Comprehensive Structured Abstraction of Pathology Reports Is Now Feasible Using Local Large Language Models

Nan-Haw Chow, 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-Ju Shiu, Ph.D.[6], Yen-Wei Chu, Ph.D.[7], Pau-Choo Chung, Ph.D.[8], 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

[8]Department of Electrical Engineering, National Cheng Kung University, Tainan, 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

Pathology reports contain the most detailed descriptions of cancer diagnoses, yet their unstructured format has long limited large-scale reuse for cancer registries and population surveillance. Prior applications of large language models (LLMs) have therefore focused on narrow extraction tasks, reflecting a persistent implementation trilemma: comprehensive abstraction, strict data privacy, and computational feasibility could not be achieved simultaneously in real-world clinical settings. Given current LLM capabilities, this trilemma can now be resolved. We show that recent open-weight LLMs enable reliable, full-length, schema-bound abstraction of pathology reports on standard on-premise hardware. We present a model-agnostic framework implemented using DSPy, a declarative framework for structured LLM pipelines, in which deterministic, programmatic prompting co-designed with pathologists enables end-to-end structured abstraction. Across 893 real-world pathology reports spanning ten major cancer types, the system achieved a mean exact-match accuracy of 94.3% across 193 CAP-aligned registry fields, including complex variable-length structures such as surgical margins, lymph nodes, and breast biomarkers. All processing was performed locally on a single workstation-class GPU, ensuring data privacy without sacrificing completeness or feasibility. Independent external validation using TCGA pathology reports confirmed robust generalizability.

1. Introduction

Cancer registries serve as critical infrastructures for public health surveillance and clinical decision-making. Their data underpin epidemiological research, policy design, and treatment evaluation[1–3]. However, the system has long relied on manual abstraction of information from unstructured pathology reports—a process that is time-consuming, error-prone, and increasingly unsustainable as case volumes surge and manpower declines. Studies have shown that over 80% of clinically relevant information resides in free-text pathology reports. Manual abstraction typically introduces months of delay and substantial error rates, degrading both surveillance and research quality. As case volumes and reporting requirements expand, the system's workload has grown exponentially, rendering manual maintenance unsustainable[4–6].

We categorize this structural crisis as a threefold systemic failure:

(1) Failure of Completeness — insufficient manpower and growing caseloads lead to missing diagnostic and treatment information, systematically under-representing certain patient groups[4,5,7]

(2) Failure of Ethics and Efficiency — manual abstraction repeatedly exposes complete medical records, heightening privacy risks and increasing operational costs[8–10]

(3) Failure of Granularity — temporal dynamics and biological detail are compressed into single field values during registration, leaving only a coarse “skeleton” of the patient record for analysis[11–13].

Together, these systemic failures highlight the urgent need for scalable, automated solutions that can match the complexity and volume of modern cancer registries. Recent studies have made meaningful progress in applying large language models (LLMs) to oncology workflows, with several reports demonstrating near-expert performance in well-defined information extraction tasks. However, these advances have largely been achieved in constrained problem settings, often tackling one narrow subtask at a time[14–16]. Existing on-premises deployments, while safeguarding patient privacy, typically address only a narrow subset of cancer registry requirements[17–20].

For example, one locally deployed system parsed surgical pathology reports but extracted only six high-level diagnostic variables across multiple tumor types[21–25]. Other multi-cancer efforts expanded disease coverage but still focused on a limited synoptic report schema, extracting only about 22 standardized data elements [26]. Even the most comprehensive single-cancer study to date—targeting breast pathology—captured 51 structured fields and required an enormous ~405-billion-parameter model to achieve expert-level accuracy, a configuration far beyond the computational capacity of routine hospital environments [14,24]. Cloud-based LLMs (e.g. GPT-4) have likewise shown strong performance on oncology benchmarks[27], yet their practical deployment remains limited. Strict data-privacy regulations (e.g. HIPAA and GDPR) generally preclude routine transmission of protected health information to external AI services[17], and current generative models can exhibit instability when handling long clinical texts, making deterministic and reproducible extraction outputs difficult to obtain[28]. Together, these findings highlight both the progress and the persistent gaps in the field: despite promising prototypes, no existing system yet provides a comprehensive, end-to-end abstraction solution suitable for real-world cancer registry operations[14].

Still, no framework simultaneously satisfies all three essentials: comprehensive scope, ethical privacy, and computational feasibility. These define the central bottleneck of registry automation and mirror the threefold systemic failures noted above. We posit that earlier systems failed to meet the three core requirements because they lacked a prompting engine co-designed with pathologists, while local LLMs at the time were limited in scale and reasoning, preventing stable clinical extraction. With advances from 2023 to 2025, local LLMs have become efficient enough to support modular, scientifically grounded prompting engines.

