Automating cancer registry abstraction with an autonomous, resource-efficient AI for multi-cancer pathology reports: a model development and validation study

Nan-Hua Chou, M.D., Ph.D.[1+], Han Chang, M.D., Ph.D.[2,3+], Hung-Kai Chen[2], Chen-Yuan Lin, B.S.[4], Ying-Lung Liu, B.S.[5], Po-Yen Tseng[2], Li-Chu Hsu, Ph.D.[6], Yen-Wei Chu, Ph.D.[7], Kai-Po Chang, M.D., Ph.D.[2,3]*

[1] Center for Precision Medicine, China Medical University Hospital, Taichung, Taiwan

[2] School of Medicine, China Medical University, Taichung, Taiwan

[3] Department of Pathology, China Medical University Hospital, Taichung, Taiwan

[4] Signal1, Toronto, Canada

[5]Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan

[6]Department of Foreign Languages and Literature, National Chi Nan University, Nantou, Taiwan

[7]Graduate Institute of Genomics and Bioinformatics, National Chung Hsing University, Taichung, Taiwan

[+]These authors contributed equally to this work

*** Corresponding Author:**

Kai-Po Chang

017179@tool.caaumed.org.tw

+886-4-2205-2121 #12678

2 Yude Rd., Taichung, Taiwan

Abstract

Background The manual abstraction of data from unstructured pathology reports for cancer registries is a major bottleneck in oncology informatics. While Large Language Models (LLMs) show promise, their adoption is hindered by computationally prohibitive resource requirements and data privacy concerns. There is an urgent need for an accessible, robust, and autonomous system deployable at the front lines of healthcare.

Methods We developed a fully autonomous, multi-stage AI workflow that operates on a single professional-grade GPU. The system first deploys a detector that triages raw clinical documents to identify cancer surgical reports requiring cancer registry entry, and a classifier that routes pathology reports to one of ten cancer-specific information extraction engines. This privacy-preserving pipeline is powered by a locally deployed open-weight LLM (gpt-oss:20b) and programmed via the DSPy framework, a self-optimizing framework for LLMs. The end-to-end system was validated on 893 uncurated, real-world cancer-surgical reports, with a gold standard generated by two board-certified pathologists.

Findings The triage detector achieved an accuracy of 96.6% and an F1 score of 97.9% on cancer surgical reports detection, and the cancer-type classifier routed reports with an accuracy of 97.8% and a macro average F1 score of 95.5%. Across ten distinct malignancies, the pipeline achieved a mean exact-match accuracy of 93.9%±8.6% for 196 registry fields. Critical oncologic information, including margin positivity and key

biomarkers for breast cancer, were extracted with acceptable fidelity (90–100%). The entire autonomous process was executed at an acceptable speed.

Interpretation This work provides a deployable, resource-efficient blueprint for a "digital registrar" robust enough for real-world clinical workflows. This marks a pivotal step toward transforming global cancer surveillance from a slow, retrospective process into a near real-time system.

Funding National Science and Technology Council (NSTC) of Taiwan

1 Introduction

Population-based cancer registries are the cornerstone of oncology surveillance and public health policy[1], yet their operational efficacy is fundamentally constrained by the manual abstraction of data from unstructured pathology reports[2]. This process is notoriously laborious, error-prone, and unscalable, creating data availability delays that limit the timeliness of research and public health interventions. For decades, this bottleneck has awaited automated solutions[3,4]. Early Natural Language Processing (NLP) methods, including rule-based systems and traditional machine learning models, offered partial solutions but struggled with the linguistic heterogeneity of clinical text and lacked the generalizability required for diverse cancer types and reporting styles.[5,6]. Subsequent neural network–based approaches, particularly convolutional and recurrent architectures, demonstrated improved capacity for learning contextual representations from large-scale text corpora. However, their dependence on extensive annotated datasets and limited interpretability constrained their widespread clinical adoption[7].

The advent of Large Language Models (LLMs) has catalyzed a paradigm shift, demonstrating a remarkable capacity for clinical information extraction that can rival

human expert performance[8]. Recent studies have validated the high performance of LLMs across various institutions and cancer types. For instance, in breast cancer pathology, both proprietary (OpenAI's GPT-4o) and open-weight (Meta's Llama 3.1 405B) models achieved accuracies (96.1% and 94.7%, respectively) statistically comparable to a human annotator (95.4%)[9]. Similarly, researchers using Google's Gemma-2 model on the nationwide Veterans Health Administration database identified colorectal dysplasia and cancer with F1-scores exceeding 95%[10]. At Memorial Sloan Kettering Cancer Center, GPT-4 and GPT-3.5 extracted key elements from hepatocellular carcinoma reports with accuracies up to 94.8%[11]. However, while these studies affirm LLM capabilities, they also highlight a crucial limitation: the majority remain focused on a single cancer type or a narrow set of features. Even advanced multi-organ systems are often limited to classification tasks, such as assigning a "tumor group" across 19 organs, rather than performing the comprehensive, multi-field data extraction required for registry automation.11 To date, a truly comprehensive, multi-cancer, multi-feature extractor suitable for end-to-end automation has not been demonstrated[4].

