npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01508-2

Continuous multimodal data supply chain and expandable clinical decision support for oncology

Jee Suk Chang[1,7], Hyunwook Kim[2,7], Eun Sil Baek[3], Jeong Eun Choi[4,6], Joon Seok Lim[5], Jin Sung Kim[1] ✉ & Sang Joon Shin[2] ✉

The study introduces a clinical decision support system (CDSS) developed at a single academic cancer center, integrating real-time clinical, genomic, and imaging data for over 170,000 patients across 11 cancer types. We have developed the Yonsei Cancer Data Library (YCDL) data integration framework to continuously collect and update multimodal datasets comprising over 800 features per case. Quality control measures, using 143 logical comparisons, addressed missing data and outliers, achieving median accuracies of 92.6% for surgical and 98.7% for molecular pathology. An Extract-Transform-Load (ETL) process with natural language processing transformed unstructured data, enabling survival analyses stratified by tumor stage, which revealed significant stage-dependent differences. The CDSS dashboard visualizes patient trajectories and key milestones. User feedback from oncology professionals showed strong acceptance, with satisfaction scores exceeding 4 out of 5. This framework demonstrates the potential of multimodal data integration to enhance clinical decision-making and patient outcomes, with future research needed to validate its generalizability and scalability.

Oncology data is multidimensional and diverse, encompassing a vast array of information such as patient characteristics, stage, tumor, and imaging data[1]. The advent of electronic medical records (EMR) and emerging data sources has caused a transformative surge in health information[2]. This data deluge often exceeds human cognitive limits for decision-making[3] and has led oncology professionals to spend more time navigating EMR than engaging with patients to seek fragmented health data from disparate sources, which exacerbates burnout[4].

Fortunately, rapid advancements in computational techniques, notably machine learning and artificial intelligence (AI), herald new possibilities for harnessing extensive and intricate medical data for individualized, data-driven care[5]. These technologies have demonstrated potential in refining imaging[6] and pathology diagnostics[7], prognosticating clinical outcomes, optimizing radiation treatment planning[8], and accelerating drug development[9,10]. AI has also significantly impacted foundational research in oncology[11].

However, challenges related to validation and generalizability[12] mean that the current methodologies for data management and model development fall short of the maturity required for broad-scale AI adoption. Transitioning from the present ad-hoc data aggregation and curation approach to a dynamic "metadata supply chain" is essential for providing contextualized, robust data in real-time[13]. By capturing pivotal data in real-time, this metadata supply chain can lay the groundwork for a clinical decision support system that vividly maps patient journeys, potentially transforming clinical workloads. Recent advancements have also led to the development of other multimodal frameworks, integrating diverse data types such as clinical, genomic, and imaging data to enhance precision in patient care, exemplified by the MEDomics framework developed by Morin et al.[14].

In this study, our objective was to present our collaborative endeavor for establishing a comprehensive data supply chain in oncology. This system seamlessly integrates clinical, genomic, and imaging data, representing a persistent, flexible, and expandable model. The infrastructure holds the potential to expedite the development of clinical decision support systems and AI applications for risk stratification, diagnosis, and

[1]Department of Radiation Oncology, Yonsei Cancer Center, Yonsei University College of Medicine, Seoul, Republic of Korea. [2]Division of Medical Oncology, Department of Internal Medicine, Yonsei Cancer Center, Yonsei University College of Medicine, Seoul, Republic of Korea. [3]Songdang Institute for Cancer Research, Yonsei University College of Medicine, Seoul, Republic of Korea. [4]Office of Data Services at Division of Digital Health, Yonsei University Health System, Seoul, Republic of Korea. [5]Department of Radiology and Research Institute of Radiological Science, Severance Hospital, Yonsei University College of Medicine, Seoul, Republic of Korea. [6]Present address: Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, Republic of Korea. [7]These authors contributed equally: Jee Suk Chang, Hyunwook Kim. ✉e-mail: JINSUNG@yuhs.ac; SSJ338@yuhs.ac

treatment in oncology, paving the way for individualized patient-centered care.

Results

Data collection and infrastructure are illustrated in Fig. 1, with detailed procedures described in the Methods section. Through this process, at the time of analysis, the DB contained records of the feature sets of 171,128 individuals diagnosed with 11 different cancers at a single academic cancer center between January 2006 and March 2022 (Table 1). For each individual, 817 essential features in the common columns and a median of 61 features (range: 38–109) in the cancer-type-specific columns were updated daily and continuously. To facilitate the extraction of structured information from unstructured medical text documents, Natural Language Processing (NLP) techniques were applied during data processing.

During the quality control (QC) process, we established a comprehensive set of 143 human-driven logical comparisons, including 70 focused on identifying missing data, 41 ensuring temporal validity (e.g., the completion date of radiotherapy should coincide with or follow its initiation date), 15 pinpointing outlier data (such as age at menarche between 8 and 20 years), 13 selecting the relevant values among multiple time points, and 4 dedicated to spotting duplicated or inconsistent data. The QC logic outcomes showed consistent results across 11 different cancer types, comprising a total of 1,523 datasets. We initially set the estimated daily QC case count to 10%, which translated to approximately 81 cases per day.

We generated survival graphs for each of the 11 distinct cancers in our dataset, segmented by tumor stages (Fig. 2). As expected, except for prostate cancer, there is a significant variation in survival rates depending on the stage of cancer; generally, higher stages are associated with lower survival rates.

The efficacy of our data framework in rapidly generating and evaluating clinical hypotheses was demonstrated in a study, focusing on rectal cancer, published in 2022[15]. Following the initiation of the study design in December 2020 and subsequent approval from the institutional review board, researchers requested baseline data on patients, tumors, and treatments, as well as peripheral blood neutrophil and lymphocyte counts, spanning the period from the initial diagnosis to the respective dates of primary rectal surgery for study participants. Data abstraction for 1386 individuals was efficiently executed using our framework, encompassing a total of 14 distinct clinical features. All features were validated through meticulous chart reviews by researchers, with head-to-head comparisons ensuring the accuracy and reliability of the data prior to its utilization in the study. User feedback further supported its reliability and effectiveness. This proficiency enabled researchers to commence a pilot analysis in January 2021, merely a month post the initial data acquisition.

The results of evaluating the accuracy of the NLP models used in our ETL process are as follows. For the first analysis, the median number of features for surgical pathology and molecular pathology was 26 (range: 20–33) and 13 (range: 9–16), respectively. The median accuracy and missing rate for surgical pathology were 92.6% (range: 86.5–98.8%) and 4.9% (range: 0.5–10.7%), respectively. For molecular pathology, the median accuracy and missing rate were 98.7% (range: 92–100%) and 0.6% (range: 0–8%), respectively (Supplementary Table 1). For the second analysis, the NLP