To address the limitations of brittle manual prompting, we utilized DSPy, a framework that abstracts language model interactions into programmable, modular primitives[29]. Building upon this programmatic approach, this study enables clinician-engineer

co-design of extraction logic and applies it to ten major cancers under CAP-standard field structures. This yields a reproducible and auditable registry architecture that overcomes the challenges of completeness, ethical integrity, granularity, feasibility, and generality:

(1) Restoring Completeness: The system extracts over 193 registry fields across ten major cancer types with a mean exact-match accuracy of 94.3%. By capturing variable-length items such as margin and lymph node status according to CAP templates, the system systematically corrects the omissions and selection biases often introduced by manual abstraction.

(2) Restoring Ethical Integrity: By enabling strictly on-premise computation, the system keeps all processing within the institution, eliminating the risks of PHI transmission. Our single-GPU inference approach balances cost efficiency with rigorous regulatory compliance.

(3) Restoring Granularity: Leveraging DSPy's multi-step reasoning capabilities, the pipeline reconstructs complex semantic dependencies and temporal details often lost in broad classification tasks. This programmatic logic ensures that fine-grained pathological features are extracted with high consistency, moving beyond simple keyword matching.

(4) Ensuring Feasibility: We demonstrated architectural independence by benchmarking the pipeline across three diverse, resource-efficient open-weight models (GPT-OSS[30], Qwen3[31], and Gemma3[32]). By focusing on this compact 20B–30B parameter class, our comparative analysis confirms that the system can be successfully deployed on standard clinical workstations (48GB VRAM), ensuring the technology is accessible without requiring high-performance datacenter infrastructure.

(5) Extending Generality: We validated the system's transportability through a rigorous Independent External Validation (IEV) using the multi-institutional TCGA (The Cancer Genome Atlas) dataset[33]. By testing on breast, colorectal, and esophageal cohorts from diverse institutions, the system maintained high extraction fidelity (>92% accuracy) despite unseen data structures and linguistic styles.

In summary, the proposed framework restores completeness, ethical integrity, and informational granularity, while establishing cross-model generality validated on standard hardware. This achievement enables the technical realization of truly comprehensive structured pathology-report abstraction, laying the groundwork for future real-time clinical data infrastructures.

2. Results

2.1 Assessment of Model Agnosticism and Computational Feasibility

To determine the optimal engine for clinical deployment, we benchmarked the extraction framework across three distinct open-weight architectures (gpt-oss:20b, qwen3-30b-A3B, and gemma3:27b). The objective was to validate the model-agnostic nature of the pipeline while identifying the configuration that best balances registry-grade accuracy with the hardware constraints of a standard medical workstation (single GPU, 48GB VRAM).

As shown in **Table 1**, the framework achieved high exact-match accuracy across all tested architectures, confirming its agnosticism. However, a decisive divergence emerged in operational efficiency. Although all three models operated within the workstation-class envelope (e.g., a 48 GB VRAM workstation GPU), the fastest model by a clear margin was gpt-oss:20b, which maintained the highest accuracy (94.3%) while processing complex reports in roughly 40–70 seconds, about 2–3× faster than its alternatives.

The performance gap is not primarily due to simple VRAM capacity exhaustion; rather, it reflects the interplay of model architecture, total parameter state, expert routing overhead (in the case of mixture-of-expert (MoE) models), and memory/ bandwidth traffic. For instance, the qwen3□30b□A3B model, despite activating only ~3.3 billion parameters per token, has a total parameter size of ~30.5 billion and uses 128 experts with 8 activated per token[31]. These aspects increase system overhead and latency, positioning it behind gpt-oss:20b in our single-GPU workstation tests.

Consequently, based on this efficiency and accuracy analysis, gpt-oss:20b was identified as the optimal candidate for the "Digital Registrar" framework—being the model that delivers registry-grade accuracy (94.3%) with high throughput on standard workstation hardware. For this reason, all subsequent results in this study—including eligibility triage, organ classification, and granular field extraction (Sections 2.2–2.4)—are derived exclusively from the gpt-oss:20b pipeline.

2.2 Automated Report Triage and Organ Classification

To ensure the pipeline selectively processes valid cancer registries, the system was first evaluated on its ability to triage unstructured pathology reports—distinguishing eligible cancer excision cases from a background of biopsies and non-malignant specimens. As detailed in **Table 2**, the eligibility classifier demonstrated robust reliability across the full validation cohort (N = 893). Crucially for a screening tool, the model exhibited near-perfect sensitivity; out of 694 eligible cancer excision reports, only a single breast excision case was missed (False Negative), confirming its

safety as an automated filter. The specificity was similarly high, with most discrepancies (29 False Positives) restricted to clinically ambiguous cases—such as large metastatic deposits or extensive biopsies—that mimic the linguistic complexity of primary resections. These results confirm that the model operates with the high sensitivity required to ensure that virtually no eligible cancer surgery report is overlooked during the initial screening process.).