This gap persists largely because transition of LLMs from research benchmarks to routine clinical practice is impeded by critical implementation barriers. A primary challenge is the risk of factual inaccuracy, or "hallucination," where models generate plausible but incorrect information—a failure mode that poses a significant threat to patient safety and data integrity[12,13]. Equally significant are the immense computational costs and data privacy concerns associated with state-of-the-art models. The highest-performing LLMs are often proprietary APIs, which require transmitting sensitive patient data to third-party vendors, raising regulatory issues[13]. Alternatively, deploying powerful open-source models locally has historically been bothered by expensive,

multi-GPU hardware clusters, a configuration that is infeasible for most healthcare institutions. These creates a critical gap between the theoretical capabilities of LLMs and their practical deployment in resource-constrained clinical environments[14].

This research is situated at the confluence of two groundbreaking developments that directly address these implementation challenges: the release of powerful, resource-efficient open-weight models and the maturation of programmatic LLM development frameworks. The first development involves an architectural shift toward computational efficiency, exemplified by sparse Mixture-of-Experts (MoE) models that increase capacity without a proportional rise in computational cost[15]. This efficiency, embodied in open-weight models like gpt-oss-20b and enhanced by advanced inference systems like MoBiLE[16,17], enables the local deployment of state-of-the-art models on consumer hardware—critical for ensuring data privacy and accessibility. Concurrently, the second development moves beyond brittle "prompt engineering". Programmatic frameworks like DSPy replace this manual process with a structured, compiler-driven approach that algorithmically optimizes prompts for specific models and tasks[18]. This paradigm of "programming, not prompting" is further enhanced by features like DSPy Assertions, which introduce self-refining capabilities to enforce reliability and correctness. The synergy of a locally deployable, high-performance model like gpt-oss and a programmatic optimization framework like DSPy provides, for the first time, the necessary toolkit to build the comprehensive, multi-cancer, multi-feature extractor that this study proposes.

This study directly addresses this implementation gap by presenting a novel, fully autonomous information extraction pipeline that achieves state-of-the-art performance across ten solid-organ cancers and can operate entirely on a single, commercially available GPU. Our framework makes three key contributions. First, we demonstrate

that a locally deployed, open-source LLM can achieve registry-grade accuracy across a comprehensive set of features for multiple cancer types, ensuring data privacy and security. Second, we introduce a robust, multi-stage workflow built using the DSPy framework, a programmatic approach that replaces brittle, manual prompt engineering with a more principled, self-optimizing, and declarative AI development process. Finally, by validating our system on a large cohort of uncurated, real-world pathology reports, we provide a deployable and economically viable blueprint that is truly comprehensive, breaking the resource barrier to enable the widespread automation of cancer surveillance.

2 Methods

2.1 Data Source and Annotation

This study was conducted under research ethics committee approval (REC #CMUH114-REC2-037) from China Medical University Hospital. The dataset comprised a total of 893 de-identified surgical pathology reports from the years 2023–2024. The final validation cohort consisted of 694 reports describing resected, invasive primary malignancies across ten major organ systems, with the detailed composition summarized in Table 1. An additional 199 reports describing biopsies, non-invasive tumors, or benign conditions served as negative controls for the initial classification stages of the pipeline.

To establish an expert-verified gold standard for evaluation, a semi-automated annotation process was employed. First, an initial structured JSON summary was generated for each report using a single inference pass of the gpt-oss:20b model. Subsequently, two board-certified pathologists independently reviewed and

corrected these preliminary outputs in a side-by-side interface displaying the original report text and the structured data. All discrepancies between the two reviewers were resolved through a final blinded consensus, producing the definitive reference dataset used for all performance evaluations.

Organ site	Number of reports (n)
Esophagus	75
Stomach	80
Colorectum	71
Breast	76
Uterine cervix	30
Pancreas	29
Prostate	95
Lung	64
Liver	80
Thyroid	75
Others (dual or out-of-scope cancers)	30
Total	694

Table 1 Composition of the cancer excision cohort (n = 694)

2.2 Schema definition

All data extracted by the system conform to a strictly typed, hierarchical JSON schema designed for clinical precision and computational reliability. The schema is composed of two layers: (1) a universal core layer that defines attributes common to all surgical pathology reports, such as report eligibility and primary tumor site,

and (2) ten organ-specific extension layers that capture the detailed variables unique to each malignancy.

Each organ-specific schema was developed to be in strict alignment with the College of American Pathologists (CAP) Cancer Protocols[19], ensuring that all extracted data fields, terminologies, and logical dependencies meet established standards for cancer reporting and registry submission. To enable deterministic, exact-match evaluation and prevent ambiguity, all schema fields are strictly typed using Pydantic-style models. Categorical fields are defined by enumerated string literals (Literal[...]), binary fields as booleans (bool), and numeric values as integers (int), with missing data explicitly represented as None. This approach eliminates inconsistencies arising from synonyms, rounding, or fuzzy text matching.

