values require temporal reasoning to resolve. A recent scoping review found that only 9% of end-to-end medical RAG systems employ agentic architectures (Miao et al., 2025).

We deploy ACIE (Agentic Clinical Information Extraction) at University Medicine Essen, whose FHIR repository, with nearly 2 billion resources, is among the largest in Europe. Clinicians define extraction schemas with typed targets without developer involvement. An agentic RAG pipeline reasons over complete patient contexts, grounding every value in source passages for clinician verification, running entirely on-premise. Our contributions are:

1. A clinician-verified evaluation of agentic extraction alongside an independent retrospective lymphoma registry study (74 clinician-configured fields, 99 patients, 7,326 judgments), in which nuclear-medicine physicians accept or reject every extracted value and label rejections with structured error and editorial categories.

*Corresponding author.

2. A quantified analysis of the metadata gap between what AI systems need and what clinical data exports provide.

3. Architectural decisions shaped by this data reality, illustrating the design trade-offs of building on real clinical data.

2 Related Work

From rules to domain-specific pretraining. Early clinical IE relied on engineered NLP pipelines such as cTAKES (Savova et al., 2010), where extraction targets were defined by developers. Knowledge Author (Scuba et al., 2016) enabled domain experts to define schemas through a web interface, but its rule-based backend limited expressiveness: only 76% of target concepts could be fully represented, with recall as low as 46%. Domain-specific pretraining (BioBERT (Lee et al., 2020), GatorTron (Yang et al., 2022)) improved accuracy but still required fine-tuning per extraction target. In a survey of 263 clinical IE studies, (Wang et al., 2018) found that over half targeted disease-related extraction spanning 88 unique diseases, concluding that the portability and generalizability of clinical IE systems are still limited. Community shared tasks from i2b2/VA (Uzuner et al., 2011) and n2c2 (Henry et al., 2020) to recent iterations (Lybarger et al., 2023; Yao et al., 2024) similarly operate on predefined targets. Throughout this, extraction targets remained fixed or configurability mechanisms could not meet the demands of real clinical complexity.

LLMs shift what is possible but remain largely undeployed. LLMs enabled few-shot clinical extraction without task-specific training, yet LLM-based clinical IE has rarely moved beyond research evaluations (Artsi et al., 2025). (Wiest et al., 2024) locally deploy Llama 2 for five fixed features from MIMIC-IV patient histories, demonstrating on-premise feasibility but with static, researcher-defined targets. LLM-AIx (Wiest et al., 2025) provides an open-source pipeline for structured extraction from individual documents with user-defined schemas and local inference, but processes documents independently without retrieval augmentation and has only been validated on research datasets. Deployed systems have followed a separate track. MedCAT (Kraljević et al., 2021) and MiADE (Jiang-Kells et al., 2025) are clinical NLP pipelines in production, but with developer-defined targets. Griot (Griot et al., 2025) deploys Qwen3-235B with RAG inside Epic for clinical assistance (1,028 users) and Grünig et al. (Grünig et al., 2026) deploy an on-premise LLM at a German university hospital, but neither performs structured extraction.

Agentic RAG as the emerging frontier. Retrieval-augmented generation (Lewis et al., 2020) grounds LLM outputs in retrieved evidence, and agentic frameworks like ReAct (Yao et al., 2023) enable iterative reasoning over complex information needs. i-MedRAG (Xiong et al., 2025) shows that iterative retrieval outperforms single-pass RAG for medical QA but uses a fixed iteration schedule on curated knowledge bases. Agentic clinical IE has recently emerged: CLINES (Yang et al., 2025) structures clinical concepts through a modular pipeline, HARMON-E (Gupta et al., 2025) applies hierarchical reasoning to oncology notes, and ReflecTool (Liao et al., 2025) benchmarks tool-augmented clinical agents. However, all use researcher-defined targets on benchmark datasets; none are deployed with clinician-configurable schemas. To our knowledge, no prior work has quantified the gap between the metadata AI systems need and what clinical data exports provide, or traced architectural decisions to data quality failures in a deployed system.

3 System Overview

Figure 1 illustrates the pipeline. Clinicians configure extraction targets through typed schemas, and the system handles retrieval and extraction. This section describes the deployed system. §5 traces the architectural decisions behind it. Implementation details of the extraction schema engine, agent orchestration, and document export pipeline are outside the scope of this paper.

3.1 Patient Context

A patient context consists of all clinical data available for a patient: documents (discharge letters, radiology reports, laboratory findings, referral letters, operative notes) and structured FHIR (HL7, 2019) data points (laboratory results, medications, conditions, observations) accumulated over years of care. §4.1 characterizes the scale and quality of this data.

ACIE ingests all available data via the hospital's FHIR server. Non-machine-readable documents are processed via OCR, and those falling below a quality threshold are excluded. Each document is semantically chunked at two granularities: coarse