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 rare and distinctively patterned: qualitative review revealed these were often cases with double primary malignancies—such as reports describing independent tumors in both lung lobes—which were intentionally excluded from the automated staging logic (ground truth "Others") but were occasionally interpreted by the model as belonging to the specific organ class (e.g., "Lung"). Collectively, these results confirm that foundational classification layers of the DSPy pipeline operate with registry-grade reliability, providing clean input for downstream extraction. Detailed confusion matrices for all three test runs are provided in **Supplementary Tables S12–S14**.

2.3 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.

High-performing fields (≥ 95 %) were consistent across organs expect thyroid (where frequent reference 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 “anatomic staging” and

"pathologic staging" in breast—also contributed to minor discrepancies. A representative example is provided in **Supplementary Appendix S25.3** (Sample S25.3.1–S25.3.2). 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. A representative example and extra discussion is provided in **Supplementary Appendix S25.4** (Sample S25.4.1–S25.4.2).

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**.

2.4 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 1 and 2**. 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 esophagus, 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. Examples of LN extraction performance are provided in **Supplementary Appendix S25.5**. (Samples S25.5.1.1-S25.5.2.2).

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. Representative margin-status extraction results the show complexity of pancreatic cancer are shown in **Supplementary Appendix S25.6** (Samples S25.6.1–S25.6.2). Because prostate cancer reports do not define multiple margin stations and instead document margin involvement using only overall

positivity, extent of involvement, and the specific anatomic sites involved, prostate margins were excluded from this station-based margin accuracy evaluation.

2.5 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.

2.6 External validation performance (TCGA Cohort)

In the external TCGA cohort, the model maintains performance comparable to the internal validation set, confirming strong cross-domain robustness. Across 150 reports, triage and organ classification achieved 97.3% accuracy, with near-perfect discrimination for major cancers (F1 = 1.000 for colorectal; 0.990 for esophagus). The few discrepancies were administrative rather than interpretative: three reports labeled as “Others” in the ground truth were predicted as breast cancer but were subsequently verified as synchronous bilateral breast cancers—cases intentionally excluded from organ-level staging due to the system’s single-primary design. Field-level extraction across 53 pathological variables reached a mean exact-match accuracy of 92.4% ± 11.0%, closely mirroring internal-set performance (93.9% ± 8.6%) and demonstrating that the extraction logic generalizes effectively beyond the institution where it was developed.

3. Discussion

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 193 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.

While large language models (LLMs) now demonstrate human-level accuracy for information extraction from clinical texts[24], 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[34]. 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[30]. 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[29]. 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.

The framework's demonstrated capacity for model-agnosticism—validated by the seamless interchange of the three models (gpt□oss:20b, qwen3□30b□A3B and gemma3:27b)—provides crucial flexibility for healthcare deployment. All three operate within the workstation-class envelope (e.g., a 48 GB GPU such as the NVIDIA RTX A6000) rather than exclusively requiring datacenter-scale infrastructure. Nonetheless, our comparative analysis reveals that even within this workstation context, deployment efficiency diverges sharply. While all models are technically viable, gpt-oss:20b achieved the fastest inference (≈ 40-70 seconds) on our single-GPU tests. In contrast, qwen3-30b-A3B recorded substantially higher latency (≈ 140->200 seconds) despite its low active-parameter count (~3 billion) because its 30 billion-parameter total state plus MoE routing and memory-traffic overhead pushed the system toward bandwidth/interconnect limits. gemma3:27b (≈ 27 billion parameters) fell in between (≈ 50-130 seconds) — still deployable on workstation-class hardware, but slower than gpt-oss:20b and with lower accuracy. The key takeaway is that hardware capacity is not simply 'sufficient/insufficient' — architecture (MoE vs dense), active-vs-total parameters, memory traffic and routing overhead all combine to determine practical performance.

In the use-case of extraction of structured fields from complex reports (as in our healthcare scenario), the superior accuracy of gpt-oss:20b matters materially: it maintained higher exact-match accuracy in our benchmarks. Meanwhile, although gemma3:27b's latency was comparable, its accuracy lagged behind. Notably, when extraction tasks demand high fidelity, the MoE-based models (gpt-oss:20b and qwen3-30b-A3B) may offer advantages: their slim active-parameter footprint can mean lower memory overhead and potentially more targeted inference. Thus, while gpt-oss:20b currently presents the best balance of speed and accuracy, qwen3-30b-A3B's MoE design

still merits consideration in contexts where higher model capacity might pay off and latency constraints are less stringent.

The independent external validation results confirm the model's stability when applied to a structurally and institutionally diverse patient cohort. The total mean extraction accuracy of 92.4% in the external TCGA cohort is quantitatively similar to the 94.3% reported for the internal development cohort. This finding is significant as it indicates that the model has learned fundamental pathological principles rather than mimicking institutional reporting artifacts, retaining high performance without the need for site-specific fine-tuning.