The complete, versioned CAP protocol references, along with full field definitions, enumerated values, validation constraints, and specifications for variable-length lists (e.g., for surgical margins and lymph nodes) for all ten organ-specific schemas, are provided in Supplementary Tables S1-S11.

2.3 Extraction pipeline

The deterministic extraction workflow was implemented in DSPy 2.1 using dspy.Predict() module, with predefined dspy.Signature that contains essential instructions and definitions, and executed on-premises using the gpt-oss:20b model served via Ollama.

No model training, fine-tuning or few-shot prompting was applied; each report was processed in a single forward pass.

2.4.1 Eligibility classification

The EligibilitySignature predicts whether a report represents a cancer surgery (cancer_excision_report: bool), the organ category (cancer_category), and a free-text descriptor when “others”.

2.4.2 Organ detection

If cancer_excision_report = True, a secondary signature (CancerTypeSignature) verifies or refines the organ label prior to extraction.

2.4.3 Organ-specific extraction

Ten sets of organ-specific DSPy modules (for example, for breastcancer, for modules, BreastCancerNonnested, BreastCancerMargins, BreastCancerLN, BreastCancerBiomarkers, was used) receive both original pathology report and the json summary, previously generated by gpt-oss:20b, as input and output a validated JSON matching their schema.

Modules will attempt to extract tumor characteristics, AJCC/TNM staging, surgical margins and lymph-node details for each cancer type. For breast and colorectal cancer, biomarker statuses are extracted as well. An overview of the deterministic extraction pipeline is shown in Figure 1 for a breast cancer report example, highlighting each computational stage and its data outputs.

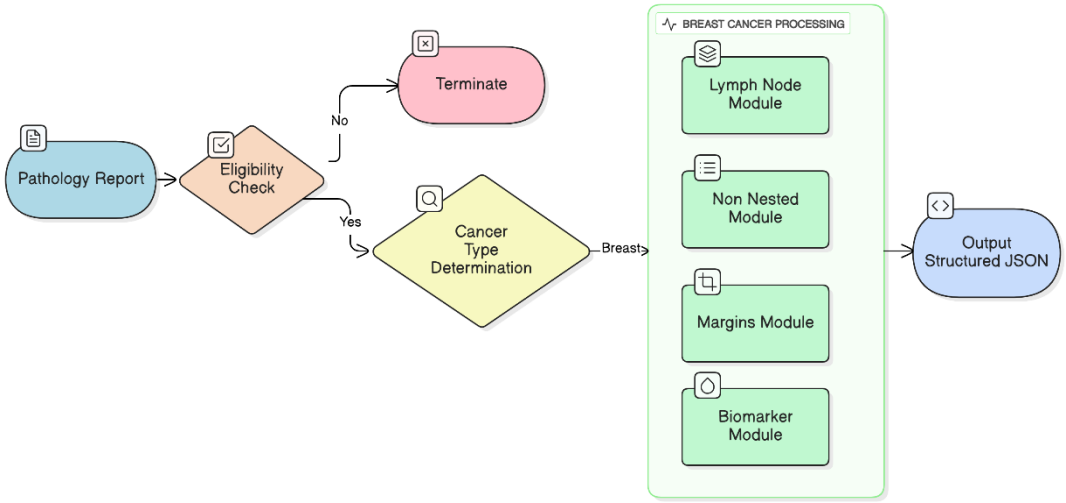

Figure 1 Extraction pipeline for structured cancer data from surgical pathology reports (breast as an example).

2.5 Evaluation metrics

Model performance was evaluated by comparing each generated JSON output against the expert-verified reference standard. Given that all schema variables are strictly typed (i.e., Literal, bool, int, or None), evaluation was conducted using deterministic, exact-string matching rather than similarity-based metrics. For list-type variables, such as surgical margins and lymph node stations, corresponding elements were aligned by field name before comparison. All statistical analyses were performed using the scikit-learn library (version 1.5.0) in Python 3.11.

The primary metrics included accuracy, precision, recall, and the F1 score. The F1 score is calculated as the harmonic mean of precision and recall, providing a single, balanced measure of a test's accuracy. For multiclass classification tasks, such as the organ-type identification performed by the CancerTypeSignature classifier, the macro-average F1 score was computed as the unweighted mean of the F1 scores for each class, thereby giving equal weight to each category regardless of

its frequency. Overall accuracy was defined as the proportion of all extracted fields that exactly matched the reference standard across the validation cohort. The specific definitions for composite metrics applied to the extraction modules, such as margin status correctness and total lymph node accuracy, are detailed in Table 2.