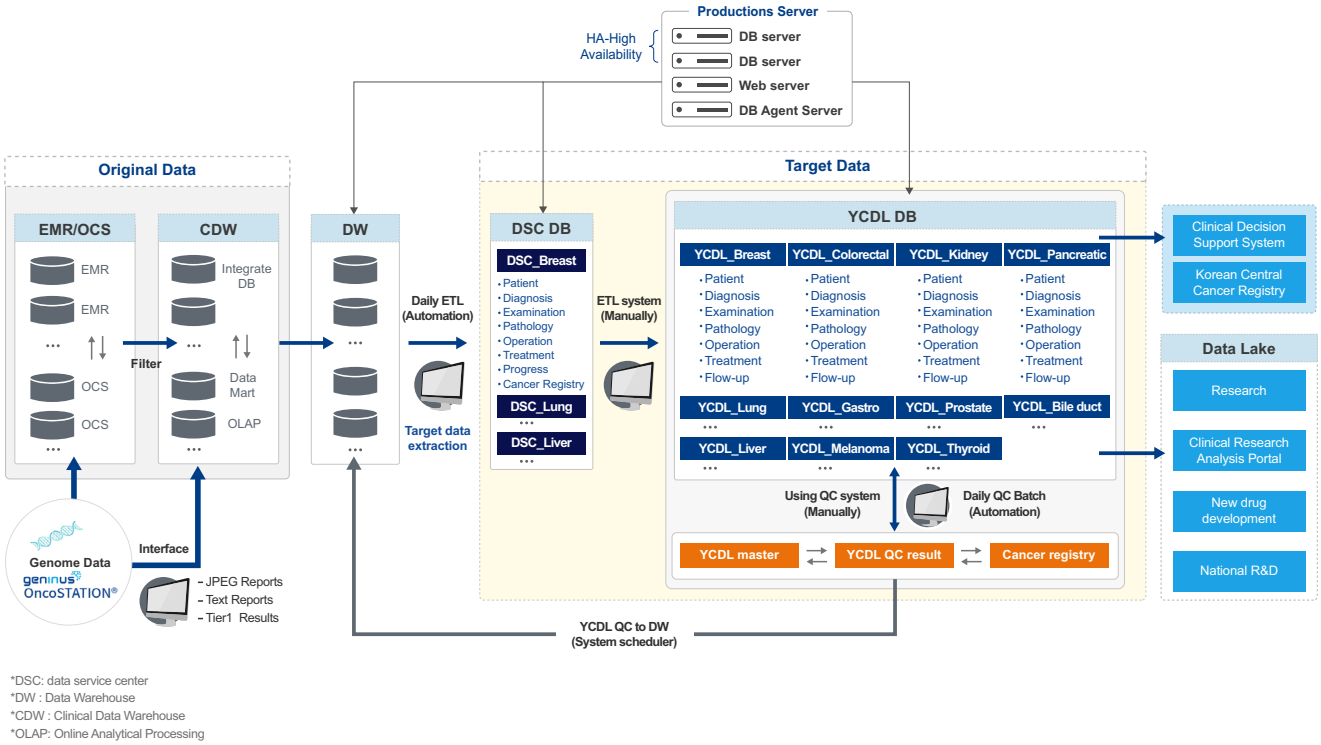

Fig. 1 | Overview of the YCDL framework. The original data from the EMR/OCS, which contains clinical data for all patients visiting the hospital, serves as the source data for the cancer-specific YCDL database. After the original data is transferred to the DW server, data marts are created from the DW tables in the DSC database, grouped by related topics. Separate databases are established for each cancer type, named "DSC_cancertype," to prevent excessive time spent on complex SQL query execution. The DSC database condenses data from 18 tables and 433 columns by integrating relevant tables from the DW database and joining with code master or terminology tables to include code-code name columns for immediate comprehension of codes. Similarly, the YCDL_DB maintains separate physical databases for each cancer type, where data is loaded. A patient-centric data model was developed, underpinned by patient identification numbers dispensed by the hospital information system, serving as a linchpin for linking anonymized datasets. All data processing, transfer, and storage were performed within the network infrastructure of the hospital. The YCDL site allows the execution of individual Data Manipulation Language (DML) to load YCDL data. The transfer of data from the original source to the DW/DSC DB is automated, with ETL processes running daily at 10 AM, transferring cancer-specific target data.

Table 1 | Number of patients added each year and, in every cohort

	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022.3	Total
Breast	764	811	737	837	959	880	891	740	938	1078	1358	1409	1811	1754	1526	1671	253	18,417
Colorectal	882	942	1006	1158	1090	1207	1143	979	1154	1254	1438	1413	1433	1387	1149	1046	222	18,903
Lung	712	740	756	768	817	840	884	902	943	1143	1257	1409	1653	1922	1737	1909	334	18,726
Gastric	1829	1815	1944	1865	1914	1964	1836	1696	1899	1915	1949	1918	1862	1522	1267	1433	254	28,882
Liver	657	589	656	737	679	655	621	569	589	609	657	610	616	646	565	489	112	10,056
Melanoma	49	43	73	74	87	67	58	75	86	99	113	91	108	107	123	105	24	1382
Kidney	207	245	265	311	305	300	317	301	358	358	440	439	466	588	430	436	41	5807
Prostate	378	529	621	755	697	760	774	713	711	779	1007	1105	1195	1432	1289	1,152	206	14,103
Thyroid	1535	2372	2841	2748	2783	2619	2909	2615	2037	1792	1975	2206	3,058	3201	2805	3575	695	41,766
Pancreas	246	254	274	298	290	313	362	324	326	442	546	455	526	605	545	631	174	6611
Bile duct	324	330	316	340	338	418	397	331	360	409	463	472	519	453	470	423	112	6475
Total	7583	8670	9489	9891	9959	10,023	10,192	9245	9401	9878	11,203	11,527	13,247	13,617	11,906	12,870	2427	171,128

classification model demonstrated accuracy across 1000 individual CT reports, with multilabel selection applied as follows: Complete Response/No Evidence of Disease (CR/NED), Partial Response (PR), Stable Disease (SD), Progressive Disease (PD), and Indeterminate (232) achieving an AUROC of 0.956 and an F1 score of 0.823. The model showed 72.3% (95% CI, 59.5–85.1) accuracy in predicting the day of disease progression within a ± 30-day window and 55.3% (95% CI, 41.1–69.5) accuracy in predicting the best response category and its timing within ±45 days. Notably, the model was more accurate in predicting CR/NED at 72% compared to SD and PR, which had accuracies of 27.3% and 15.4%, respectively.

We successfully developed a clinical decision support system with four layouts. In the upper-left layout, shown in Fig. 2, three selected image series are displayed alongside their corresponding three-dimensional tumor visualizations, using DICOM files of individually, manually contoured lesions (Supplementary Movie 1, Fig. 3a). In the middle-upper layout, the output of the longitudinal tumor tracking is in the form of a graph (Fig. 3b). The section with the hope of predicting individual patient outcomes has been reserved for future integration of any potential model (Fig. 3c). The lower layout presents a comprehensive overview of a patient's healthcare journey, allowing readers to intuitively understand the chronological sequence of events and progression of the patient's treatment (Fig. 3d, Supplementary Movie 2). This offers holistic and interactive patient summaries on a graphical timeline anchored by real-time data captured within our framework. Users can easily assess patient data in a temporal context with a single click, and the depth of information can be fine-tuned using zoom features and pop-up boxes.

To assess the satisfaction of using CDSS with EMR, we surveyed 33 healthcare oncology providers, including professors, residents, and physician assistants, for five randomly selected cases (breast, colorectal, lung, gastric, and liver cancer). The median EMR usage experience among the participants was 9 years (range: 1–19 years). The satisfaction scores of patient chart assessment using EMR along with CDSS are detailed in Table 2. The results showed that, in almost all areas, the scores averaged above 4 out of 5 points, with 5 being the highest possible score. While those with longer EMR usage experience (≥ 10 years) reported lower satisfaction with the user interface, there were no significant differences in other aspects based on the years of EMR usage experience.

Discussion

We successfully developed a cancer-specific information technology (IT) infrastructure designed to facilitate the longitudinal collection of comprehensive health data, an accomplishment realized through extensive cross-departmental collaboration. Using our IT infrastructure, we created a database that automatically updates the data on a regular basis, each with over 800 unique characteristics. Manually collating such an expansive array of features is challenging. To ensure data integrity, we initially implemented rigorous data QC methods, starting with manual logic applications and subsequently transitioning to an automated management system. This approach, conducted within closed-loop systems, led to a steady enhancement in data precision. while we did not verify the accuracy of all features, the simpler ETL processes demonstrated high accuracy, whereas more complex NLP tasks such as RECIST categorization highlighted the need for expert correction and ongoing model improvement and maintenance. For example, the suboptimal results in SD and PR suggest the limitations of evaluating RECIST criteria solely from textual information without incorporating imaging analysis. Another potential issue could be the unbalanced dataset across individual categories, as cases in SD and PR represent the smallest absolute numbers, which may affect predictive performance. Of practical significance, our system highlights the potential for real-time capture of disease state and treatment data, exemplified by a proof-of-concept for rapid clinical hypothesis testing and offering a holistic view of a patient's journey with a single click. This not only alleviates the clinical burden but also optimizes the research workflow. Survey results from oncology healthcare providers revealed generally high user satisfaction and strong expectations for the potential of CDSS (Table 2). Our evaluation