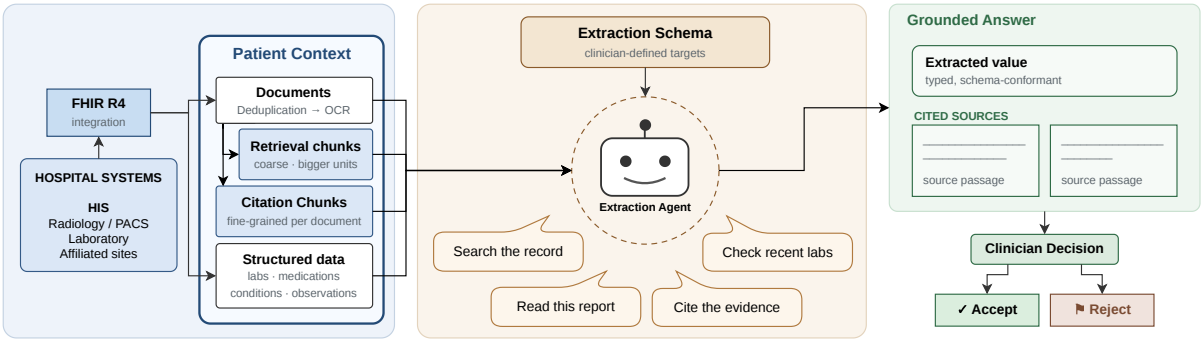

Figure 1: ACIE system overview. Clinical data from multiple hospital systems, accessed via a FHIR server, is organized into a patient context. Each document is chunked at two granularities for retrieval and citation. For each extraction target, an agent iteratively searches and inspects the patient context, returning a grounded answer that the clinician verifies against cited source passages. The structured-data labels indicate FHIR resource types.

retrieval chunks that preserve surrounding context, and fine-grained passages serving as atomic units for source citation. Clinical documents interleave narrative, tabular, and form-like content in heterogeneous layouts. Chunking their serialized text yields many short, low-information fragments that dense retrievers systematically favor (Fayyaz et al., 2025). We apply a length-penalized retrieval score:

$$s = \mathrm{sim}(q, c) \cdot p(\ell), \quad p(\ell) = \min\left(\frac{\ell}{\tau}, 1\right) \cdot \frac{2}{3} + \frac{1}{3} \tag{1}$$

where $\mathrm{sim}(q, c)$ is the cosine similarity between the query q and chunk c, ℓ is the chunk length in characters, and $\tau = 40$. The penalty $p(\ell)$ dampens scores of short fragments without eliminating them. We fixed τ and the blend weights on a development subset and held them constant across tasks.

3.2 Agentic Extraction

Clinicians define extraction targets through a typed schema. The same system serves different clinical use cases (pre-procedure protocols, retrospective study data collection, clinical documentation) through schema configuration alone, without code changes.

For each extraction target, a tool-calling agent (Yao et al., 2023) searches the patient context. Standard retrieve-then-generate pipelines with metadata-based filters proved insufficient because the metadata they depend on is unreliable or absent (§4.1, §5). The agent's tools allow searching by semantic similarity across the full patient context, listing documents with query-relevant summaries, inspecting a specific document in detail, and querying structured data directly. When listing documents, summaries are generated on the fly by assembling the highest-scoring citation chunks per document in document order until at least 200 words are accumulated, producing a chronologically coherent, content-based relevance preview.

The agent iterates until it has gathered sufficient evidence, then returns an answer with every value attributed to specific source passages. This grounding is a safety requirement: clinicians review each extracted value against the cited passages and accept or reject it before it enters clinical documentation.

3.3 Deployment

ACIE runs entirely on-premise on hospital infrastructure, deployed as a web application on Kubernetes. Patient data never leaves the hospital network. The extraction model is Qwen 3.6 35B-A3B (Qwen Team, 2026), a mixture-of-experts model. Scanned documents are processed by PaddleOCR-VL 1.5 (Cui et al., 2026). For this evaluation, the extraction model was served on 4×H100 GPUs and the OCR model on a single H100 GPU.

4 Evaluation

ACIE is deployed at University Medicine Essen, whose FHIR R4 server, conforming to a national interoperability core-dataset specification, integrates nearly 2 billion resources across 1.7 million patients from the hospital's primary information system, radiology, laboratory, and affiliated hospitals, making it one of the largest clinical FHIR repositories in Europe. Despite this scale, document-level metadata remains sparse (§4.1). Table 1 summarizes the corpus, Table 2 characterizes patient contexts from 10,000 randomly sampled patients.

Category	Count
Patient records	5,598,272
Unique individuals	1,747,135
Orders and requests	852M
Lab values and observations	675M
Clinical documents	84M
Medication orders	77M
Diagnoses and conditions	40M
Medication administrations	33M
Clinical encounters	29M
Other	182M
Total FHIR resources	**1.97B**

Table 1: Clinical data corpus. Only the most populated resource types are shown.

	Med. (IQR)	P99	Max
Docs	52 (14–140)	937	2,542
Dedup. (%)	33.5 (20.0–43.0)	–	54.6
Struct. resources	406 (38–1,922)	37,074	119,191
Encounters	18 (6–46)	323	1,207
OCR rej. (%)	10.3 (6.8–17.8)	–	52.0

Table 2: Per-patient statistics (n=10,000), sampled randomly from 2025. Docs = deduplicated documents; Dedup. = fraction of raw documents removed by deduplication; Struct. resources = non-document FHIR resources (lab values, medications, conditions, etc.); OCR rej. = fraction of documents rejected by OCR quality filtering.

4.1 FHIR Data Quality Analysis

We characterize the challenges clinical data poses for automated extraction across 10,000 patients ($\sim$1.2M deduplicated documents).

Encounter linkage and distribution. FHIR groups clinical activities into encounters, which could in principle organize a patient's documents by episode of care. In this export, encounters follow a three-level hierarchy (case, department, stay), but documents link exclusively to case-level encounters, the broadest administrative unit. Of these, 13.7% hold no documents at all, and the remainder are highly non-uniform: a single encounter holds a median of 47.5% of a patient's documents (Table 11), dropping to 14.7% for patients with 20+ encounters, which suggests the hierarchy distributes documents as intended (Appendix E). Yet at P99, coverage remains 53.5% even for patients with 20+ encounters, and the concentration index reaches 14.83 (Table 11): the most complex patients are precisely those where encounter structure fails to partition documents meaningfully. Linkage is also temporally imprecise: 56.5% of linked documents carry timestamps entirely outside their encounter's period (median delta 14.0 days). Heuristics may partially recover episode-level structure, but cannot guarantee reliable scoping, which is why we bypass encounter-based scoping altogether (§5).