Metric	Definition
Margin status correctness	Exact match of both margin_category and margin_involved for every reported margin element.
Margin distance correctness	Exact match of margin_category, margin_involved, and distance (integer, mm) for each margin element.
Total lymph-node accuracy	Exact match of examined and involved integer counts within each organ's regional_lymph_node list.
Lymph-node station accuracy	Exact match of station_name, examined, and involved values for each lymph-node group ("station").
Accuracy of other fields	Proportion of all remaining scalar fields (e.g., histological_type, tumor_size, grade, biomarker results, and stage-group entries) that exactly match the expert reference.

Table 2 Evaluation metrics and definitions

Each metric represents a binary exact-match comparison between model output and expert-annotated reference. No partial credit or token-level similarity was used. Overall accuracy for each category equals the ratio of correctly matched fields to the total number of evaluated fields.

2.6 Computational environment and reproducibility

All experiments were conducted on a NVIDIA RTX A6000 Ada Generation GPU (48 GB VRAM). The pipeline was implemented in Python 3.11 using DSPy 2.1 for declarative program synthesis and the open-source gpt-oss:20b model (Apache 2.0 license) for rough JSON generation and LLM inference.

2.7 Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, writing of the report, or in the decision to submit the paper for publication.

2.8 Ethical oversight

This study was approved by the research ethics committee of China Medical University & Hospital, under REC number of CMUH114-REC2-037. All reports were fully de-identified and processed on encrypted institutional servers. No re-identification was attempted.

3 Results

3.1 Eligibility and organ classification

The eligibility classifier reliably distinguished cancer surgery reports from non-cancer-surgery documents across three independent test runs (N = 694). Accuracy consistently exceeded 96 % (0.966–0.969), with average accuracy of 96.6% and a macro average F1 score of 97.9%, confirming near-perfect sensitivity and minimal false-negative risk. Only a single breast-excision report was

misclassified in Run 1, while remaining false positives (28–30 per run) were confined to the “others” category. These results indicate that the model can safely operate as the first triage stage in an automated pipeline, ensuring that no eligible cancer surgery report is missed.

The subsequent organ-labeling sub-model demonstrated similarly strong performance, achieving precision F1-scores ≥ 0.955 across all ten target organs, with an average accuracy of 0.978 and a macro average F1 score of 0.955. Eight organ classes—breast, cervix, colorectal, liver, pancreas, prostate, stomach, and thyroid—achieved perfect recall (1.000), while both lung and esophagus reached F1 ≥ 0.955. Misclassifications were mostly attributable to cases with double primary malignancies—for example, reports describing independent tumors in both lung lobes, each staged as a separate primary cancer. Such cases, which were intentionally excluded from automated staging logic and assigned to the heterogeneous “others” category, were occasionally interpreted by the model as belonging to the “lung” class. Collectively, these results confirm that foundational classification layers of the DSPy pipeline operate with registry-grade reliability, providing clean input for downstream extraction of tumor characteristics, margins, and lymph-node (LN) data. Detailed confusion matrices for each of the three test runs are provided in the Supplementary Tables S12–S14.

Building on the model’s demonstrated stability in both eligibility determination and organ classification—even under stochastic sampling (temperature = 0.7)—subsequent evaluations focused on the extraction components. Accordingly, accuracy assessments for per-organ extraction, lymph-node (LN) extraction, margin extraction, and biomarker extraction were conducted using results from Run 3 exclusively.

3.2 Per-organ extraction accuracy for various registry item

After classification, 893 cancer surgery reports were parsed through organ-specific extraction modules, each aligned to a JSON schema of variable fields according to different organ system (procedure, histology, grade, AJCC T/N/M categories, and vice versa, please see Supplementary Tables S14–S23 for detail). Exact-match accuracies across all ten organ systems (totally 193 fields) ranged between 88.9 % and 96.2 %, with an overall mean of 94.3%±8.32%. Table 3 summarizes the cohort size, field count, and mean performance for each organ. Surgical margins, lymph nodes and biomarkers are extracted by separate modules and discussed in later section.

Organ	n Reports	# Fields Evaluated	Mean Accuracy ± SD (%)
Breast	75	26	**93.5 ± 9.7**
Uterine cervix	29	16	**90.9 ± 9.7**
Colon and rectum	70	20	**96.6 ± 5.3**
Esophagus	73	16	**96.2 ± 4.8**
Liver – HCC	79	17	**94.1 ± 9.0**
Lung	63	20	**92.5 ± 7.1**
Pancreas	28	15	**91.7 ± 12.9**
Prostate	94	25	**97.0 ± 3.4**
Stomach	79	18	**93.2 ± 11.6**
Thyroid	74	20	**92.3 ± 9.5**

Table 3. Per-organ extraction accuracy for various registry item

High-performing fields (≥ 95 %) were consistent across organs expect thyroid (where frequent referrance to other reports caused some model difficulty) and included procedure histology, grade, and presence of vascular invasion, indicating that the pipeline captures core oncologic variables with near-perfect fidelity even in organs with extensive schema complexity (≥ 24 fields).