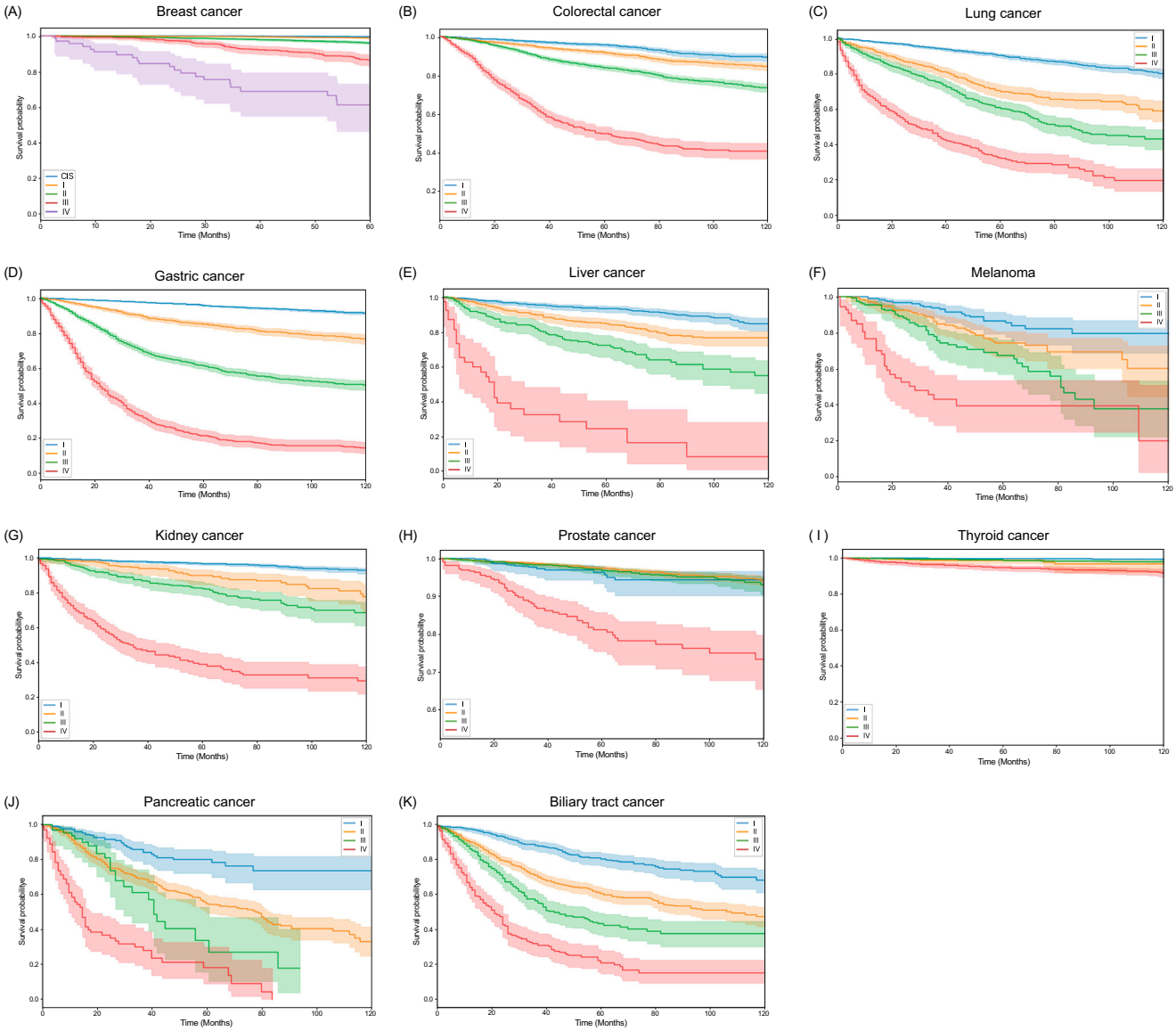

Fig. 2 | Kaplan–Meier survival curves by tumor stage for 11 cancer types. Kaplan–Meier survival curves for patients with cancer in YCDL target data, stratified by tumor stage. Survival rates were compared across 11 distinct cancer types: **A** breast cancer, **B** colorectal cancer, **C** lung cancer, **D** gastric cancer, **E** liver cancer, **F** melanoma, **G** kidney cancer, **H** prostate cancer, **I** thyroid cancer, **J** pancreatic cancer, and **K** biliary tract cancer. The curves illustrate stage-dependent survival variations, with generally lower survival rates at higher stages, except for prostate cancer where no significant variation was observed.

employs questionnaires designed for diverse healthcare professionals and encompasses multiple usability metrics—efficiency, effectiveness, and the ability to identify user errors. Nonetheless, further evaluations across different institutional and national settings, incorporating methods such as user trials, interviews, and heuristic evaluations[16], are essential to fully validate the system's utility and generalizability.

Understanding the crucial role of a reliable automatic data supply chain, several research groups have collaboratively developed frameworks to capture and transport oncologic data[14,17]. Morin et al.[14] introduced MEDomics, an information technology infrastructure that integrates seamlessly with multiple EHR DBs to ensure uniform data collection. Their research amassed data from nearly 175,000 patients with cancer at the University of California, San Francisco, between 2010 and 2019. Employing rule-based selection techniques, they identified individuals with high-quality data, narrowing them down to 3782 breast cancer and 2054 lung cancer cases. Lower-quality data were more prevalent among individuals located further away from the institution; a trend associated with increased mortality rates. Jung et al.[17] showcased ROOT, an auto-updating data warehouse that consolidates comprehensive clinical data of 67,617 individuals diagnosed with head and neck, thoracic, and esophageal cancers at the Samsung Medical Center in Korea between 2008 and 2020. These endeavors underscore the importance of data governance and active participation of all stakeholders. Considering geographic disparities and practice variations, in-house development might be best positioned to cater to the specific needs of end users.

Building an automated data warehouse using oncology EMR data poses inherent challenges because of the varying degrees of data completeness, inconsistencies, and conflicting or evolving records[18]. In this context, a nationwide initiative was launched to create a comprehensive cancer data library aimed at standardizing terminology and classification within our country. Concurrently, institutional efforts aimed to gather extensive feedback and integrate preexisting registries from diverse cancer groups. The recent proposal of Operational Ontology for Oncology (O3) seeks to achieve multi-institutional and multi-stakeholder consensus, lowering the barriers for collaborative information aggregation[19]. Our next task involves identifying differences and similarities between our defined features and the variables proposed by O3, and if possible, updating the

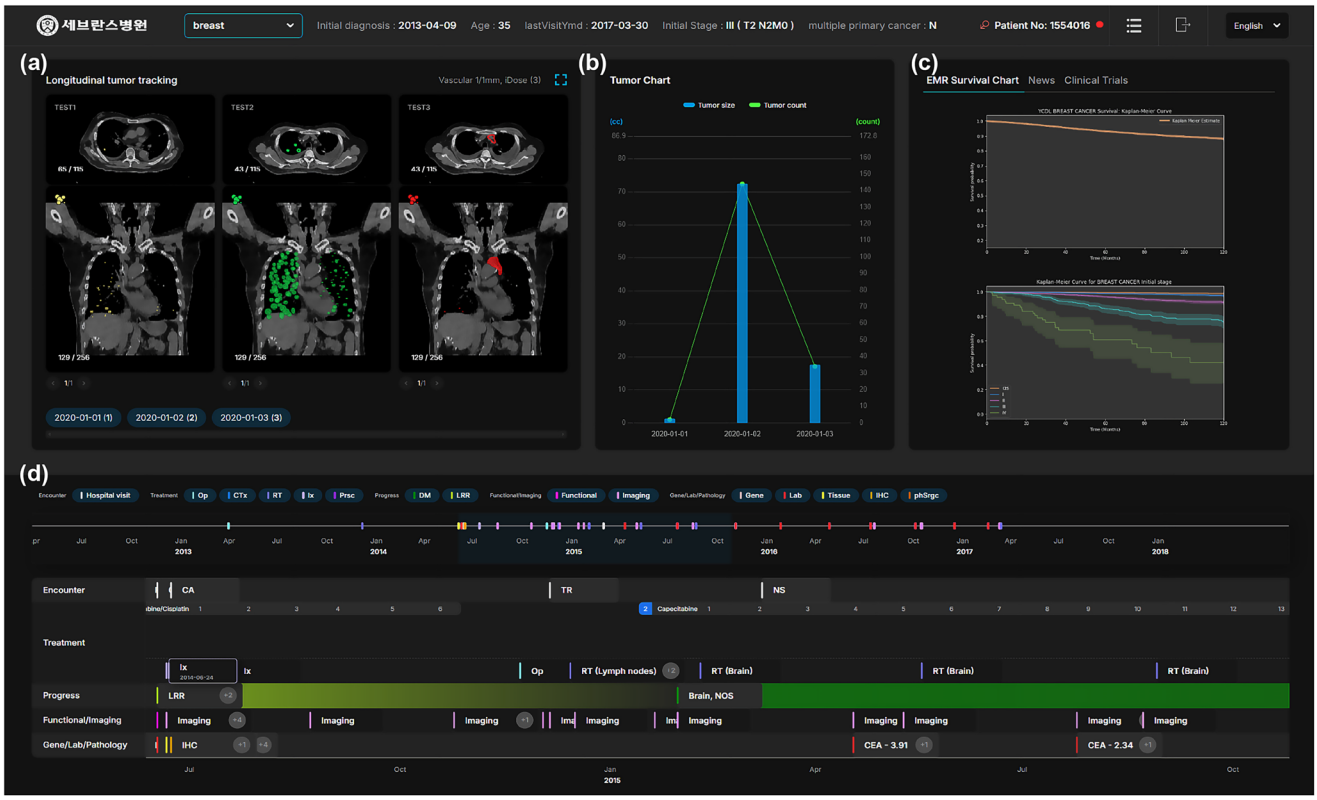

Fig. 3 | Proposed clinical decision support system with four layouts. a Three-dimensional display of overall disease burden, with individual lesions contoured manually or automatically in advance. **b** Longitudinal tumor tracking output in the form of a graph. **c** Section displaying survival curves for assessing and predicting individual patient outcomes by integrating any potential model. **d** Comprehensive overview of a patient's cancer journey including treatment history, follow-up, and disease status.