Document quality and metadata. FHIR provides mechanisms for document relationships and unique identifiers but only recommends them: just 0.52% and 27.8% of documents carry them. Content-level deduplication removes a median 33.5% of documents per patient (up to 54.6%; Table 2). Metadata on the FHIR document reference is otherwise sparse: authorship appears for 1.9%, subtypes for 41.87%, structured conclusions for 0.45%. Provenance fields are better populated on other resource types (e.g., 97.5% on diagnostic reports; Appendix, Table 14), but these do not cover the full document corpus. Crucially, no document-level summary or abstract exists that a retrieval or agentic pipeline could use as a content preview to decide whether a document is worth reading. Over 1,000 document categories are used, many differing only in wording, and OCR rejection reaches 52.0% for the worst patient (median 10.3%; Table 2).

Timestamp reliability. Documents carry several metadata timestamps (report finalization, file creation, record update, encounter). The most clinically meaningful, the encounter timeframe, is absent from the export, and not propagated to the document reference. Even when inferred from the linked encounter resource, 56.5% of documents carry timestamps outside their encounter period. We therefore resolve each document's primary date from the available timestamps via a priority cascade. To assess whether the resolved timestamp reflects the actual clinical date, we compared it against the date extracted from document content via OCR and an LLM. When no date can be identified from the content, the system falls back to the resolved timestamp, so reported agreement is an upper bound. Only 58.8% agreed on the same day, and 36.5% diverged by more than one day. Agreement stays near 59% whichever field supplies the date (Appendix H), so no document-level timestamp reliably represents the clinical date or orders a patient's context in time.

Patient context scaling. Patient contexts span orders of magnitude: after deduplication, document counts range from 1 to over 2,500, structured data points from 0 to over 119,000, and document lengths from 24 to over 900,000 characters (Appendix D, Table 9). The top 1% (P99) define the

hardest cases any deployed system must handle without degradation. These patient contexts hold at least 937 documents and over 37,000 structured resources. Appendix F provides a breakdown by history length.

4.2 Clinical Study Extraction

We evaluate ACIE alongside an independent retrospective registry study of lymphoma patients undergoing molecular imaging (Appendix A). Its electronic case report form (eCRF), designed by two nuclear-medicine physicians and a hematologist, predates ACIE and was defined independently of it, so the extraction targets were not shaped by what the tool can do. A nuclear-medicine specialist with over four years of training configured all 74 AI-extracted fields (45 categorical, 9 numerical, 8 Boolean, 6 date, 3 free-text, 3 tabular; Appendix A), choosing their typed specifications and refining them on a subset of cohort patients, with no engineering effort. The fields cover clinical classification, immunohistochemical and molecular markers, longitudinal treatment and imaging history, and outcomes.

ACIE extracted these fields for 99 patients (7,326 values). Each value and its cited passages were verified by one of two nuclear-medicine physicians against the clinical systems used in routine work, so acceptance estimates verified correctness rather than agreement with a fixed key. Rejections are typed as *extraction errors* (wrong, fabricated, extraneous, or missed), *editorial adjustments* (acceptable value, but more or less detail wanted), or *form configuration* issues (Table 4). We report pooled rates alongside the patient-level distribution.

Reliability. The blended acceptance rate (7,073 of 7,326, 96.5%) combines two behaviours. Where the system committed to a value (4,440 fields) clinician-verified precision is 96.4%; where it returned nothing (2,886 fields) 96.8% were correct abstentions, leaving 92 in which a value did in fact exist (per type, Appendix C, Table 6). It thus neither produces extraneous values for fields that should be empty, a common LLM failure, nor misses present ones in bulk, and is consistent across patients (mean 96.5%, median 97.3%, range 82.4–100%; 78 of 99 patients $\geq$95%, 7 patients with no rejections).

Field type drives accuracy. Acceptance varies far more by data type than the 74-field count suggests (Table 3). The categorical, numerical, Boolean, and free-text fields are accepted at 96.0–98.6%, whereas the two weak types are dates at 84.3% and tabular fields at 79.8% (71.2% among non-empty tables), and the rejected cases show why. Tabular fields are rejected even when their evidence is directly stated in the sources: the difficulty is assembling the multi-row tabular timeline itself, with missing or guessed dates, dropped rows, and extraneous ones. Dates fail in two directions, when a value is returned the reviewer often judges it the wrong clinical event among several candidates, and when none is returned a value frequently did exist, making date the one type whose abstentions are unreliable (empty answers accepted only 69.8%, against 96.8% across types). The two hardest individual fields are of exactly these kinds, the date of death or last follow-up (accepted for 55 of 99 patients) and the treatment-timeline table (59 of 99). The driver of error is thus how much temporal reasoning the answer demands over the full record, assembling a timeline for tables or selecting the right event for dates.

Field type	n	**Acc.%**	**Empty%**
Categorical	4,455	98.6	35.1
Numerical	891	98.3	76.1
Boolean	792	98.6	21.8
Free text	297	96.0	59.9
Date	594	84.3	34.0
Tabular	297	79.8	31.0
Total	**7,326**	**96.5**	**39.4**

Table 3: Clinician acceptance by field type. "Empty%" is the share of fields where the system returned no value, typically a legitimately absent value.