In contrast, performance on TNM staging fields showed greater variability across organs, reflecting the inherent complexity and inter-organ heterogeneity of staging schemas. Occasional ambiguities in clinical phrasing—such as cases labeled "Nx" but inferred by the model as "N0"—also contributed to minor discrepancies. Nevertheless, overall accuracy for T, N, and M fields remained high, typically exceeding 90% across test runs.

Lower-accuracy fields (60–90 %) were limited to poorly documented or quantitative attributes—such as surgical technique in stomach and cervix or tumor percentage and prostate weight in prostate reports—where textual variability is high.

In summary, the organ-specific modules sustain ≥ 90 % mean exact-match accuracy across varying schema sizes, reaffirming the scalability of the one-shot LLM pipeline for comprehensive extraction of structured oncologic data. Full per-field extraction results for each organ, including exact-match accuracy for all schema items, are provided in the Supplementary Tables S15–S24.

3.3 Lymph Node and Surgical Margin Status

The pipeline demonstrated strong performance in extracting prognostically critical data on lymph node (LN) and surgical margin status, as summarized in Figures 2 and 3. For LN evaluation, the system achieved a mean total-node

concordance of over 80% across all ten organ systems. Performance was highest in organs with well-defined nodal basins such as the breast and liver, while it was more variable in anatomically complex sites like the lung and pancreas, where occasional omissions were the primary source of error rather than misclassification. Detailed per-organ LN metrics, including both total and station-level analyses, are provided in Supplementary Table S25.

For surgical margins, the model achieved a mean accuracy of 91.2% for determining margin positivity. Concordance was excellent (≥95%) for several major cancers, including breast, colorectal, and lung. In cases with lower accuracy, such as pancreas resections, discrepancies were primarily attributable to challenges in mapping variably named margin categories rather than incorrect identification of tumor involvement. As expected, the accuracy for extracting the exact margin distance was lower, reflecting the inherent ambiguity of inequality notations and unit variations in the source reports.