While overall performance was high, error analysis highlighted challenges inherent to multi-institutional data curation. The highest variability was observed in fields such as surgical_technique in colorectal cancer (48.0% accuracy) and pm_category (distant metastasis) in esophageal cancer (62.5% accuracy). These discrepancies largely reflect known limitations in the ground truth definitions within the TCGA metadata (e.g., inconsistent M1 status annotations) rather than model failure. Furthermore, classification errors were largely attributable to complex bilateral breast cancer cases which, while correctly identified by the model as breast tissue, fell outside the current single-primary schema scope. On core oncological variables such as histology and grade, the model maintained 100% accuracy, supporting its utility as a reliable "digital registrar."

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[35]. 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[36]. 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[37].

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.

4. Methods

4.1 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 4**. 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.

4.2 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[38], 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 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-S10**.

4.3 Model architecture and workflow.

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.

4.3.1 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”.

4.3.2 Organ detection

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

4.3.3 Organ-specific extraction

Ten sets of organ-specific DSPy modules (for example, for breast cancer, 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. The pipeline is illustrated in **Figure 3**. 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. The readers are encouraged to see **Supplementary Appendix S25** for examples of model output.

4.3.4 Validation of Model Agnosticism and Hardware Feasibility

To validate the architectural independence of the extraction framework and assess the feasibility of deployment on standard workstation hardware, the identical pipeline structure was instantiated and executed using two alternative open-weight transformer architectures: qwen3-30b (Sparse Mixture-of-Experts) and gemma3:27b (Dense). The foundational design of the AI pipeline explicitly embraces principles of model agnosticism, utilizing the DSPy 2.1 framework to decouple the extraction logic from specific model weights.

All comparative models were executed on-premises using the same dedicated NVIDIA RTX A6000 Ada Generation GPU (48 GB VRAM). This hardware

constraint was strictly enforced to evaluate whether models could operate efficiently within the memory envelope of a standard high-end medical workstation, without requiring multi-GPU datacenter configurations.

4.4 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 5**.

4.5 Independent external validation

To assess the robustness and transportability of the extraction model to external, non-institutional data, we utilized the TCGA-Reports dataset [33]from The Cancer Genome Atlas (TCGA) repository. This dataset serves as a rigorous gold standard for external validation, comprising machine-readable pathology reports from 32 tissues/cancer types generated by a wide variety of institutions, ensuring high structural and linguistic heterogeneity.

The model was applied in an "off-the-shelf" capacity, without any further training or fine-tuning, to a validation subset of 150 reports spanning Breast, Colorectal, Esophagus, Stomach, and "Others" categories. We evaluated both the classification (triage) performance and the extraction fidelity for 53 specific pathological fields across the Breast (22 fields), Colorectal (17 fields), and Esophagus (14 fields) cohorts.

4.6 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.

4.7 Usage of large language models in 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 and gpt-oss:120b 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.

5. Reference

1. Piñeros, M. *et al.* The role and utility of population-based cancer registries in cervical cancer surveillance and control. *Preventive Medicine* **144**, 106237 (2021).

2. Lancet, T. Cancer registries: the bedrock of global cancer care. *The Lancet* **405**, 353 (2025).

3. Nair, S., Ngwa, W., Addai, B. W., Addai, A. O. & Oti, B. A. The role of a national cancer registry in effective implementation and monitoring of the Global Breast Cancer Initiative (GBCI) framework in Ghana: a narrative review. *Translational Breast Cancer Research* **6**, (2025).

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* e2400110 (2024) doi:10.1200/CCI.24.00110.

5. Hands, I. & Kavuluru, R. A survey of NLP methods for oncology in the past decade with a focus on cancer registry applications. *Artif Intell Rev* **58**, 314 (2025).

6. Hailer, L., Miller, J. & Chapman, S. Cancer Registrar Workload and Staffing Study: Guidelines for Hospital Cancer Registry Programs. (2024).

7. Penberthy, L. T., Rivera, D. R., Lund, J. L., Bruno, M. A. & Meyer, A.-M. An overview of real-world data sources for oncology and considerations for research. *CA: A Cancer Journal for Clinicians* **72**, 287–300 (2022).

8. Moor, M. *et al.* Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).

9. Singhal, K. *et al.* Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).

10. Tertulino, R., Antunes, N. & Morais, H. Privacy in electronic health records: a systematic mapping study. *J Public Health (Berl.)* **32**, 435–454 (2024).

11. Hernandez-Boussard, T., Blayney, D. W. & Brooks, J. D. Leveraging Digital Data to Inform and Improve Quality Cancer Care. *Cancer Epidemiol Biomarkers Prev* **29**, 816–822 (2020).