Errors and safety. Of 253 rejections (Table 4), 241 are extraction errors, 3 are editorial adjustments, and 9 are form configuration issues (e.g., dropdown options not matching clinical reality) rather than extraction failures; excluding the latter, extraction-attributable acceptance is 96.7%. By direction of failure, 161 rejections correct a value the system produced and 92 supply one it left empty (overwhelmingly dates). The errors are also highly concentrated: ten of the 74 fields account for 182 of the 253 rejections (Appendix C), led by the death-or-last-follow-up date and the treatment timeline. No value was hallucinated, i.e. produced without support from any cited passage. In a single case the system returned a value where it should have abstained (one extraneous extraction). The dominant residual risk is thus a wrong or missing value that a reviewer corrects, not an invented one.

Rejection category	Count (%)
Extraction errors	
Incorrect value (needs correction)	221 (87.4)
Fully incorrect (replaced)	7 (2.8)
Missed extraction (false negative)	9 (3.6)
Missing reference	3 (1.2)
Extraneous extraction (false positive)	1 (0.4)
Hallucinated	0 (0.0)
Editorial adjustments	
Missing information	2 (0.8)
Excess information	1 (0.4)
Form configuration	
Configuration error	9 (3.6)

Table 4: The 253 rejected fields by category (% of all rejections); category definitions in Appendix B. Each rejected field carried exactly one flag. Extraction errors dominate; no content was hallucinated.

5 Lessons from Deployment

We report two lessons from deploying ACIE on real clinical data, quantifying the specific mechanisms we encountered.

L1: Clinical data quality falls far short of what AI systems require.

Prior work documents clinical data quality challenges (Vorisek et al., 2022) and heterogeneity across institutions (Palm et al., 2025). §4.1 quantifies this gap from the perspective of an AI system that must retrieve and extract based on this metadata. The deployment site operates a large-scale clinical FHIR repository, yet document-level metadata remains sparse: fields like encounter periods, document relationships, and authorship are absent from the document reference or too sparsely populated for filtering. Clinicians navigate this sparsity through institutional knowledge, but AI systems cannot. In a large primary care database, only 13% of clinical concepts in free-text notes had structured counterparts in coded fields (Seinen et al., 2025), and where timestamps are populated, they are not necessarily correct (§4.1). These gaps are likely not site-specific: independent single-site studies document data quality problems of similar severity in other settings, from pervasive duplication in US clinical notes (Steinkamp et al., 2022) to decades of erroneous administrative entries at a regional German hospital (Förstel et al., 2024). We hypothesize that this reflects data infrastructure historically optimized for billing rather than clinical coherence. Any system deployed on such data must compensate architecturally.

L2: Architectural decisions shaped by data.

The data quality gaps in §4.1 shaped three design decisions in ACIE's architecture.

Agentic retrieval over static filtering. We first deployed a retrieve-then-generate pipeline with metadata-based filters (encounter-scoped retrieval, date-range and category filters). Each filter depended on metadata that §4.1 shows is unreliable or absent. After exhausting static filter combinations, we concluded that reliable retrieval requires an agent that reasons about which documents matter based on content, not a pipeline that filters by metadata.

Query-relevant document summaries. To make content-based triage tractable over hundreds of documents, the agent cannot read every document in full. Instead, it previews documents through the query-relevant summaries of §3.2 and decides from content whether full inspection is warranted. Fine-grained chunking compounds the problem by producing fragments that vary by orders of magnitude in length, which dense retrievers systematically favor when short (Fayyaz et al., 2025). The length penalty in Equation 1 corrects this bias.

Markdown over JSON serialization. Data serialization affects extraction quality for smaller models (Pator, 2026), and output format constraints degrade reasoning (Tam et al., 2024). We encountered a related failure on the *input* side: when patient metadata was presented in JSON format, specific patients consistently triggered malformed tool calls. The failures were deterministic and patient-specific, suggesting that particular JSON structures from heterogeneous clinical metadata interfered with tool-calling. Switching to markdown eliminated all failures.

6 Conclusion

Clinical IE research typically treats the data as given and asks how capable the model is. Deploying ACIE taught us the inverse: the data dictates the architecture. The document-level metadata needed for retrieval and triage is largely absent, unreliable or not propagated to the document level even at a site with large-scale FHIR integration, so retrieval cannot filter by structure, and reasoning must move into the retrieval loop. Evaluated alongside an independent lymphoma registry study whose extraction targets predate the system, ACIE reached 96.5% acceptance with no hallucinated content. The residual errors were governed by the temporal reasoning

each target demanded. Grounding every value in source passages shifts the clinician's role from compiling to verifying. Extraction can run in batches outside working hours, and the study physicians reported roughly three times faster completion per patient. These metadata gaps are well-documented across institutions (Steinkamp et al., 2022; Förstel et al., 2024), and we expect any system deployed on real hospital data to face similar constraints. Deployed clinical extraction therefore rests on two supports: architectures that reason over content, and human verification of grounded outputs.

Limitations

Our evaluation is a single retrospective study, one disease area, one hospital, one language, and 99 patients, so generalization to other settings is untested, and each field was graded by a single expert reviewer, leaving inter-rater reliability unmeasured. The retrospective setting is more permissive than point-of-care use, where a value directly informs an intervention rather than populating a research cohort; the headline acceptance also reflects the large share of fields whose correct answer is legitimately absent (39.4%), with precision on produced values at 96.4% and lower on the weakest types (84.3% dates, 79.8% tables). The absence of flagged hallucinations was judged by the same reviewers under source-grounded review rather than by independent adjudication of every passage. We did not compare against a non-agentic or commercial baseline, so we do not isolate the contribution of the agentic design, and the data-quality findings (§4.1) come from a single hospital's FHIR repository, so the specific rates may differ elsewhere. Extraction quality also remains bounded by the on-premise model.