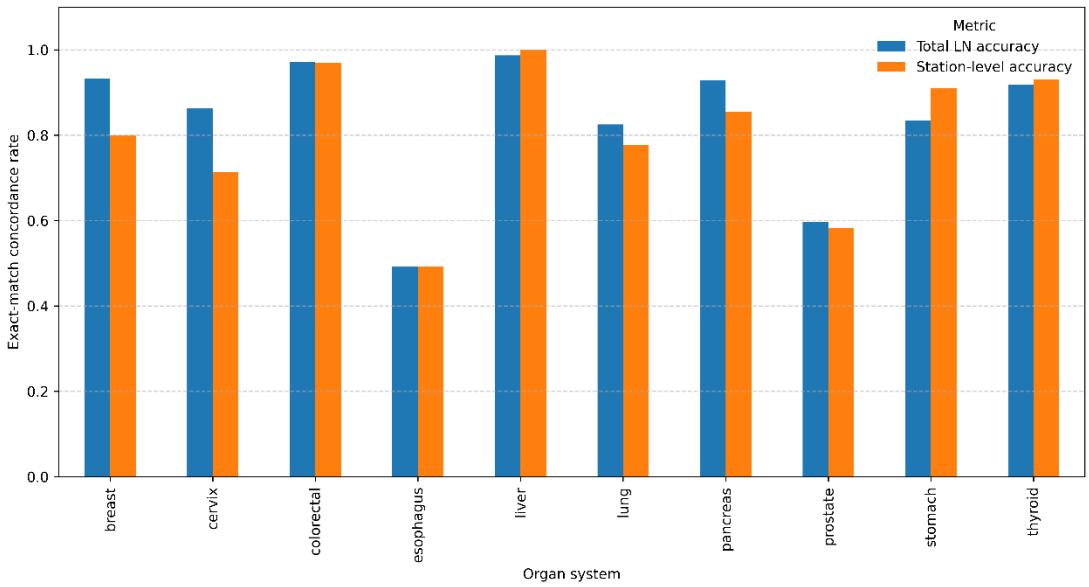

Figure 2. Lymph node status extraction accuracy by organ

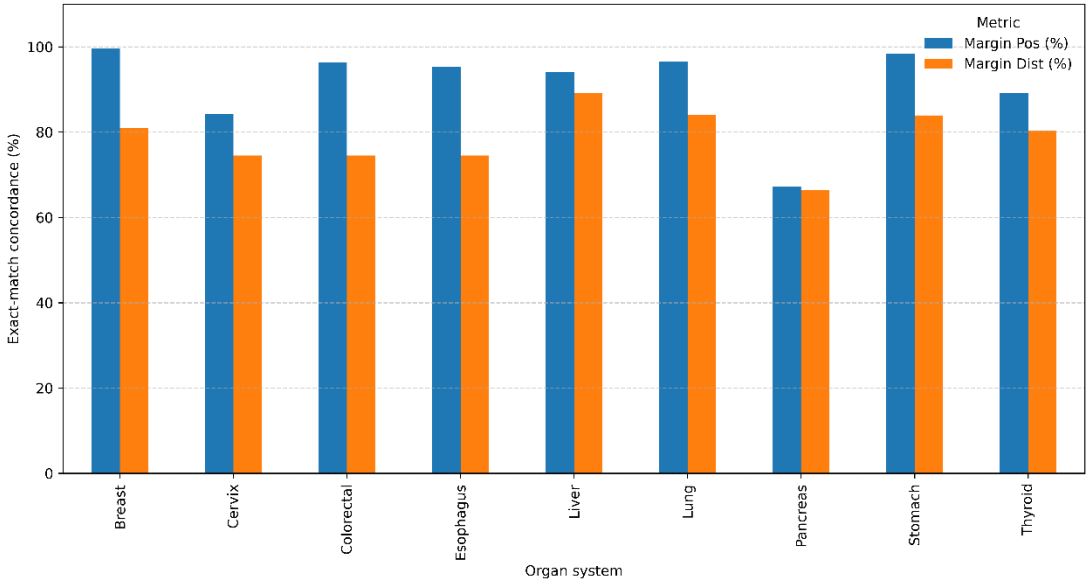

Figure 3. Surgical margin status extraction accuracy by organ.

3.4 Breast biomarker extraction performance

Breast cancer biomarker fields were extracted with near-perfect accuracy across 75 reports. ER positivity and percentage were correct in 75/75 and 73/75 cases, respectively. PR positivity and percentage were correct in 74/75 and 73/75 cases. HER2 score was correctly extracted in all 75 cases, and Ki-67 percentage in 74/75 cases. These results confirm the robustness of field-level extraction and validation under exact string matching. The results are summarized in Table 4.

Biomarker	Field type	Correct (n)	Accuracy (%)
Estrogen receptor (ER)	Positivity	75 / 75	100.0
Estrogen receptor (ER)	Percentage	73 / 75	97.3
Progesterone receptor (PR)	Positivity	74 / 75	98.7
Progesterone receptor (PR)	Percentage	73 / 75	97.3
HER2	Score	75 / 75	100.0
Ki-67	Percentage	74 / 75	98.7

Table 4. Accuracy of breast cancer biomarker extraction (N = 75)

3.5 Overall Performance Summary

The DSPy pipeline achieved registry-grade reliability: >96 % accuracy for classification, >94 % for per-organ extraction, >80 % for lymph-node concordance, and >90 % for margin positivity. Breast biomarker extraction exceeded 97 %. Overall, the system demonstrated robust, scalable, and clinically interpretable performance across diverse organ systems.

4 Discussion

4.1 Principal Findings

This study demonstrates the successful development and validation of a resource-efficient, privacy-preserving, and autonomous AI pipeline that achieves cancer-registry-grade accuracy for multi-cancer data extraction from unstructured pathology reports. By operating on a single GPU and leveraging a locally deployed open-weight model, our system provides a practical solution to the long-standing informatics bottleneck that has constrained the scalability and timeliness of cancer

surveillance for decades. The pipeline's robust performance—achieving a mean exact-match accuracy of 93.9%±8.6% across ten distinct malignancies and 196 registry fields—underscores its potential for real-world clinical deployment, offering a comprehensive, end-to-end workflow that encompasses document triage, cancer-type classification, and granular data abstraction.

4.2 Research Context and Comparison with the State of the Art

While large language models (LLMs) now demonstrate human-level accuracy for information extraction from clinical texts[8], their clinical deployment is hindered by a critical trilemma: proprietary models raise significant data privacy and cost concerns, while open-source alternatives have historically required computationally prohibitive hardware[20]. Our study directly addresses this implementation gap by presenting a fully autonomous pipeline that is simultaneously accurate, privacy-preserving, and resource-efficient. This work is situated at the confluence of two recent technological shifts. First, the emergence of powerful, open-weight models like gpt-oss:20b that are optimized for local deployment on a single GPU, which resolves the hardware barrier and ensures data sovereignty[17]. Second, the maturation of programmatic LLM development frameworks like DSPy, which enables a paradigm shift from brittle prompt engineering to the construction of robust, reliable, and scalable AI systems[18]. By integrating these advances, our work moves beyond single-cancer benchmarks to deliver a comprehensive, multi-organ abstraction system. It provides an economically viable and generalizable blueprint for a "digital registrar," demonstrating a practical path to overcoming the technological and organizational barriers that have stalled widespread AI adoption in healthcare informatics.

4.3 Limitations

The principal limitation of this study is its reliance on data from a single academic medical center in Taiwan for both development and validation. This context is crucial for interpreting the findings; while the pipeline has proven effective on reports generated within a specific institutional and linguistic setting, its generalizability across different healthcare systems, with their inherent variability in formatting conventions and reporting styles, remains unverified. Therefore, future work must include independent external benchmarking on multi-institutional and international datasets to confirm the system's robustness and adaptability before widespread clinical adoption can be recommended.