12. Bagherzadeh, P., Sultanem, K., Batist, G. & Abbasinejad Enger, S. An automatic pipeline for temporal monitoring of radiotherapy-induced toxicities in head and neck cancer patients. *npj Precis. Onc.* **9**, 40 (2025).

13. Paiva, B. *et al.* A New Natural Language Processing–Inspired Methodology (Detection, Initial Characterization, and Semantic Characterization) to Investigate Temporal Shifts (Drifts) in Health Care Data: Quantitative Study. *JMIR Medical Informatics* **12**, e54246 (2024).

14. Saluja, R. *et al.* Cancer type, stage and prognosis assessment from pathology reports using LLMs. *Sci Rep* **15**, 27300 (2025).

15. Chen, D., Alnassar, S. A., Avison, K. E., Huang, R. S. & Raman, S. Large Language Model Applications for Health Information Extraction in Oncology: Scoping Review. *JMIR Cancer* **11**, e65984 (2025).

16. Hung, T. K. W. *et al.* Human vs. AI: A comparative effectiveness study of large language models for automated biomarker extraction. *J Clin Oncol* **43**, e13605–e13605 (2025).

17. Wiest, I. C. *et al.* A software pipeline for medical information extraction with large language models, open source and suitable for oncology. *npj Precis. Onc.* **9**, 313 (2025).

18. Bartels, S. & Carus, J. From text to data: Open-source large language models in extracting cancer related medical attributes from German pathology reports. *Int J Med Inform* **203**, 106022 (2025).

19. Wiest, I. C. *et al.* Privacy-preserving large language models for structured medical information retrieval. *npj Digit. Med.* **7**, 257 (2024).

20. Bednarczyk, L. *et al.* Scientific Evidence for Clinical Text Summarization Using Large Language Models: Scoping Review. *Journal of Medical Internet Research* **27**, e68998 (2025).

21. Tripathi, A. *et al.* Employing Consensus-Based Reasoning with Locally Deployed LLMs for Enabling Structured Data Extraction from Surgical Pathology Reports. *medRxiv* 2025.04.22.25326217 (2025) doi:10.1101/2025.04.22.25326217.

22. Bumgardner, V. K. C., Mullen, A., Armstrong, S., Hickey, C. & Talbert, J. Local Large Language Models for Complex Structured Medical Tasks. Preprint at https://doi.org/10.48550/arXiv.2308.01727 (2023).

23. Lee, D. *et al.* Using Large Language Models to Automate Data Extraction From Surgical Pathology Reports: Retrospective Cohort Study. *JMIR Formative Research* **9**, e64544 (2025).

24. Balasubramanian, J. B. *et al.* Leveraging large language models for structured information extraction from pathology reports. Preprint at https://doi.org/10.48550/arXiv.2502.12183 (2025).

25. Grothey, B. *et al.* Comprehensive testing of large language models for extraction of structured data in pathology. *Commun Med* **5**, 96 (2025).

26. Rajaganapathy, S. *et al.* Synoptic Reporting by Summarizing Cancer Pathology Reports using Large Language Models. *medRxiv* 2024.04.26.24306452 (2024) doi:10.1101/2024.04.26.24306452.

27. Hao, Y. *et al.* Large language model integrations in cancer decision-making: a systematic review and meta-analysis. *npj Digit. Med.* **8**, 450 (2025).

28. Atil, B. *et al.* Non-Determinism of 'Deterministic' LLM Settings. Preprint at https://doi.org/10.48550/arXiv.2408.04667 (2025).

29. Khattab, O. *et al.* Dspy: Compiling declarative language model calls into state-of-the-art pipelines. in *The Twelfth International Conference on Learning Representations* (2024).

30. Kumar, D., Yadav, D. & Patel, Y. GPT-OSS-20B: A Comprehensive Deployment-Centric Analysis of OpenAI's Open-Weight Mixture of Experts Model. Preprint at https://doi.org/10.48550/arXiv.2508.16700 (2025).

31. Yang, A. *et al.* Qwen3 Technical Report. Preprint at https://doi.org/10.48550/arXiv.2505.09388 (2025).

32. Team, G. *et al.* Gemma 3 Technical Report. Preprint at https://doi.org/10.48550/arXiv.2503.19786 (2025).

33. Kefeli, J. & Tatonetti, N. TCGA-Reports: A machine-readable pathology report resource for benchmarking text-based AI models. *Patterns (N Y)* **5**, 100933 (2024).

34. Dennstädt, F., Hastings, J., Putora, P. M., Schmerder, M. & Cihoric, N. Implementing large language models in healthcare while balancing control, collaboration, costs and security. *NPJ Digit Med* **8**, 143 (2025).

35. Waqas, A., Tripathi, A., Ramachandran, R., Stewart, P. & Rasool, G. Multimodal data integration for oncology in the era of deep neural networks: a review. *Frontiers in Artificial Intelligence* **7**, (2023).