Table 2 | Satisfaction scores from 33 healthcare providers for comprehensive patient assessment using CDSS with EMR in 5 random cases

	Total (165 responses)		EMR usage < 10 years		EMR usage ≥ 10 years		
Satisfaction measures	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	***p***
Ease of system use	4.14	0.83	4.20	0.86	4.05	0.78	NS
Results understanding	4.26	0.75	4.29	0.83	4.22	0.60	NS
Terminology understanding	4.02	0.89	4.02	0.99	4.02	0.70	NS
Usefulness of the system	4.28	0.78	4.32	0.86	4.20	0.64	NS
User interface	3.97	0.93	4.09	1.00	3.78	0.76	<.001
Information accuracy	4.05	0.89	3.99	1.00	4.15	0.67	NS
Information timeliness	4.25	0.74	4.24	0.81	4.26	0.62	NS
Information reliability	4.01	0.87	3.98	1.01	4.06	0.58	NS
Up-to-datedness	4.35	0.81	4.38	0.90	4.31	0.64	NS
Decision support	4.21	0.79	4.22	0.84	4.20	0.71	NS
Processing time	4.09	0.93	4.17	0.99	3.97	0.81	NS
Task satisfaction	4.18	0.77	4.23	0.84	4.11	0.66	NS

SD standard deviation, *NS* not significant.

necessary parts. To manage the vast variability of data sources and types, we devised algorithms that harness structured data from diverse origins and process unstructured content using ETL procedures. ETL operations present unique challenges, especially when dealing with components presenting multiple ETL-related complications. Collaboration with team members well-versed in treatment workflows and medical informatics, combined with close cooperation with the IT department, was pivotal in understanding the system functionality and nuances of data interpretation. Both data governance and ethical deliberation are instrumental in ensuring data security and patient privacy.

In the absence of formalized frameworks, challenges may arise in query fulfillment and data management[20]. However, our data supply chain addresses this issue through an end-to-end workflow for data quality assurance, ensuring continual evaluation and improvement. Conflicting, missing, or incorrect data were identified through human-driven logical comparisons and rectified by making logical corrections or adjusting the algorithms. Since its implementation, the quality assurance workflow has been continuously refined, accumulating data checks across multiple cycles. This iterative process enhances data quality and reduces the need for human intervention. Engagement with various groups familiar with the data sources and limitations is essential.

Our YCDL framework has numerous potential clinical and research applications. Although limited data have evaluated clinically relevant outcomes in oncology care, emerging evidence suggests that CDSS using EMR data can positively impact care quality[21]. The most actively researched area is non-knowledge-based CDSS, which leverages machine learning and AI to predict patient outcomes[22], as follows: A recent randomized controlled trial by Hong et al.[23] demonstrated accurate triaging of patients with cancer and reduced acute care rates using an EMR-based machine learning algorithm. Coombs et al.[24] showed that a proposed machine learning tool using real-world EMR data could identify patients with cancer at risk for a 60-day emergency department visit. Another potential application is the generation and rapid testing of clinical hypotheses, as suggested by Morin et al.[14], which would not have been feasible using traditional data approaches. The YCDL enabled the collection of a vast amount of data, including laboratory results and patient

features, thereby facilitating the first pilot analysis. Additionally, automatic flagging of eligible patients for clinical trials shows promise[24].

Our study demonstrates that data consolidation and a continuous multimodal data supply chain can automatically generate visual timelines and enhance decision support—characteristics of a knowledge-based CDSS[22]. With advancements in systemic drugs, patients with stage IV cancer now live longer and have complex treatment histories[25]. A quick overview of a patient's cancer journey allows physicians to efficiently characterize both the disease and the individual, potentially reducing burnout and ensuring quality care[26]. Commercial clinical decision support software such as NAVIFY Oncology Hub[27], Syapse[28], and Flatiron Assist[29], are undergoing evaluation for integration into the EMR system to provide a comprehensive view of a patient's journey. Chen et al. demonstrated the potential of AI-assisted summarization tools using medical records, particularly when the input data is accurate[30]. With emerging local therapies[31], AI can play a significant role in detecting and segmenting normal tissues and tumors[32], as well as tracking lesions over time in relation to treatment[33]. However, further research on tumor auto-segmentation is warranted.

This study has several limitations that should be considered when interpreting our results. First, our method represents the experience of a single institution, and large-scale adjustments may be necessary for implementation elsewhere. Our system was not developed with direct consideration for interoperability standards such as HL7 and FHIR. Given that our ETL processes are based on our hospital's data, we have primarily focused on optimizing performance within our own hospital environment. However, as the national FHIR standard evolves and is finalized, we plan to expand our system accordingly to ensure compliance. Second, the data supply chain approach is designed as an expandable infrastructure that accommodates updated ontologies and evolving demands. Establishing a strong leadership in data governance, implementing sharing agreements, and promoting open science practices are essential for a robust metadata supply chain. This requires dedicated departments to ensure job security. Collaborative efforts such as workshops and knowledge transfers promote an understanding of the benefits offered by the metadata supply chain and AI technologies. Future work will incorporate additional cancer types such as brain tumors and rare malignancies. Once the ETL process is finalized, we aim to make it publicly accessible. Our hospital primarily diagnoses and follows up with patients within our institution; however, inter-hospital data sharing may become necessary in certain cases. The NLP models used in our ETL process, such as logic-based segmentation, demonstrated sufficiently good accuracy, and we believe that applying language models could further enhance these outcomes. While large language models have made remarkable progress in terms of performance and are likely to perform well in data center environments, their adoption may be limited by concerns over cost-effectiveness. In such scenarios, smaller language models—with their reduced computational requirements and reliance on less training data—could represent a practical and efficient option for addressing specific tasks[34]. Additionally, our study did not demonstrate whether multimodal data is superior to single-modal data in predicting patient outcomes[35]. Finally, the current version of the YCDL framework only captures survival and recurrence data despite the growing recognition of the importance of quality of life and toxicity profiles as critical outcomes.

In conclusion, this study underscores the critical role of developing a streamlined data integration framework to organize and visualize the substantial volume of oncology patient data, supporting and enhancing clinical decision-making. The integration of real-time updates into our framework is particularly significant, enabling the incorporation of evolving treatment trends and up-to-date information. This collaborative effort to establish a robust data infrastructure highlights its potential to advance personalized care, accelerate the adoption of AI-driven applications, and refine clinical workflows. Furthermore, adopting comprehensive data supply chains and AI technologies requires a commitment to strong data governance, the embrace of open science principles, and strengthened collaboration within the medical community.

Methods

Development of multimodal data supply chain

Our research was conducted in accordance with the Declaration of Helsinki after approval of the protocol by the Institutional Review Board of Severance Hospital (Reference Number: 4-2021-1241). The need for informed consent was waived by the ethics committee because the study involved retrospective analysis of anonymized data, posing minimal risk to participants. First, we established a development server using a Windows-based, 12-core computer with 64 GB of memory and Serial Attached SCSI (SAS) disk drives of 100 GB and 2 TB, and four RTX 5000 GPUs. Operational servers, constituting High Availability (HA) systems, included a database (DB) server (2Ea) with a 10-core CPU, 128 GB memory, OS SSD 100 GB of storage/SCL 2019, DB Safer, Hiware, EMS, and Backup (DB), and a web-based server with a 12-core CPU, 64 GB memory, OS SSD 100 GB of storage/Hiware, EMS, and Backup (File).