4.4 Further research direction and intrinsic challenges

A key future direction is the extension of this framework into a multimodal system. While our system excels at abstracting data from textual reports, the "digital registrar" of the future must integrate information from various data streams to build a comprehensive patient model. The integration of genomic data and digital pathology images, as highlighted in recent reviews of multimodal AI in oncology, represents the next frontier in automated cancer surveillance[21]. The development of pathology-specific multimodal LLMs, such as PathChat+, which can reason directly over high-resolution pathology images, represents a pivotal technological enabler for this vision[22,23]. A future system built on our framework could leverage such models to correlate histologic patterns with molecular markers and clinical outcomes at an unprecedented scale, moving beyond mere data extraction to true integrated diagnostics and prognostication[24].

4.5 Conclusion

In conclusion, by solving the critical implementation challenges of resource efficiency and data privacy, this work provides a deployable and economically viable blueprint that democratizes access to state-of-the-art AI for cancer surveillance. This research marks a pivotal step in transforming this essential public health function from a slow, retrospective, and manual process into a near-real-time, intelligent system. Such a system is poised to accelerate clinical research, enhance the precision of public health action, and ultimately improve patient outcomes on a global scale.

5 Ethical Statement

This study was approved by the research ethics committee of China Medical University & Hospital, under REC number of CMUH114-REC2-037. All reports were fully de-identified and processed on encrypted institutional servers. No re-identification was attempted.

6 Data Sharing Statement

The de-identified dataset used for the development and validation of the model, including the full text of the pathology reports and the expert-annotated ground truth data, will be made available. The complete data dictionary defining each schema field will also be provided. The source code for the AI extraction pipeline is also available for sharing. These materials will be made available to qualified academic researchers for non-commercial purposes, such as for the replication or extension of the present work, beginning 3 months and ending 5

years following article publication. Interested parties must submit a formal research proposal outlining the intended use of the data and/or code to the corresponding author for review. Access will be granted upon the execution of a data use agreement and a software license agreement to ensure the protection of patient privacy and intellectual property.

7 Declaration of interests

We declare no competing interests.

8 Funding

This work was supported by the National Science and Technology Council (NSTC) of Taiwan, under grant number NSTC 113-2221-E-039-017, and the NSTC College Student Research Grant, number 114-2813-C-039-132-E.

9 Acknowledgement and generative AI statement

We are deeply grateful to Dr. Ting-Sung Hsieh of UT Southwestern Medical Center, as well as Dr. Ching-Hang Chien and Dr. Yi-Chung Dzeng of the Research Center for Applied Sciences, Academia Sinica. Their insightful feedback and critical suggestions were invaluable to the intellectual development and final refinement of this manuscript.

This manuscript was prepared with partial drafting assistance from large language models. OpenAI GPT-5 was used for non-confidential sections to improve structure and clarity; gpt-oss:20b was used to draft and revise paragraphs describing proprietary schema design and extraction logic under a controlled,

offline environment; and Gemini Pro 2.5 with Deep Research was used to identify and summarize peer-reviewed literature cited in the Introduction. All analytical workflows, data interpretation, and final text approval were performed by the authors. No LLM generated, altered, or fabricated experimental data.

10 Reference

1. Parkin DM. The role of cancer registries in cancer control. Int J Clin Oncol. 2008 Apr;13(2):102–11.

2. Dahl S, Bøgsted M, Sagi T, Vesteghem C. Performance of Natural Language Processing for Information Extraction From Electronic Health Records Within Cancer: Systematic Review. JMIR Med Inform. 2025 Sept 12;13:e68707.

3. Midthune DN, Fay MP, Clegg LX, Feuer EJ. Modeling Reporting Delays and Reporting Corrections in Cancer Registry Data. Journal of the American Statistical Association. 2005 Mar 1;100(469):61–70.

4. Gondara L, Simkin J, Arbour G, Devji S, Ng R. Classifying Tumor Reportability Status From Unstructured Electronic Pathology Reports Using Language Models in a Population-Based Cancer Registry Setting. JCO Clin Cancer Inform. 2024 Nov;(8):e2400110.

5. Park P, Choi Y, Han N, Park YL, Hwang J, Chae H, et al. Leveraging natural language processing for efficient information extraction from breast cancer pathology reports: Single-institution study. PLOS ONE. 2025 Feb 18;20(2):e0318726.

6. Abedian S, Sholle ET, Adekkanattu PM, Cusick MM, Weiner SE, Shoag JE, et al. Automated Extraction of Tumor Staging and Diagnosis Information From Surgical Pathology Reports. JCO Clinical Cancer Informatics. 2021 Dec;(5):1054–61.

7. Alawad M, Gao S, Qiu JX, Yoon HJ, Blair Christian J, Penberthy L, et al. Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks. J Am Med Inform Assoc. 2020 Jan 1;27(1):89–98.