36. Chen, C. *et al.* Evidence-based diagnostic reasoning with multi-agent copilot for human pathology. Preprint at https://doi.org/10.48550/arXiv.2506.20964 (2025).

37. Witowski, J. *et al.* Multi-modal AI for comprehensive breast cancer prognostication. *ArXiv* **abs/2410.21256**, (2024).

38. Torous, V. F. *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* 47–55 (2021) doi:10.1200/CCI.20.00104.

6. Data availability statement

6.1 Internal Development Cohort

The primary dataset consists of de-identified surgical pathology reports from China Medical University Hospital. Due to patient privacy protections, these data are not publicly available but may be accessed by qualified researchers upon reasonable request to the corresponding author, subject to a Data Use Agreement (DUA).

6.2 External Validation Cohort (TCGA Benchmark)

To facilitate reproducibility and benchmarking, the complete external validation dataset—including the 150 selected TCGA pathology reports, the expert-annotated ground truth labels, and the model outputs—has been deposited in Zenodo and is publicly available at https://doi.org/10.5281/zenodo.17689361 . The original source data remains accessible via the NCI Genomic Data Commons (GDC) Data Portal.

6.3 Code availability

The complete source code for the "Digital Registrar" extraction pipeline, including all DSPy signatures, organ-specific schema definitions, and the orchestration logic used for this study, is openly available on GitHub at https://github.com/kblab2024/digitalregistrar . The repository includes instructions for reproducing the environment and deploying the specific open-weight models (gpt-oss:20b, qwen3-30b, gemma3:27b) utilized in the benchmarking and validation experiments. Upon acceptance, a permanent DOI will be minted via Zenodo to ensure long-term archival stability.

7. Acknowledgments

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.

9. Author contributions

K.P.C. conceived the study, oversaw the project, and wrote the original draft. H.K.C. developed the core deep learning algorithms and performed the technical implementation. C.Y.L. and Y.L.L. provided technical consultation and optimized the software architecture. N.H.C. and H.C. provided clinical resources, pathological data, and supervised the clinical interpretation. P.Y.T. assisted in data curation and validation. P.C.C. provided critical guidance on the methodology and revised the

manuscript. Y.W.C. contributed to the bioinformatics analysis. L.J.S. assisted with manuscript editing and linguistic refinement. All authors read and approved the final manuscript.

10. Competing Interests

We declare no competing interests.

11. 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 114-2813-C-039-132-E, and China Medical University Hospital, under grant number DMR-115-078. The funder played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

12. Figure Legends

Figure 1. Per-organ classification performance metrics. Bar chart summarizing the classification fidelity across ten target organ systems (N=893). The model achieved perfect recall (1.0) for eight distinct organ classes. High precision and F1-scores (>0.95) were maintained across all categories, confirming the model's reliability as a routing engine for downstream extraction modules. Slight performance dips in the "Lung" and "Esophagus" categories were largely attributable to the administrative exclusion of complex double-primary malignancies (labeled as "Others" in ground truth).

Figure 2. Lymph-node extraction accuracy across organ systems. Bar plots show the model’s performance in extracting lymph-node information across ten organ systems. The blue bars represent total lymph-node accuracy, defined as the exact-match concordance rate for the aggregated examined and positive lymph-node counts within each organ’s regional nodal basin. The red bars represent station-level accuracy, which requires an exact match for every individual lymph-node station, including station name, examined count, and positive count.

Figure 3. Surgical margin extraction accuracy across organ systems. Bar plots compare two levels of surgical-margin extraction performance across ten organ systems. The blue bars represent margin-positivity accuracy, defined as the exact-match concordance for determining whether each reported surgical margin is involved by tumor. The red bars represent margin-distance accuracy, a stricter metric that requires exact matching of both the margin involvement status and the reported tumor-to-margin distance (in millimeters) for every margin element. Because prostate cancer reports do not define multiple margin stations and instead document margin involvement using only overall positivity, extent of involvement, and the specific anatomic sites involved, prostate margins were excluded from this station-based margin accuracy evaluation.

Figure 4. Extraction pipeline for structured cancer data from surgical pathology reports.

13. Tables

Table 1. Comparative Performance and Efficiency of Locally Deployed LLMs
Benchmarks were conducted on a single NVIDIA RTX A6000 (48GB VRAM). Note that while Qwen3 offers high accuracy, its memory footprint exceeds the optimal efficiency envelope for single-GPU workstations, resulting in significantly higher latency.