The dataflow and computational modules are illustrated in Fig. 3. The original data from the EMR/OCS, which contains clinical data for all patients visiting the hospital, serves as the source data for the cancer-specific YCDL database. Data access is strictly managed to ensure security and privacy. Only authorized personnel can access the data, and even then, they must go through two layers of identity verification within the hospital's internal network, which is isolated from the internet to prevent external threats. For research purposes, all data provided is either anonymized or pseudonymized to prevent the identification of individual patients, ensuring the protection of personal information while enabling research activities. Data are managed in compliance with ISO 27001 (international certification) and ISMS (domestic certification) standards for data protection regulations. The first data transfer from source data is governed by three conditions: the Target Patient Number (AlsUnitNo), a de-identified number assigned to each patient based on a diagnosis code matching the ICD-10 code of the primary cancer; the Target Encounter Number (AlsChosNo), a de-identified number generated during patient visits, which is selected as the encounter number for the primary cancer if the patient number from the first condition matches the target diagnosis code related to the primary cancer; and the creation of a Main Target Table, which distinguishes cancer types by the target patient and encounter numbers. If the target patient number exists in the tables on the DW server, all data related to the target encounter number are transferred. Cancer types are not distinguished in the DW tables. Additionally, the latest data from code master or terminology tables are also transferred. To improve the data quality and mitigate the risks associated with erroneous or omitted data, we tailored the selection approaches for each cancer type. The selection was based on the International Classification of Diseases for Oncology (ICD) and physician-assigned ICD-M codes as well as validity criteria designated by the cancer registration program. A comprehensive breakdown of the selection methodologies for each cancer type is provided in Table 3. We received authorization for access to all digital records from the EMR system and billing data from the Oncology Care System.

After the original data is transferred to the DW server, data marts are created from the DW tables in the Data Science Center (DSC) database, a proprietary name, grouped by related topics. Separate databases are established for each cancer type, named "DSC_cancertype," to prevent excessive time spent on complex SQL query execution. The DSC database condenses the data from 18 tables and 433 columns by integrating relevant tables from the DW database and joining with code master or terminology tables to include code-code name columns for immediate comprehension of codes. Similarly, the YCDL_DB maintains separate physical databases for each cancer type. Each cancer-specific database has tables where data is loaded. In this process, a patient-centric data model was developed, underpinned by the patient identification numbers dispensed by the hospital information system. This served as a linchpin for linking the anonymized datasets. In the clinical data extraction stage, we developed an Extract-Transform-Load (ETL) process, which includes NLP, for each feature (Fig. 4). It facilitated the daily movement of data from the DSC source DB to the YCDL target DB. The DSC DB is a repository that contains unstructured and semi-structured

Table 3 | Selection methods for each cancer type

Cancer Id	Cancer Type	DBName	Criteria
01	Breast cancer	YCDL_BRST	(1) Cancer Registry : ICDOCd[a]=C50% AND available=Y AND ICDOCdM[b] <M9590
02	Colorectal cancer	YCDL_CLRC	(1) Cancer Registry: ICDOCd = (C18%, C19%, C20%) AND ICDOCdM=M81403(Adenocarcinoma) AND available=Y
03	Lung cancer	YCDL_LUNG	(1) Cancer Registry : ICDOCd=C34% AND available=Y AND ICDOCdM < M9590 AND ICDOCdM NOT LIKE '%/2'
04	Gastric cancer	YCDL_GSTR	(1) Cancer Registry : ICDOCd=C16% AND available=Y AND available=Y AND ICDOCdM<M9590 AND ICDOCdM NOT LIKE '%/2'
05	Liver cancer	YCDL_LVER	(1) Cancer Registry : ICDOCd=C22.0 AND available=Y AND ICDOCdM < M9590 AND ICDOCdM NOT LIKE '%/2'
06	Melanoma	YCDL_MLNM	(1) Cancer Registry: ICDOCdM_EngNm (pathology) LIKE '%Melanoma%' AND available=Y(2) The cancer diagnosis group = D0023(Malignant melanoma) in CAP system[c].(3) There are records of '%Melanoma%', '%Malignant Spitz%' in the pathology diagnosis results.(4) There are records of '%Melanoma%', '%Malignant Spitz%' in the imaging test. (excluded '%r/o%')
07	Kidney cancer	YCDL_KDNY	(1) Cancer Registry : ICDOCd=C64% AND available=Y AND ICDOCdM < M9590 AND ICDOCdM NOT LIKE '%/2'
08	Prostate cancer	YCDL_PRST	(1) Cancer Registry : ICDOCd=C61% AND available=Y AND available=Y AND ICDOCdM<M9590 AND ICDOCdM NOT LIKE '%/2'
09	Thyroid cancer	YCDL_THRD	(1) Cancer Registry : ICDOCd=C73% AND available=Y AND ICDOCdM <M9590 AND ICDOCdM NOT LIKE '%/2'
10	Pancreatic cancer	YCDL_PNCT	(1) Cancer Registry : ICDOCd=C25% AND available=Y AND ICDOCdM < M9590 AND ICDOCdM NOT LIKE '%/2'
11	Bile duct cancer	YCDL_BLDT	(1) Cancer Registry : ICDOCd = (C22.1, C23.9, C24.0, C24.1, C24.8, C24.9) AND available=Y AND ICDOCdM < M9590 AND ICDOCdM NOT LIKE '%/2'

[a]ICDOCd = ICD-O (International Classification of Diseases for Oncology) Codes.
[b]ICDOCdM = Morphology section of the ICD-O Code.
[c]CAP system = Chemotherapy Assistance Program for ordering oncology medications.

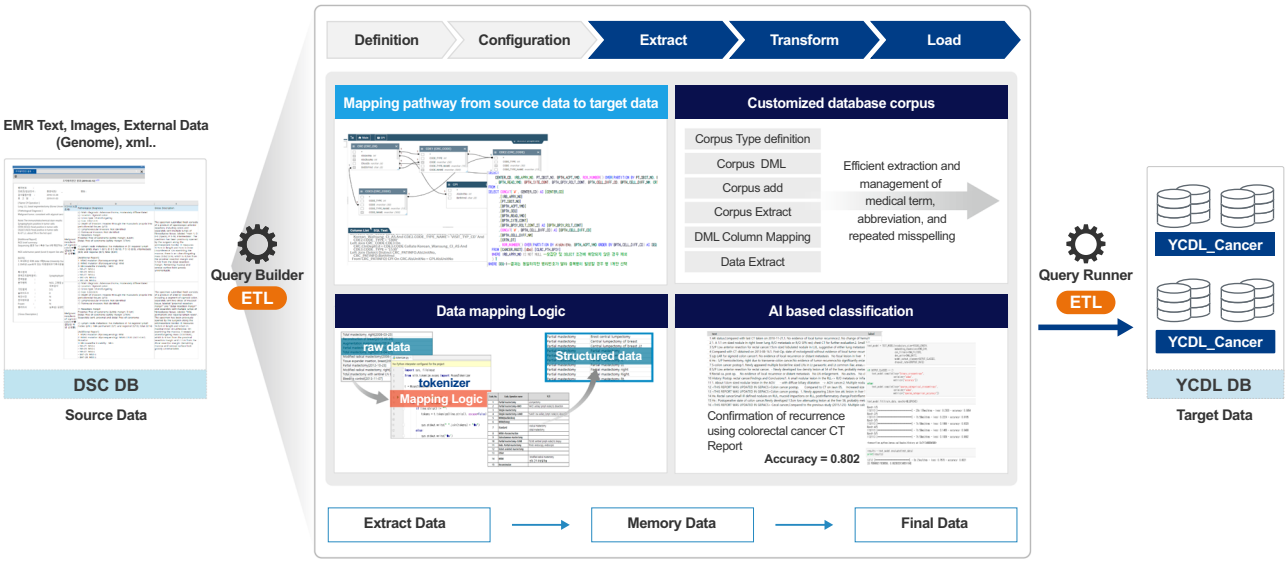

Fig. 4 | The Extract-Transform-Load (ETL) process within the YCDL framework. In the clinical data extraction stage, we developed an ETL process, which includes Natural Language Processing (NLP), for each feature. The DSC DB serves as a reservoir containing raw medical text, (semi-) unstructured data, imaging files, next-generation sequencing (NGS) results, and Extensible Markup Language (XML) formats. In the initial phase of data processing, we tailored the database corpus from the DSC DB, optimizing the extraction and management of medical terminology, abbreviations, and recurrent misspellings (e.g., within pathology reports). Subsequently, the procured data underwent transformation through a specialized ETL algorithm designed to harmonize terminology based on assertions and the interrelationships of medical concepts. NLP was instrumental in utilizing CT and MRI interpretation counts from follow-up visits as criteria for individual selection.

data, including medical record text, imaging files, and next-generation sequencing (NGS) results. The YCDL site enables the execution of individual database queries to extract data from the source system and load it into the YCDL target system. Examples of data extraction from DSC DB to YCDL are provided in the Supplementary Figs. 1-9, using SQL queries, MS-SQL user-defined functions, and Python user-defined functions. All data processing, transfer, and storage were performed within the network infrastructure of the hospital. The transfer of data from the original source to the DW/DSC DB is automated, with ETL processes running daily, transferring cancer-specific target data. Access to sensitive data is strictly controlled through a role-based permission system, which assigns access levels based on user roles and responsibilities. To ensure system reliability and data integrity, fail-safe mechanisms are in place, including automated daily backups, geographically redundant storage, and real-time system monitoring to detect and address issues proactively. The user management system further enforces security through role-based access control, enabling administrators to define and customize specific roles, permissions, and access rights tailored to each user group's needs.