8. Balasubramanian JB, Adams D, Roxanis I, de Gonzalez AB, Coulson P, Almeida JS, et al. Leveraging large language models for structured information extraction from pathology reports. Journal of Pathology Informatics. 2025 Oct 10;100521.

9. Ntinopoulos V, Rodriguez Cetina Biefer H, Tudorache I, Papadopoulos N, Odavic D, Risteski P, et al. Large language models for data extraction from unstructured and semi-structured electronic health records: a multiple model performance evaluation. BMJ Health Care Inform. 2025 Jan 19;32(1):e101139.

10. Johnson B, Bath T, Huang X, Lamm M, Earles A, Eddington H, et al. Large language models for extracting histopathologic diagnoses of colorectal cancer and dysplasia from electronic health records. BMJ Open Gastroenterol. 2025 Sept 18;12(1):e001896.

11. Geevarghese R, Sigel C, Cadley J, Chatterjee S, Jain P, Hollingsworth A, et al. Extraction and classification of structured data from unstructured hepatobiliary pathology reports using large language models: a feasibility study compared with rules-based natural language processing. J Clin Pathol. 2025 Jan 17;78(2):135–8.

12. Cheng J. Applications of Large Language Models in Pathology. Bioengineering. 2024 Apr;11(4):342.

13. Busch F, Hoffmann L, Rueger C, van Dijk EH, Kader R, Ortiz-Prado E, et al. Current applications and challenges in large language models for patient care: a systematic review. Commun Med (Lond). 2025 Jan 21;5:26.

14. Lorencin I, Tankovic N, Etinger D. Optimizing Healthcare Efficiency with Local Large Language Models. In: Intelligent Human Systems Integration (IHSI 2025): Integrating People and Intelligent Systems [Internet]. AHFE Open Acces; 2025 [cited 2025 Oct 19]. Available from: https://openaccess.cms-conferences.org/publications/book/978-1-964867-36-6/article/978-1-964867-36-6_58

15. Mu S, Lin S. A Comprehensive Survey of Mixture-of-Experts: Algorithms, Theory, and Applications [Internet]. arXiv; 2025 [cited 2025 Oct 24]. Available from: http://arxiv.org/abs/2503.07137

16. Zhao Y, Qin Y, Wang Y, Yang X, Han H, Wei S, et al. MoBiLE: Efficient Mixture-of-Experts Inference on Consumer GPU with Mixture of Big Little

Experts [Internet]. arXiv; 2025 [cited 2025 Oct 24]. Available from: http://arxiv.org/abs/2510.12357

17. Kumar D, Yadav D, Patel Y. GPT-OSS-20B: A Comprehensive Deployment-Centric Analysis of OpenAI's Open-Weight Mixture of Experts Model [Internet]. arXiv; 2025 [cited 2025 Oct 19]. Available from: http://arxiv.org/abs/2508.16700

18. Khattab O, Singhvi A, Maheshwari P, Zhang Z, Santhanam K, Haq S, et al. Dspy: Compiling declarative language model calls into state-of-the-art pipelines. In: The Twelfth International Conference on Learning Representations [Internet]. 2024 [cited 2025 Oct 19]. Available from: https://openreview.net/forum?id=sY5N0zY5Od

19. Torous VF, Simpson RW, Balani JP, Baras AS, Berman MA, Birdsong GG, et al. College of American Pathologists Cancer Protocols: From Optimizing Cancer Patient Care to Facilitating Interoperable Reporting and Downstream Data Use. JCO Clinical Cancer Informatics. 2021 Dec;(5):47–55.

20. Dennstädt F, Hastings J, Putora PM, Schmerder M, Cihoric N. Implementing large language models in healthcare while balancing control, collaboration, costs and security. NPJ Digit Med. 2025 Mar 6;8:143.

21. Zhang B, Wan Z, Luo Y, Zhao X, Samayoa J, Zhao W, et al. Multimodal integration strategies for clinical application in oncology. Front Pharmacol. 2025 Aug 20;16:1609079.

22. Chen C, Weishaupt LL, Williamson DFK, Chen RJ, Ding T, Chen B, et al. Evidence-based diagnostic reasoning with multi-agent copilot for human pathology [Internet]. arXiv; 2025 [cited 2025 Oct 24]. Available from: http://arxiv.org/abs/2506.20964

23. Waqas A, Tripathi A, Ramachandran RP, Stewart PA, Rasool G. Multimodal data integration for oncology in the era of deep neural networks: a review. Front Artif Intell [Internet]. 2024 July 25 [cited 2025 Oct 24];7. Available from: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2024.1408843/full

24. Witowski J, Zeng K, Cappadona J, Elayoubi J, Chiru ED, Chan N, et al. Multi-modal AI for comprehensive breast cancer prognostication. ArXiv [Internet].

2024;abs/2410.21256. Available from:
https://api.semanticscholar.org/CorpusID:273654929