Ethical Considerations

ACIE runs entirely on-premise, so patient data never leaves the hospital network. It is assistive, not autonomous: every value is grounded in cited passages and must be verified by a clinician before it enters documentation. This mitigates but does not remove the risk of automation bias, where a frictionless interface invites uncritical acceptance; mandatory source review, and the rejection workflow we evaluate, are the safeguard. Extraction quality may degrade for patient groups underrepresented in the records or in the underlying model, a risk clinician verification is intended to catch. The registry study was conducted in accordance with the applicable institutional and regulatory requirements for the retrospective use of clinical data at University Medicine Essen.

In line with the ACL policy on AI writing assistance, an AI assistant was used for language editing and LaTeX formatting only; all research content and claims originate from the authors, who take full responsibility for the final text.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022. Association for Computational Linguistics.

Yaara Artsi, Vera Sorin, Benjamin S. Glicksberg, Panagiotis Korfiatis, Girish N. Nadkarni, and Eyal Klang. 2025. Large language models in real-world clinical workflows: A systematic review of applications and implementation. *Frontiers in Digital Health*, 7:1659134.

Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, Yue Zhang, Yubo Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. 2026. PaddleOCR-VL-1.5: Towards a multi-task 0.9b VLM for robust in-the-wild document parsing. *arXiv preprint arXiv:2601.21957*.

Fabio Dennstädt, Janna Hastings, Paul Martin Putora, Max Schmerder, and Nikola Cihoric. 2025. Implementing large language models in healthcare while balancing control, collaboration, costs and security. *npj Digital Medicine*, 8(1):143.

Mohsen Fayyaz, Ali Modarressi, Hinrich Schuetze, and Nanyun Peng. 2025. Collapse of dense retrievers: Short, early, and literal biases outranking factual evidence. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9136–9152, Vienna, Austria. Association for Computational Linguistics.

Stefan Förstel, Markus Förstel, Markus Gallistl, Dario Zanca, Bjoern M. Eskofier, and Eva M. Rothgang. 2024. Data quality in hospital information systems: Lessons learned from analyzing 30 years of patient data in a regional German hospital. *International Journal of Medical Informatics*, 192:105636.

Maxime Griot, Jean Vanderdonckt, and Demet Yuksel. 2025. Implementation of large language models in electronic health records. *PLOS Digital Health*, 4(12):e0001141.

Aliće Grünig, Jenifer Kriebel, Julian Varghese, Tim Herrmann, Sarah Sandmann, and Christian Bruns.

2026. Implementation and user evaluation of an on-premise large language model in a German university hospital setting: Cross-sectional survey. *JMIR AI*, 5:e84362.

Shashi Kant Gupta, Arijeet Pramanik, Jerrin John Thomas, Regina Schwind, Lauren Wiener, Avi Raju, Jeremy Kornbluth, Yanshan Wang, Zhaohui Su, and Hrituraj Singh. 2025. HARMON-E: Hierarchical agentic reasoning for multimodal oncology notes to extract structured data. *arXiv preprint arXiv:2512.19864*.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Özlem Uzuner. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

HL7. 2019. Fast healthcare interoperability resources (FHIR) release 4. https://hl7.org/fhir/R4/. Accessed: 2026-06-15.

Jennifer Jiang-Kells, James Brandreth, Leilei Zhu, Jack Ross, Yogini Jani, Enrico Costanza, Maisarah Amran, Zeljko Kraljević, Xi Bai, M.M.N.S. Dilan, Jayathri Wijayarathne, Ravi Wickramaratne, Folkert W. Asselbergs, Richard J.B. Dobson, Wai Keong Wong, and Anoop D. Shah. 2025. Design and implementation of a natural language processing system at the point of care: MiADE (medical information AI data extractor). *BMC Medical Informatics and Decision Making*, 25(1):365.

Zeljko Kraljević, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A. Folarin, Angus Roberts, Rebecca Bendayan, Mark P. Richardson, Robert Stewart, Anoop D. Shah, Wai Keong Wong, Zina Ibrahim, James T. Teo, and Richard J.B. Dobson. 2021. Multi-domain clinical natural language processing with MedCAT: The medical concept annotation toolkit. *Artificial Intelligence in Medicine*, 117:102083.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Yusheng Liao, Shuyang Jiang, Yanfeng Wang, and Yu Wang. 2025. ReflecTool: Towards reflection-aware tool-augmented clinical agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13507–13531, Vienna, Austria. Association for Computational Linguistics.

Kevin Lybarger, Meliha Yetisgen, and Özlem Uzuner. 2023. The 2022 n2c2/UW shared task on extracting social determinants of health. *Journal of the American Medical Informatics Association*, 30(8):1367–1378.

Yiqun Miao, Yuhan Zhao, Yuan Luo, Huiying Wang, and Ying Wu. 2025. Improving large language model applications in the medical and nursing domains with retrieval-augmented generation: Scoping review. *Journal of Medical Internet Research*, 27(1):e80557.

Sungrim Moon, Luke A. Carlson, Ethan D. Moser, Bhavani Singh Agnikula Kshatriya, Carin Y. Smith, Walter A. Rocca, Liliana Gazzuola Rocca, Suzette J. Bielinski, Hongfang Liu, and Nicholas B. Larson. 2022. Identifying information gaps in electronic health records by using natural language processing: Gynecologic surgery history identification. *Journal of Medical Internet Research*, 24(1):e29015.

Julia Palm, Kutaiba Saleh, André Scherag, and Danny Ammon. 2025. Leveraging interoperable electronic health record (EHR) data for distributed analyses in clinical research: Technical implementation report of the HELP study. *JMIR Medical Informatics*, 13(1):e68171.

Sanjoy Pator. 2026. Serialisation strategy matters: How FHIR data format affects LLM medication reconciliation. *arXiv preprint arXiv:2604.21076*.