In the initial phase of data processing, we tailored the database corpus from the DSC DB, optimizing the extraction and management of medical terminology, abbreviations, and recurrent misspellings (e.g., within pathology reports). Subsequently, the procured data underwent transformation through a specialized ETL algorithm designed to harmonize terminology based on assertions and the interrelationships of medical

Table 4 | Tables in the DSC database

DB No.	DB Name	DB code	Table No.	Table Name	Table Description
1	Patients	PT	1	CNCR_PATINFO	Patient basic information
1	Patients	PT	2	CNCR_BODYINFO	Body measurement information
2	Diagnosis	DG	3	CNCR_DX	Diagnoses relating to a hospital visits
2	Diagnosis	DG	4	CNCR_CRDINFO	Copayment Decreasing Policy
2	Diagnosis	DG	5	CNCR_CSLT	Consultant Information
3	Examination	EM	6	CNCR_LAB	Events relating to laboratory tests
3	Examination	EM	7	CNCR_IMAGE	Events relating to Imaging test
4	Pathology	PH	8	CNCR_PATHOLOGY	Events relating to Pathology
5	Operation	OP	9	CNCR_OP	Surgery
6	Treatment	TX	10	CNCR_REGIMEN	Chemo-therapy
6	Treatment	TX	11	CNCR_RT	Radiation-therapy
6	Treatment	TX	12	CNCR_DRUG	Medicines prescribed
6	Treatment	TX	13	CNCR_PROC	Procedure (included medical operation)
7	Progress	TE	14	CNCR_FRM	Clinical Forms
8	[a]Cancer registry	TM	15	CNCR_TUMOR_RGT	Tumor Registry (personal details and cancer diagnosis)
8	Cancer registry	TM	16	CNCR_TUMOR_TRANS	Tumor Registry (included cancer recurrence/metastasis)
8	Cancer registry	TM	17	CNCR_TUMOR_TRC	Tumor Registry (included cancer patient follow-up)
8	Cancer registry	TM	18	CNCR_TUMOR_TRET	Tumor Registry (included cancer treatment)

[a]Cancer registry = database of information on cancer patients.

concepts. To enhance user convenience in handling extensive cancer patient data, we developed a model based on imaging reports to determine the best responses and the timing of disease progression[36]. Imaging reports from 6574 patients were gathered, amounting to 97,119 CT readings. Among these, 9000 CT reports corresponding to 2859 patients were randomly subjected to multilabel manual labeling by four radiology experts, based on the RECIST version 1.1 classification (CR/NED, PR, SD, PD). The pre-trained BERT-base-uncased model was employed and fine-tuned for the downstream tasks of multilabel classification. 6765 reports were used for training, while the remaining 1000 reports were divided equally between the validation and test sets. The subsequent preprocessing phase employed tokenization techniques to structure the extracted data. SQL queries were harnessed to mine data from the primary DSC DB, facilitated by a DML management interface. For certain datasets requiring intricate extraction protocols, bespoke ETL strategies were devised using Python scripts crafted for each specific operation (Supplementary Fig. 10).

The overall process of NGS analysis has been detailed in our previous publication[37]. Targeted DNA and RNA sequencing was performed using the TruSight Tumor 170 (TST170, Illumina, San Diego, CA) and TruSight Oncology 500 (TSO500, Illumina) panels. Secondary analysis, including read alignment, variant calling, and variant annotation, utilized the TST170 Local App and the TSO500 Local App, respectively. For tertiary analysis, which involves additional annotation, variant filtering, prioritization, and producing interpretable output, we utilized an in-house pipeline designed to discard false positive variants, germline variants, and SNPs, ensuring the accuracy and reliability of the results. Variant interpretations were manually reviewed by institutional pathologists in accordance with guidelines from the Association for Molecular Pathology, the American Society of Clinical Oncology, and the College of American Pathologists[38]. Pathogenic variants categorized as Tier 1 were automatically and systematically collected and transferred to YCDL. A substantial proportion of the procedural steps were automated using OncoSTATION (Geninus, Seoul, Korea), as shown in Supplementary Fig. 11.

Data quality control and accuracy assessment

After development, we implemented this system with our electronic health data, beginning with records from 2006. The profiles were updated using electronic health records, ensuring a comprehensive view of relevant oncological components over time. The present analysis is based on data collected up to March 2022. Key constituents of these profiles included demographics, diagnoses, clinical examination reports, pathology reports, treatment histories, and encounter specifics (Tables 4 and 5). The structures of these individual profiles were categorized into common, cancer-specific, and index columns. The common features held universal information across multiple cancer types (e.g., age, sex, and cancer diagnosis date) and accounted for 817 features, which was nearly 80% of the total. The cancer-specific features contained data relevant only to specific cancer types (for instance, pulmonary function test in lung cancer) and comprised approximately 20% of the tables (Table 6). Data feature definitions and formats for all features across all cancer types have been included in the Supplementary data 1, with necessary translations from Korean to English.

We developed a web-based computational platform for QC of data that scrutinizes potential data defects both automatically and manually on a daily basis, focusing on minimizing the role of the human component (Fig. 5). All data extracted and stored in the YCDL_cancer data repository were continuously evaluated and optimized to establish high-quality data outputs, adhering to standardized data and terminology. Programs for logical checks were configured to evaluate the distribution and continuity of data extracted by the SCL. Based on the QC results, the ETL code was continuously modified, thereby refining the QC logic to enhance the quality and accuracy of the automation. We examined four data quality measures (completeness, timeliness and usefulness, consistency, and accuracy) for all variables, in accordance with established data standards and pertinent aspects of data quality (Table 7). For instance, the logic was set such that the birth date of individuals would precede the date of the initial diagnosis. The analyses revealed that the batch processing method accurately identified erroneous data points, aligned with the established logic. Each piece of data was meticulously reviewed and optimized by a Quality Control Manager. Significant discrepancies or inaccuracies prompted an in-depth examination of the source data and respective ETL processes. Moreover, a hierarchy of data sources was established to resolve conflicts. The QC steps were continuously iterated within distinct closed-loop systems, adhered to operational ontology, and executed by independent QC personnel. This methodology gradually enhanced the accuracy of the cleansed target data with minimal intervention (Supplementary Fig. 12). We assessed the completeness of each individual's accumulated features, including fundamental characteristics

Table 5 | Tables in the YCDL database and number of variables

	Table Category		Table Name	Table Description	Common	Breast	Colorectal	Lung	Gastric	Liver	Melanoma	Kidney	Prostate	Thyroid	Prostate	Bile duct	Total
1	PT	Patient	PT_BASIC	Patient Basic Information	20												20
2	PT	Patient	PT_PHIS	Past History	39	9	12						1	2			63
3	PT	Patient	PT_SHIS	Smoking History	29												29
4	PT	Patient	PT_DRNK	Drinking History	33												33
5	PT	Patient	PT_FMHS	Family History	41												41
6	PT	Patient	PT_BDMS	Body Measurement	27												27
7	DG	Diagnosis	DG_INFO	Visit Information	27												27
8	DG	Diagnosis	DG_ECHI	Copayment Policy	27												27
9	DG	Diagnosis	DG_CNCR	Cancer Diagnosis	37												37
10	DG	Diagnosis	DG_CONS	Consultant	28												28
11	EM	Examination	EM_LAB	Laboratory Test	30												30
12	EM	Examination	EM_IMEX	Imaging Test	28			2		2					2	2	36
13	EM	Examination	EM_GENE	Genetic Test	28												28
14	EM	Examination	EM_FCLT	Function Test	28	6		5		2				6	2	2	51
15	PH	Pathology	PH_BPSY	Biopsy	29	7	13	4	13	5	5	3	7	3	4	4	97
16	PH	Pathology	PH_SRGC	Histopathology	32	29	36	20	19	34	23	24	16	25	20	33	311
17	PH	Pathology	PH_IMML	Immuno-histology	29	13	17	9	10	10	17	15	8	15	12	12	167
18	OP	Operation	OP_INFO	Operation Information	28												28
19	OP	Operation	OP_OPNN	Operation opinion	38	27	31	13	19	46	6	5	6	12	4	14	221
20	OP	Operation	OP_COMP	Operation Complication	25												25
21	TX	Treatment	TX_CHTH	Chemotherapy	36												36
22	TX	Treatment	TX_RTH	Radiotherapy	41												41
23	TX	Treatment	TX_PRSC	Drug	29												29
24	TX	Treatment	TX_MOPR	Procedure	33												33
25	TE	Follow-Up	TE_MTST	Follow-Up Metastasis	27												27
26	TE	Follow-Up	TE_RCRN	Follow-Up Relapse	23												23
27	TE	Follow-Up	TE_DEAD	Dead	25												25

such as date of birth, initial diagnosis date, age, diagnosis code (ICD), TNM and overall stages, and ICDO morphology code.