Model	Architecture Type	Total Parameters	Mean Exact-Match Accuracy (193 Registry Fields)	Typical Processing Speed(Per Complex Report)
gpt-oss:20b	Sparse MoE	20B	94.30%	40 – 70 seconds
qwen3-30b-A3B	Sparse MoE	30B	92.90%	140 – >200 seconds
gemma3:27b	Dense	27B	89.80%	50 – 130 seconds

Table 2. Confusion Matrix of the Eligibility Triage Classifier (Run 1)

The classifier was evaluated on 893 real-world pathology reports. "Positive" indicates the report was identified as a valid cancer excision case eligible for registry abstraction.

	Predicted Eligible (Cancer Excision)	**Predicted Ineligible (Biopsy / Benign / Other)**	**Total**
Actual Eligible	693 (TP)	1 (FN)	694
Actual Ineligible	29 (FP)	170 (TN)	199
Total	722	171	893

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

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 4. Composition of the cancer excision cohort (n = 694)

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 5. Evaluation metrics and definitions

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.

14. Figures

Figure 1

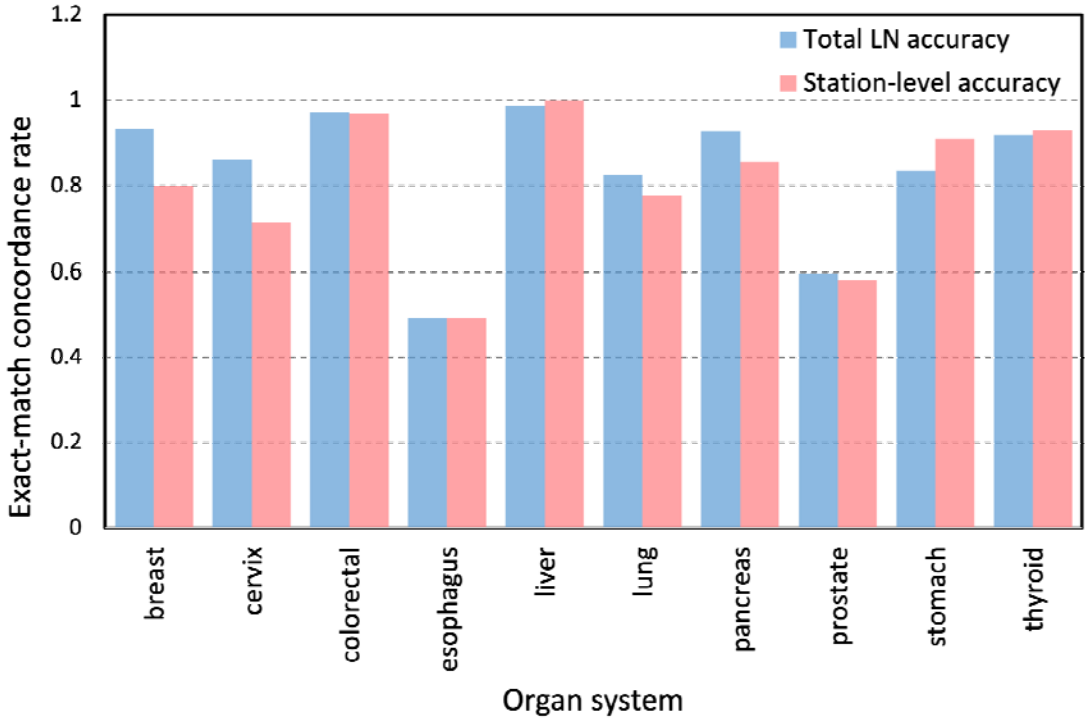


Figure 2

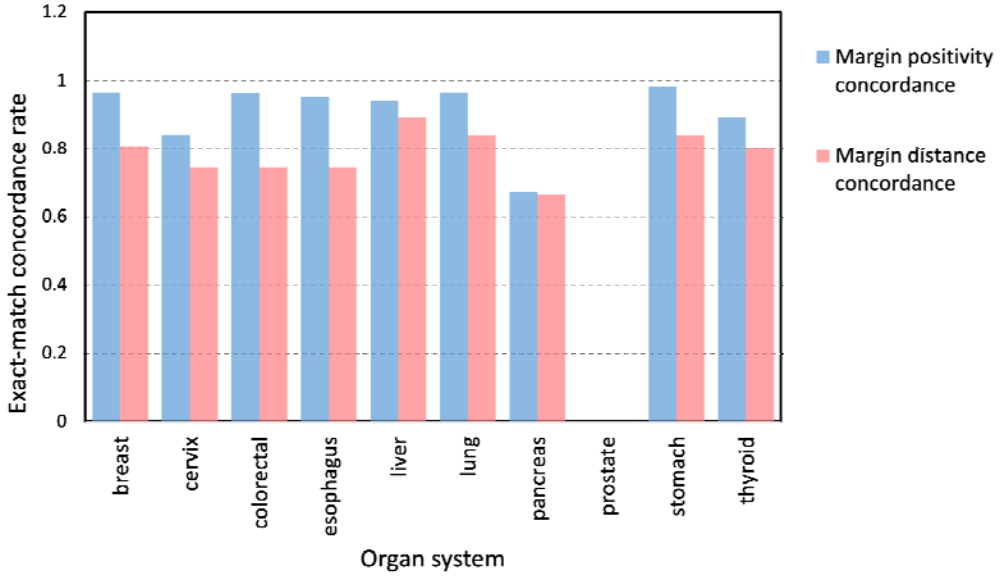
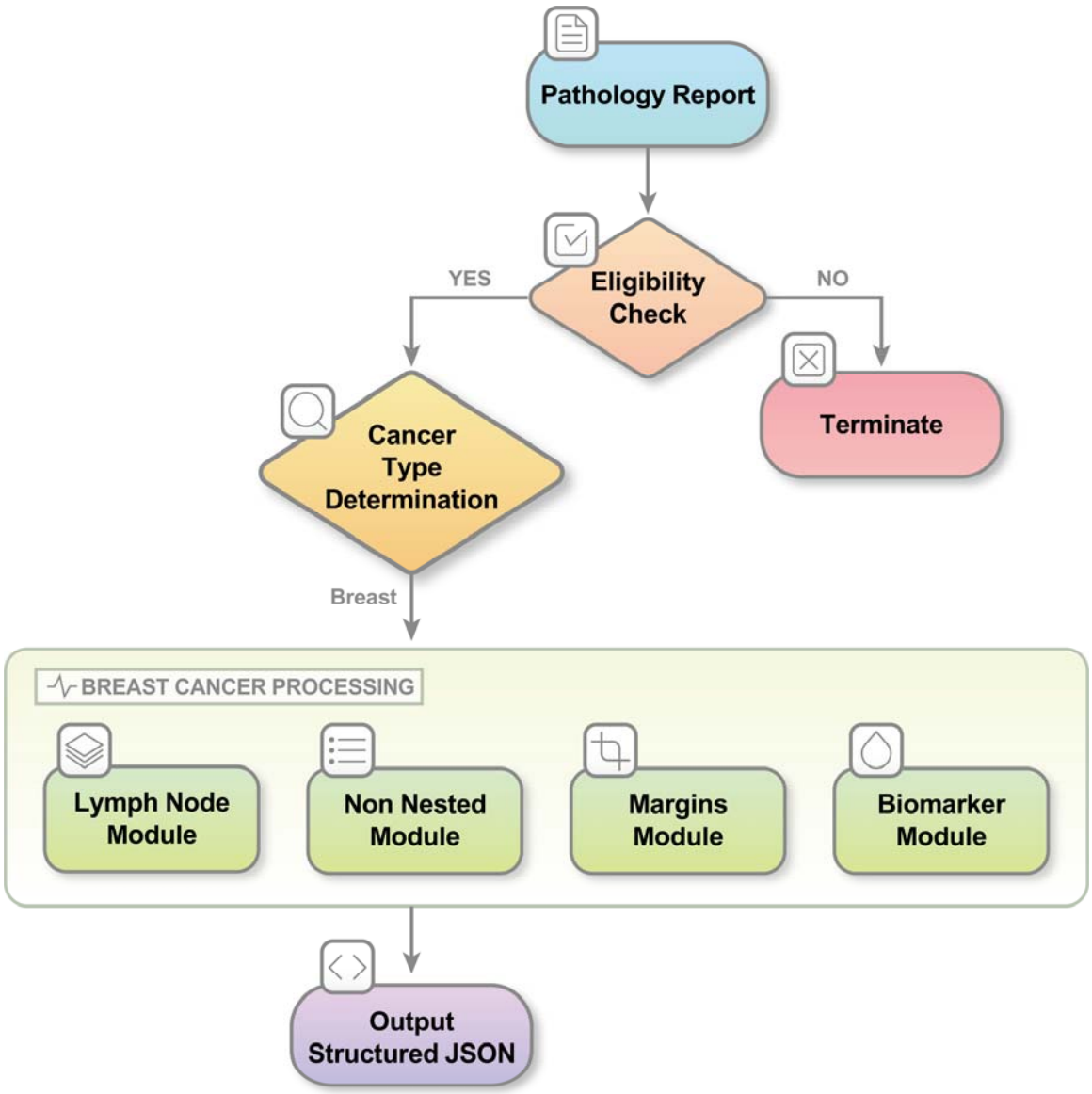

1.2
1
0.8
0.6
0.4
0.2
0
Exact-match concordance rate
breast
cervix
colorectal
esophagus
liver
lung
pancreas
prostate
stomach
thyroid
Organ system
Margin positivity concordance
Margin distance concordance

Figure 3


Pathology Report
Eligibility Check
YES
NO
Terminate
Cancer Type Determination
Breast
BREAST CANCER PROCESSING
Lymph Node Module
Non Nested Module
Margins Module
Biomarker Module
Output Structured JSON