Qwen Team. 2026. Qwen3.6-35B-A3B: Agentic coding power, now open to all. Model card: https://huggingface.co/Qwen/Qwen3.6-35B-A3B. Accessed: 2026-06-11.

Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. 2010. Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

William Scuba, Melissa Tharp, Danielle Mowery, Eugene Tseytlin, Yang Liu, Frank A. Drews, and Wendy W. Chapman. 2016. Knowledge author: Facilitating user-driven, domain content development to support clinical information extraction. *Journal of Biomedical Semantics*, 7:42.

Tom M. Seinen, Jan A. Kors, Erik M. van Mulligen, and Peter R. Rijnbeek. 2025. Using structured codes and free-text notes to measure information complementarity in electronic health records: Feasibility and validation study. *Journal of Medical Internet Research*, 27:e66910.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamber, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Blaise

Aguera y Arcas, and 12 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Jackson Steinkamp, Jacob J. Kantrowitz, and Subha Airan-Javia. 2022. Prevalence and sources of duplicate information in the electronic medical record. *JAMA Network Open*, 5(9):e2233348.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. 2024. Let me speak freely? A study on the impact of format restrictions on large language model performance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1218–1236, Miami, Florida, US. Association for Computational Linguistics.

Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Carina Nina Vorisek, Moritz Lehne, Sophie Anne Ines Klopfenstein, Paula Josephine Mayer, Alexander Bartschke, Thomas Haese, and Sylvia Thun. 2022. Fast healthcare interoperability resources (FHIR) for interoperability in health research: Systematic review. *JMIR Medical Informatics*, 10(7):e35724.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. 2018. Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, 77:34–49.

Isabella Catharina Wiest, Dyke Ferber, Jiefu Zhu, Marko van Treeck, Sonja K. Meyer, Radhika Juglan, Zunamys I. Carrero, Daniel Paech, Jens Kleesiek, Matthias P. Ebert, Daniel Truhn, and Jakob Nikolas Kather. 2024. Privacy-preserving large language models for structured medical information retrieval. *npj Digital Medicine*, 7:257.

Isabella Catharina Wiest, Fabian Wolf, Marie-Elisabeth Lessmann, Marko van Treeck, Dyke Ferber, Jiefu Zhu, Heiko Boehme, Keno K. Bressem, Hannes Ulrich, Matthias P. Ebert, and Jakob Nikolas Kather. 2025. A software pipeline for medical information extraction with large language models, open source and suitable for oncology. *npj Precision Oncology*, 9:313.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. 2025. Improving retrieval-augmented generation in medicine with iterative follow-up questions. In *Pacific Symposium on Biocomputing*, volume 30, pages 199–214.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B. Costa, Mona G. Flores, Ying Zhang, Tanja Magoc, Christopher A. Harle, Gloria Lipori, Duane A. Mitchell, William R. Hogan, Elizabeth A. Shenkman, Jiang Bian, and Yonghui Wu. 2022. A large language model for electronic health records. *npj Digital Medicine*, 5:194.

Zongxin Yang, Hongyi Yuan, Raheel Sayeed, Amelia Li Min Tan, Enci Cai, Mohammed Moro, Xiudi Li, Huaiyuan Ying, Nicholas Brown, Griffin Weber, Sheng Yu, Isaac Kohane, and Tianxi Cai. 2025. CLINES: Clinical LLM-based information extraction and structuring agent. *medRxiv*. Preprint.

Liang Yao, Harry Hochheiser, Wonjin Yoon, Shyam Goldner, and Guergana Savova. 2024. Overview of the 2024 shared task on chemotherapy treatment timeline extraction. In *Proceedings of the 6th Clinical NLP Workshop*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*.

A Clinical Study Extraction Schema

Table 5 lists the 74 AI-extracted fields of the lymphoma registry study, grouped by data type; the type determines how the agent is prompted and how its output is validated. Four further demographic fields are read directly from FHIR and are excluded from the evaluation in §4.2. The eCRF was designed by two nuclear-medicine physicians and a lymphoma hematologist, and it spans the full disease trajectory, from diagnosis and histopathology through molecular characterization, treatment, imaging follow-up, and outcomes. Related single-marker fields are grouped into one row where they share a common form; every field name is listed explicitly.

B Rejection Categories

When a reviewer rejects an extracted value, they assign a category describing the nature of the problem. The categories fall into three groups, mirroring Table 4.

Extraction errors mean the value is factually wrong, unsupported, or absent. *Incorrect value*: the value must be partially corrected. *Fully incorrect*: the value is wrong and must be replaced wholesale, though it remains attributed to a cited passage, which distinguishes it from a hallucination. *Outdated*: a once-correct value that is no longer current. *Missed extraction*: the system returned nothing although the information is present in the sources. *Extraneous extraction*: the converse, a value was produced for a field that should have

