RWD36

ontada

Accelerating EHR Insights: NLP-Driven Data Abstraction in Gallbladder Cancer

Alisha Monnette Kimble, Avi Raju, Gayathri Namasivayam, Junxin Shi, Wendy Haydon, Wanmei Ou, Robert Reid
Author Affiliations: Ontada, Boston, MA, USA

The US Oncology Network

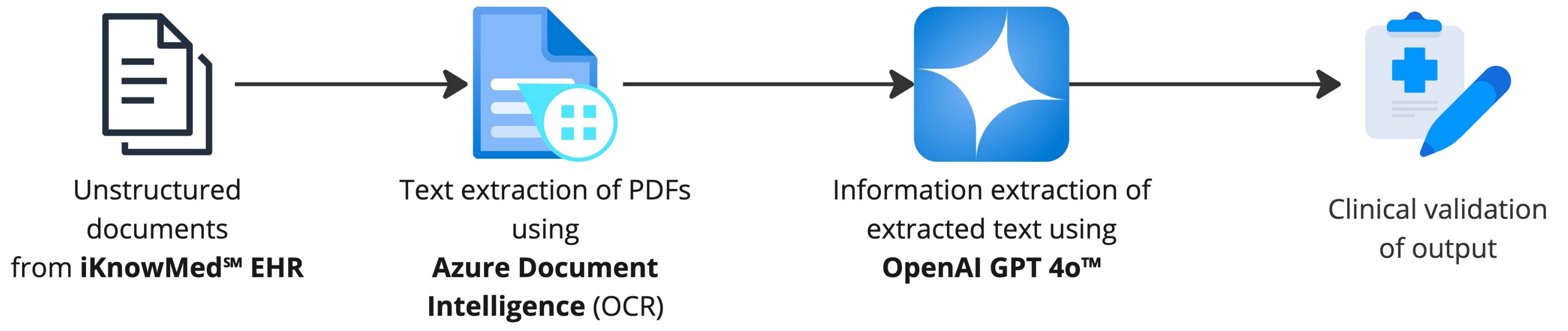
Background

- Gallbladder cancer (GBC) is a rare and aggressive malignancy with limited treatment options
- Accurate staging and histological classification are essential for guiding care and research
- However, TNM staging and histologic subtype are often missing from structured fields in electronic health records (EHRs)
- To improve data completeness and accuracy, we utilized natural language processing (NLP) to extract these variables from unstructured EHR documents

Methods

- NLP was used to extract TNM values and histology data from unstructured clinical notes and scanned pathology reports within iKnowMed (iKM), an EHR system used by US Oncology-affiliated clinics
- TNM values were sourced from progress notes while histology data was extracted from pathology reports using optical character recognition and large language models (Azure Document Intelligence and OpenAI GPT-4o)
- In a development set, the model achieved F1 scores of 0.85 for individual TNM values and 0.83 for histology
  - In NLP and named entity recognition, the F1 score measures how well a model identifies and classifies entities by balancing precision (correctly predicted entities) and recall (all actual entities), providing a single metric for overall performance.

Figure 1: NLP pipeline workflow to extract clinical data


Unstructured documents from **iKnowMed℠ EHR**
Text extraction of PDFs using **Azure Document Intelligence** (OCR)
Information extraction of extracted text using **OpenAI GPT 4o™**
Clinical validation of output

Citation:
1. SEER*Explorer: An interactive website for SEER cancer statistics [Internet]. Surveillance Research Program, National Cancer Institute; 2024 Apr 17. [updated: 2024 Nov 5; cited 2025 Mar 25]. Available from: https://seer.cancer.gov/statistics-network/explorer/.

Results

- 2,019 patients with GBC were identified between 2014 and 2022:
  - Missing all TNM values - 51.3% (N=1,035)
  - Missing histology - 56.7% (N=1,144)
- NLP identified at least one value, within 90 days of diagnosis, for:
  - TNM - 208 of 1,035 patients (20%)
  - Histology - 771 of 1,144 patients (67.4%)

Table 1: Variable completion rate before and after NLP

	Variable completion rate	
	Before NLP	**After NLP**
TNM	48.7%	59%
Histology	43.3%	81.5%

- Histology distributions aligned more closely to national averages and published literature
  - Majority in this dataset (71.6%) had adenocarcinoma (vs. reference SEER 76%) **(Table 2)**

Strengths and Limitations

- NLP substantially improved histology completeness and modestly increased TNM availability.
- The smaller TNM gains may reflect multiple challenges:
  - TNM elements are often not documented in clinical text, some pathology reports were unavailable or unreadable, and certain components (like N and M) may be harder for NLP to extract.

Conclusions

- NLP improved data completeness for staging and histology, supporting more comprehensive cohort development for research.
- While gains varied by variable, these results highlight both the promise and current limitations of NLP in rare cancer data enhancement.

Table 2: Descriptive disease characteristics before and after NLP

Variable	Before NLP	After NLP
Total patient count	2019	2019
T Staging at diagnosis, n (%)		
T0	10 (0.5%)	11 (0.5%)
T1	89 (4.4%)	136 (6.7%)
T2	392 (19.4%)	556 (27.5%)
T3	348 (17.2%)	456 (22.6%)
T4	61 (3.0%)	70 (3.5%)
TX	127 (6.3%)	134 (6.6%)
Tis	20 (1.0%)	28 (1.4%)
Not documented	**972 (48.1%)**	**628 (31.1%)**
N Staging at diagnosis, n (%)		
N0	370 (18.3%)	496 (24.6%)
N1	254 (12.6%)	330 (16.3%)
N2	128 (6.3%)	138 (6.8%)
NX	271 (13.4%)	386 (19.1%)
Not documented	**996 (49.3%)**	**669 (33.1%)**
M Staging at diagnosis, n (%)		
M0	591 (29.3%)	678 (33.6%)
M1	423 (21.0%)	462 (22.9%)
MX	49 (2.4%)	124 (6.1%)
Not documented	**956 (47.4%)**	**755 (37.4%)**
TNM Staging at diagnosis		
Not documented (all TNM values missing)	**1035 (51.3%)**	**827 (41.0%)**
Histology at diagnosis, n (%)		
Adenocarcinoma	736 (36.5%)	1446 (71.6%)
Adenocarcinoma NOS	*639 (31.6%)*	*1186 (58.7%)*
Biliary type adenocarcinoma	*2 (0.1%)*	*114 (5.6%)*
Intestinal type adenocarcinoma	*76 (3.8%)*	*98 (4.9%)*
Mixed intestinal/mucinous	*1 (0.1%)*	*4 (0.2%)*
Mucinous carcinoma	*18 (0.9%)*	*44 (2.2%)*
Adenosquamous carcinoma	0	19 (0.9%)
Clear cell carcinoma	32 (1.6%)	33 (1.6%)
Other	93 (4.6%)	118 (5.8%)
Sarcomatoid carcinoma (carcinosarcoma)	1 (0.1%)	3 (0.1%)
Signet ring cell carcinoma	13 (0.6%)	27 (1.3%)
Not documented	**1144 (56.7%)**	**373 (18.5%)**

Acknowledgements

The authors thank the investigators, study team, and the patients and their families at each of the practices.

Correspondence: Alisha Monnette Kimble
alisha.monnette@mckesson.com