To verify the accuracy of the NLP models used in our ETL process, we first evaluated the segmentation accuracy of logic-based NLP in both surgical and molecular pathology reports. We randomly selected 50 items from the top 30% in data completeness for each cancer type, where data completeness was defined as the ratio of columns filled with segmented data to the total number of pathology report features for each cancer type. Assessing the timing of the best response category is critical in understanding clinical outcomes during retrospective cancer patient data analysis. Rapid access to this information improves the efficiency of evaluating disease progression and treatment responses. Accordingly, we assessed the NLP model's accuracy in automatically categorizing RECIST criteria[36], predicting the timing of disease progression events within a ± 30-day window, and identifying the best response and its timing within a ± 45-day window.

Table 6 | Column characteristics by cancer type

	Number of Common Columns (A)	Number of Cancer-specific Columns (B)	Number of Index Columns (C)	Number of Total Columns (D)	Percentage of Cancer-specific Columns (B/D)
Breast	817	91	459	908	10%
Colorectal	817	109	459	926	12%
Lung	817	53	459	870	6%
Gastric	817	61	459	878	7%
Liver	817	99	459	916	11%
Melanoma	817	51	459	868	6%
Kidney	817	47	459	864	5%
Prostate	817	38	459	855	4%
Thyroid	817	63	459	880	7%
Pancreatic	817	44	459	861	5%
Bile duct	817	67	459	884	8%

The developed data warehouse showcased survival graphs by tumor stages and demonstrated the framework's ability to expedite data collection for quick clinical hypothesis testing. Kaplan–Meier survival graphs were generated in all cancer types according to tumor stage with 95% confidence intervals. Survival time was defined as the time interval between initial diagnosis and death or the last follow-up. To demonstrate the efficiency of our data framework as a proof-of-concept for swiftly generating and evaluating clinical hypotheses, we present a detailed chronological progression of a previously published retrospective study. The clinical question chosen by one of the authors was whether the peripheral blood neutrophil-to-lymphocyte ratio before, during, or after neoadjuvant chemoradiotherapy for locally advanced rectal cancer is associated with an increased risk of distant metastases after primary rectal cancer surgery.

CDSS development and evaluation

To underscore the capabilities of our data framework for clinical applications, we developed a CDSS with a comprehensive and modular architecture to support efficient digital content creation and management, utilizing data from the YCDL server. The current User Interface (UI) front-end components are as follows: (1) patient information, (2) DICOM image visualization for PACS-integrated three-dimensional tumor display (viewing and interactive visualization of medical images and segmentation information in DICOM format), and (3) a longitudinal view of the complete patient journey (timeline visualization of patient data over time, highlighting key events and data points). Additionally, components for survival prediction based on data from previously treated patients, personalized news/journals, and clinical trial information are being developed for integration. The CDSS front end was built using JavaScript frameworks such as Svelte or React. It provides an interactive UI for users and handles the visualization of data received from the CDSS backend. Key functionalities include user interaction management and data visualization module processing. The CDSS backend is implemented using FastAPI, a high-performance web framework for building APIs with Python. It processes data requests from the front end and performs computations for various modules. Key functionalities include handling RESTful API requests, DICOM data processing, web scraping services, and visualization of DICOM data using the VTK library. The Front-End Modules include a 2D/3D Visualization Module and a Timeline Module. The 2D/3D Visualization Module visualizes DICOM

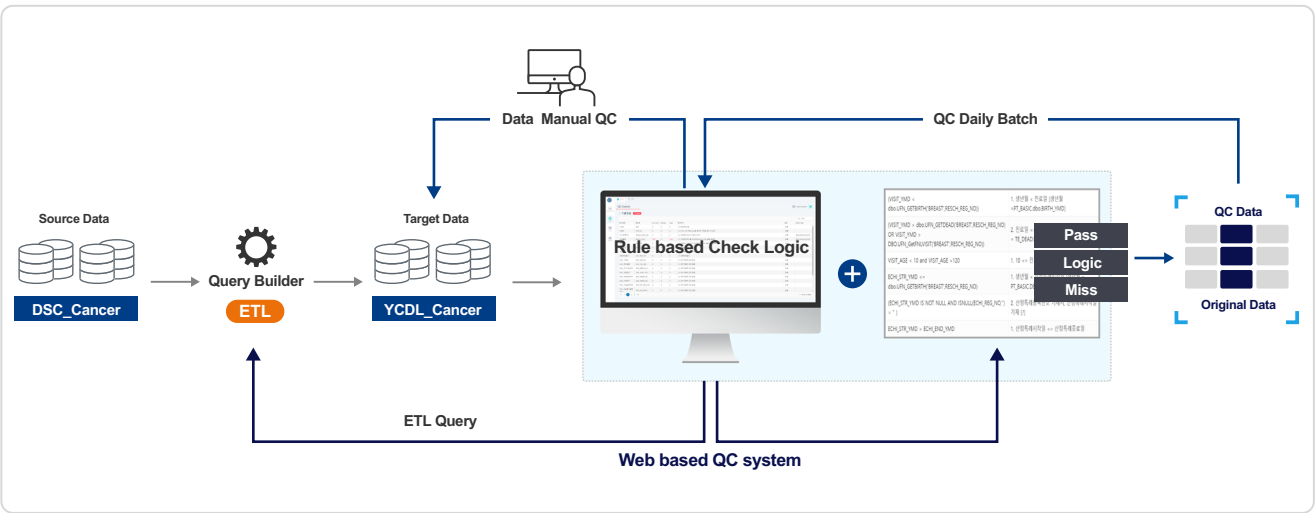

Fig. 5 | Quality management of data in the YCDL framework. We developed a web-based computational platform for data quality control (QC) that scrutinizes potential data defects both automatically and manually on a daily basis, focusing on minimizing the role of the human component. All data extracted and stored in the YCDL_cancer data repository were continuously evaluated and optimized to establish high-quality data outputs, adhering to standardized data and terminology. Programs for logical checks were configured to evaluate the distribution and continuity of data extracted by the SCL. Based on the QC results, the ETL code was continuously modified, thereby refining the QC logic to enhance the quality and accuracy of the automation. The analyses revealed that the batch processing method accurately identified erroneous data points, aligned with the established logic. Each piece of data was meticulously reviewed and optimized by a Quality Control Manager.

Table 7 | Data quality check criteria

Quality Indicators	Detailed Quality Indicators	Diagnostic Targets	Remarks
Completeness	Individual Completeness	Columns or input values defined to exist but are Null	
	Conditional Completeness	Checking for NOT NULL constraints	
	Structural Completeness	Implementation based on the physical model designed from the schema, including data types	Verified at the DB design stage
Validity	Code Validity	Whether codes defined in the common code are used	
	Format Validity	Errors in data format	Verified at the DB design stage
	Boolean Validity	Diagnosis based on columns with Y/N, 0/1 criteria	
	Date Validity	Errors based on date formats	
	Range Validity	Diagnosis based on Min, Max, and Normal range of the column	
	Temporal Relationship Validity	Diagnosis of data that deviates from predetermined sequential relationships	
Consistency	Referential Integrity	Diagnosis of operation rules for PK (Primary Key) items	Verified at the DB design stage
Accuracy	Logical Relationship Accuracy	Data diagnosis according to logical relationships, e.g., when item A is n, item B should be at least m	
	Derived Item Accuracy	Diagnosis of derived data, e.g., whether the sum of item A and item B is equal	

images received from the Back-End as VTK images, supporting both 2D and 3D visualization of medical imaging data. The Timeline Module, based on EMR data, visualizes patient data as a timeline, arranging data chronologically to show key events and data points, and supporting interactions like zooming and dragging. The backend modules include an AAA Module and a DICOM Image Transformation Module. The AAA Module handles authentication, authorization, and accounting using JWT (JSON Web Token) for secure processing. The DICOM Image Transformation Module converts and processes DICOM images into VTK images using the VTK library. The overview of the described architecture is depicted in Supplementary Fig. 13.

The data flow is as follows: Users initiate requests through the web UI. The front end processes the user's request and routes it to the appropriate frontend module. The backend processes data requests and interacts with the required backend modules. The latter processes EMR and DICOM data to generate or retrieve necessary information. The processed data is returned through each layer, ultimately displaying results in the User Interface. This design bolsters the accessibility of the system, guarantees platform independence, and ensures that users can access services across various device types.