Type (n)	Fields
Categorical (45)	
WHO classification & subtyping (9)	Primary lymphoma category; entity-specific subtype for LBCL, DLBCL cell-of-origin, follicular, Hodgkin, mantle-cell, marginal-zone, Burkitt, and peripheral T-/NK-cell lymphoma
Histology & qualitative IHC (9)	Biopsy sites; necrosis; positive/negative immunohistochemistry for p53, PD-L1 (tumour cells), MYC, BCL2, CD20, CD30, and EBV (EBER ISH)
Cytogenetics & FISH (9)	Cytogenetics performed; karyotype available; complex karyotype; FISH status; MYC / BCL2 / BCL6 rearrangement; del(17p)/TP53; 9p24.1 amplification
NGS mutation status (17)	Overall NGS status; TP53, MYD88 (L265P), NOTCH1, NOTCH2, EZH2, CD79B, ARID1A, MEF2B, EP300, FOXO1, CREBBP, CARD11, RHOA (G17V), TET2, DNMT3A, IDH2 (R172)
Outcome (1)	Overall survival event
Numerical (9)	Height; weight at diagnosis; months from ASCT to relapse or progression; Ki-67 index; IHC expression percentage for p53, PD-L1 (tumour cells), PD-L1 (immune cells), MYC, and BCL2
Boolean (8)	Significant comorbidities (CIRS); B-symptoms; progression event; relapse event; prior ASCT before relapse/progression; histological transformation; bone-marrow involvement; large-cell component
Date (6)	Diagnosis; most recent biopsy; transformation; progression or last follow-up; relapse or last follow-up; death or last follow-up
Free text (3)	Original pathology-report diagnosis; extranodal site specification; biopsied body-fluid specification
Tabular (3)	Treatment timeline (each therapy line, transplant, surgery, and radiation with dates and cycles); PET examinations (dates, indication, tracer); MRD assessments (method, result)

Table 5: The 74 AI-extracted fields of the lymphoma study schema, grouped by data type. Parenthesized counts give the number of fields. Related single-marker categorical fields are summarized by group; every field name is listed.

been left empty. *Hallucinated*: content not supported by the cited source. *Missing reference*: the value or a crucial detail is given without attribution to a source passage.

Editorial adjustments mean the value is acceptable but the amount or form of information differs from what the field wanted. *Missing information*: available, relevant detail was omitted. *Excess information*: more was returned than the field asked for. *Reformatting*: the value is correct but formatted incorrectly.

Form configuration covers rejections that reflect the form rather than the extraction. *Configuration error*: the field's options or definition did not match clinical reality, a form-design issue rather than an extraction failure.

	Returned a value		Returned empty	
Field type	n	Acc.%	n	Acc.%
Categorical	2,892	98.3	1,563	99.1
Numerical	213	98.1	678	98.4
Boolean	619	98.4	173	99.4
Free text	119	93.3	178	97.8
Date	392	91.8	202	69.8
Tabular	205	71.2	92	98.9
Total	**4,440**	**96.4**	**2,886**	**96.8**

Table 6: Acceptance split by whether the system returned a value or abstained. Tabular fields fail when they produce a value (71.2%) but abstain reliably; dates are the reverse, abstaining unreliably (69.8%). All other types are strong in both modes.

C Error Analysis

This appendix breaks the 253 rejections down three ways. Table 6 separates each field type's acceptance into the case where the system returned a value (precision) and the case where it abstained (abstention reliability). Table 7 locates each rejection category within the field types. Table 8 shows how concentrated the errors are: ten of the 74 fields account for 182 of the 253 rejections.

D Per-Patient Context Distributions

Table 9 provides the full distribution of per-patient context statistics. The mean consistently exceeds the median across all dimensions, reflecting heavy right skew: a minority of patients accumulate disproportionately large contexts. The P1/P99 columns bound the central 98% of patients. The gap from P99 to Max defines the hardest cases the

Field type	Rej.	Incorrect	Fully incorr.	Miss. info	Excess	Missed	Extran.	Miss. ref.	Config.
Categorical	62	48	1	1	–	3	1	1	7
Numerical	15	12	–	–	–	2	–	1	–
Boolean	11	6	2	1	–	–	–	–	2
Date	93	89	2	–	–	1	–	1	–
Free text	12	11	–	–	–	1	–	–	–
Tabular	60	55	2	–	1	2	–	–	–
Total	**253**	**221**	**7**	**2**	**1**	**9**	**1**	**3**	**9**

Table 7: Rejection category by field type (counts). Categories not triggered anywhere (hallucinated, outdated, reformatting) are omitted; definitions in Appendix B. Configuration errors occur only in categorical and Boolean fields (form-option mismatches); dates are almost entirely incorrect-value.

Field	Type	Rej.
Date of death / last follow-up	Date	44
Treatment timeline	Tabular	40
Date of progression / last PET	Date	22
Overall survival event	Categorical	16
MRD assessments	Tabular	15
Weight at diagnosis	Numerical	12
Date of relapse / last PET	Date	11
Biopsy sites	Categorical	8
Diagnosis date	Date	7
Extranodal site specification	Free text	7

Table 8: The ten most-rejected fields (each evaluated on all 99 patients), accounting for 182 of the 253 rejections. Date and table fields dominate the error mass.

agentic framework must handle. The top 1% of patients hold at least 937 documents and 37,074 structured resources, reaching up to 2,542 and 119,191 respectively. The history length maximum (739,726 days) reflects corrupt timestamps in the source data. P99 (9,775 days, ∼26.8 years) provides a more realistic upper bound.

E Encounter Coverage by Patient Complexity

Table 10 shows how top-encounter coverage varies with the number of encounters per patient. For patients with few encounters ($\leq$5), a single encounter typically holds all documents. As encounters increase, coverage disperses: for patients with 20+ case-level encounters, the median drops to 14.7%, suggesting the hierarchy distributes documents across episodes as intended. Yet at P99, a single encounter still holds 53.5% of documents. The concentration index (Table 11) confirms this pattern, reaching 14.83 at P99: the patients with the most complex records are precisely those where documents cluster most unevenly. Encounter-based scoping would therefore degrade for exactly the patients that depend most on thorough retrieval.

Table 11 gives the full distribution of top-encounter coverage and of the concentration index used in §4.1.

F Patient History Length

Tables 12 and 13 break down document counts and total FHIR resources by patient history length. Both show heavy right skew, with the gap from P99 to Max again highlighting the extreme cases the system must handle.

G Metadata and Timestamp Population

Tables 14 and 15 report the population rates of document-level metadata and timestamp fields across the two FHIR resource types that carry clinical documents. These rates quantify the metadata sparsity discussed in §4.1.

H Timestamp Cross-Validation

Table 16 compares the FHIR metadata timestamp resolved for each document against the clinical date extracted from the document content via OCR and an LLM (n=15,142 documents). The alignment is consistent at roughly 59% regardless of which FHIR field provides the resolved date, confirming that no single metadata timestamp reliably represents when clinical activity occurred.

	Min	P1	Q1	Med.	Q3	P99	Max	Mean
Raw docs	1	2	22	77	193	1,269	3,029	164.6
Deduped	1	1	14	52	140	937	2,542	120.4
Encounters	0	3	6	18	46	323	1,207	42.0
Structured	0	1	38	406	1,922	37,074	119,191	2,804.1
History (days)	0	15	330	1,305	4,145	9,775	739,726	2,699.7
Doc. length (chars)[†]	24	–	1,387	2,915	11,035	–	907,179	10,788

Table 9: Per-patient context statistics (n=10,000). "Structured" counts all non-document FHIR resources (lab values, medications, conditions, etc.). Table 2 in the main text shows the compressed version. [†]The last row is per document, not per patient, over the OCR-processed documents (P1/P99 unavailable); document length spans four orders of magnitude, so a fixed chunking or truncation budget cannot serve both ends.

Case enc.	n	Min	P1	Q1	Median	Q3	P99	Max	Mean
1	1,930	9.5	42.9	100.0	100.0	100.0	100.0	100.0	96.6
2–5	3,334	2.1	23.8	46.7	59.0	75.0	100.0	100.0	61.3
6–20	3,099	6.9	12.0	24.3	32.6	44.0	80.2	100.0	35.5
20+	1,637	2.5	3.9	10.0	14.7	21.1	53.5	85.5	17.1

Table 10: Top-encounter document coverage (%) by number of case-level encounters per patient. Shows the percentage of a patient's documents held by their single busiest encounter. Even for patients with 20+ encounters, P99 coverage remains 53.5%, indicating persistent clustering.

	Min	P1	Q1	Med.	Q3	P99	Max	Mean
Top-encounter coverage (%)	2.1	6.3	26.7	47.5	80.0	100.0	100.0	52.9
Concentration index	0.04	0.73	1.27	2.24	3.75	14.83	105.1	3.07

Table 11: Per-patient distribution of top-encounter document coverage (n=10,000) and the concentration index (n=9,957 patients with at least one case-level encounter). The concentration index is the ratio of the top encounter's document share to the uniform expectation across case-level encounters; 1.0 indicates even distribution.

History	n	Min	P1	Q1	Med.	Q3	P99	Max	Mean
$<$ 1 yr	2,709	1	1	4	10	26	190	497	23.7
1–3 yr	1,995	1	2	18	49	102	572	1,476	86.0
3–5 yr	934	2	6	34	82	162	777	1,360	132.6
5–10 yr	1,602	1	6	50	107	235	1,217	2,542	187.9
10+ yr	2,760	1	3	43	114	252	1,174	2,439	196.8

Table 12: Deduplicated documents per patient by history length (n=10,000).

History	n	Min	P1	Q1	Med.	Q3	P99	Max	Mean
$<$ 1 yr	2,709	1	5	17	52	291	13,040	52,462	711.4
1–3 yr	1,995	6	8	105	490	1,645	29,889	102,412	2,241.7
3–5 yr	934	12	20	225	931	2,820	30,848	83,760	2,996.0
5–10 yr	1,602	5	19	352	1,177	4,061	47,994	100,673	4,288.7
10+ yr	2,760	6	18	365	1,303	4,939	48,929	121,542	5,086.6

Table 13: Total FHIR resources per patient by history length (n=10,000). Includes all non-document resources (lab values, medications, conditions, encounters, etc.).

Resource	Metadata Field	Pop. %
Identification and relationships		
DocRef	Unique identifier	27.8
DocRef	Related documents	0.52
Authorship and provenance		
DocRef	Author	1.9
DocRef	Authenticator	16.2
DocRef	Custodian	100.0
DiagRep	Performer	27.1
DiagRep	Results interpreter	97.5
Content description		
DocRef	Description	99.7
DocRef	Attachment title	0.0
DocRef	Content format	98.1
DiagRep	Title	60.6
DiagRep	Structured conclusion	0.45

Table 14: Document metadata population rates. DocRef: DocumentReference (n=636,534); DiagRep: DiagnosticReport (n=567,110). "Pop. %" is the fraction of resources where the field is non-empty.

Resource	Timestamp Field	Pop. %
DocRef	Report date	99.99
DocRef	File creation date	100.0
DocRef	Encounter period	0.0
DocRef	Release date	76.5
DocRef	Print date	24.7
DocRef	Record last updated	100.0
DiagRep	Effective date	95.2
DiagRep	Issued date	78.0
DiagRep	File creation date	97.7
DiagRep	Record last updated	100.0

Table 15: Timestamp field population rates. DocRef: DocumentReference (n=636,534); DiagRep: DiagnosticReport (n=567,110). "Pop. %" is the fraction of resources where the field is non-empty.

FHIR Field	n	Same day	$\pm$1 day	$>$1 day
All fields	15,142	58.8%	4.8%	36.5%
Report date	8,037	58.6%	5.2%	36.2%
Effective date	6,240	59.0%	3.7%	37.3%
Issued date	859	59.1%	7.9%	32.9%
File creation	6	50.0%	0.0%	50.0%

Table 16: Alignment between FHIR metadata timestamps and clinical dates extracted from document content. Rows show the FHIR field that provided the resolved date for each document. Overall same-day agreement: 58.8%.