Manual tumor segmentation data are required to use the three-dimensional tumor display with a longitudinal tumor-tracking function. If deep learning-based tumor auto-segmentation algorithms are developed, these models can be integrated into the pipeline[39,40]. The PACS-integrated method enables physicians to comprehensively track changes in overall trajectory patterns over a long period, fosters an environment that better explains the disease course to patients, and facilitates communication with referring physicians. Longitudinal changes in the overall disease burden were automatically generated using the prepared manual contours and displayed as graphs. The images were de-identified; however, if another image of the same patient was transferred later, the new images were allocated the same de-identified number, facilitating tumor tracking.

To investigate the effectiveness of the CDSS, we conducted a mock simulation with physicians ($n = 33$), comparing EMR-only assessments to assessments using both the EMR and CDSS. The simulation incorporated user evaluations to gather physician feedback on the patient assessment process. After selecting the five randomly selected cases of patients with breast cancer, colorectal cancer, lung cancer, stomach cancer, and liver cancer, participants were requested to complete a survey for each case after assessment of cases using EMR and CDSS. Using the evaluation framework from Kim et al.[41], we investigated a total of 12 measures. For system quality, we assessed ease of system use, results understanding, terminology understanding, usefulness of the system, and user interface. For information quality, we evaluated information accuracy, information timeliness, information reliability, and up-to-dateness. For support factors, we examined decision support, processing time, and task satisfaction. All measures were coded on a 6-point scale, with 5 being the highest score and 0 being the lowest score.

Data availability

The Yonsei University Health System (YUHS) inaugurated the Severance Data Portal (SDP), a comprehensive medical big data platform, on 2 May 2023 (available at: https://sobig.yuhs.ac/portal). The SDP provides an accessible portal tailored for the research community, with a focus on medical investigations. It is supported by a 'Data Lake' search portal that empowers researchers to locate and harness extensive data sets aligned with their specific research goals. In the forthcoming expansion phase, YUHS intends to enhance the platform through the integration of pioneering digital medical imaging information systems, such as Picture Archiving and Communication Systems (PACS), along with digital pathology data and genomic analysis datasets. Access to the SDP is governed by stringent policies devised to safeguard patient confidentiality and to ensure adherence to all pertinent legal and ethical standards. Researchers aiming to utilize the SDP must submit an access application specifying the proposed data usage, which is then subjected to a thorough review process to ensure compliance with established data governance criteria.

Code availability

The custom code and scripts used in the generation and analysis of datasets for this study are not publicly available due to institutional restrictions. However, researchers interested in accessing the code can do so by contacting the corresponding author directly. Access to the code will be granted on a case-by-case basis, contingent upon appropriate IRB approval and institutional permissions. The specific versions of software used in this study include TensorFlow 2.10, NumPy 1.23.5, and pandas 2.0.0, Microsoft SQL Server 2019, JavaScript (ES11, ES2020), and Python version 3.10.8. The version of YCDL used in this study was v1.4. The key variables and parameters used in the data analysis are uploaded in the supplementary data (Excel file format).

Received: 14 January 2024; Accepted: 9 February 2025;
Published online: 27 February 2025

References

1. Figueiredo, E. B. D., Dametto, M., Rosa, F. D. F. & Bonacin, R. A multidimensional framework for semantic electronic health records in oncology domain. In *2021 IEEE 30th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)* 165-170 (2021).
2. Heart, T., Ben-Assuli, O. & Shabtai, I. A review of PHR, EMR and EHR integration: A more personalized healthcare and public health policy. *Health Policy Technol.* **6**, 20–25 (2017).

3. Abernethy, A. P. et al. Rapid-learning system for cancer care. *J. Clin. Oncol.* **28**, 4268–4274 (2010).
4. Shanafelt, T. D. et al. Relationship Between Clerical Burden and Characteristics of the Electronic Environment With Physician Burnout and Professional Satisfaction. *Mayo Clin. Proc.* **91**, 836–848 (2016).
5. Davenport, T. & Kalakota, R. The potential for artificial intelligence in healthcare. *Future Health. J.* **6**, 94–98 (2019).
6. Huang, S. C. et al. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ Digit. Med.* **3**, 136 (2020).
7. Cui, M. & Zhang, D. Y. Artificial intelligence and computational pathology. *Lab Investig.***101**, 412–422 (2021).
8. Huynh, E. et al. Artificial intelligence in radiation oncology. *Nat. Rev. Clin. Oncol.* **17**, 771–781 (2020).
9. Perez-Lopez, R., Reis-Filho, J. S. & Kather, J. N. A framework for artificial intelligence in cancer research and precision oncology. *NPJ Precis Oncol.* **7**, 43 (2023).
10. Shreve, J. T., Khanani, S. A. & Haddad, T. C. Artificial intelligence in oncology: current capabilities, future opportunities, and ethical considerations. *Am. Soc. Clin. Oncol. Educ. Book* **42**, 1–10 (2022).
11. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
12. Ramspek, C. L. et al. External validation of prognostic models: what, why, how, when and where? *Clin. Kidney J.* **14**, 49–58 (2021).
13. Chung, C. & Jaffray, D. A. Cancer needs a robust "Metadata Supply Chain" to realize the promise of artificial intelligence. *Cancer Res* **81**, 5810–5812 (2021).
14. Morin, O. et al. An artificial intelligence framework integrating longitudinal electronic health records with real-world data enables continuous pan-cancer prognostication. *Nat. Cancer* **2**, 709–722 (2021).
15. Yang, G. et al. Association of neutrophil-to-lymphocyte ratio, radiotherapy fractionation/technique, and risk of development of distant metastasis among patients with locally advanced rectal cancer. *Radiat. Oncol.* **17**, 100 (2022).
16. Wohlgemut, J. M. et al. Methods used to evaluate usability of mobile clinical decision support systems for healthcare emergencies: a systematic review and qualitative synthesis. *JAMIA Open* **6**, ooad051 (2023).
17. Jung, H. A. et al. Real-time autOmatically updated data warehOuse in healThcare (ROOT): an innovative and automated data collection system. *Transl. Lung Cancer Res.* **10**, 3865–3874 (2021).
18. Kanas, G. et al. Use of electronic medical records in oncology outcomes research. *Clinicoecon. Outcomes Res.* **2**, 1–14 (2010).
19. Mayo, C. S. et al. Operational Ontology for Oncology (O3) - A Professional Society Based, Multi-Stakeholder, Consensus Driven Informatics Standard Supporting Clinical and Research use of "Real -World" Data from Patients Treated for Cancer: Operational Ontology for Radiation Oncology. *Int, J. Radiat. Oncol. Biol. Phys.* https://doi.org/10.1016/j.ijrobp.2023.05.033 (2023).
20. Khare, R. et al. Design and Refinement of a Data Quality Assessment Workflow for a Large Pediatric Research Network. *EGEMS (Wash. DC)* **7**, 36 (2019).
21. Pawloski, P. A., Brooks, G. A., Nielsen, M. E. & Olson-Bullis, B. A. A systematic review of clinical decision support systems for clinical oncology practice. *J. Natl. Compr. Canc Netw.* **17**, 331–338 (2019).
22. Sutton, R. T. et al. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digit. Med.* **3**, 17 (2020).
23. Hong, J. C. et al. System for high-intensity evaluation during radiation therapy (SHIELD-RT): a prospective randomized study of machine learning-directed clinical evaluations during radiation and chemoradiation. *J. Clin. Oncol.* **38**, 3652–3661 (2020).
24. Coombs, L. et al. A machine learning framework supporting prospective clinical decisions applied to risk prediction in oncology. *NPJ Digit. Med.* **5**, 117 (2022).
25. Colicchio, T. K., Cimino, J. J. & Del Fiol, G. Unintended consequences of nationwide electronic health record adoption: challenges and opportunities in the post-meaningful use era. *J. Med. Internet Res.* **21**, e13313 (2019).
26. Pivovarov, R. & Elhadad, N. Automated methods for the summarization of electronic health records. *J. Am. Med. Inf. Assoc.* **22**, 938–947 (2015).
27. Goh, E. et al. Remote evaluation of NAVIFY Oncology Hub using clinical simulation. *J. Clin. Oncol.* **41**, e13622–e13622 (2023).
28. Hirsch, J., Ford, J. M., Nadauld, L. & Hsu, A. Design and implementation of an informatics infrastructure for actionable precision oncology. *J. Clin. Oncol.* **33**, e17521–e17521 (2015).
29. Maniago, R. et al. Implementation of an EHR-embedded decision support tool in community oncology practices. *J. Clin. Oncol.* **39**, 274–274 (2021).
30. Chen, P.-H. C. C. et al. AI-assisted clinical summary and treatment planning for cancer care: A comparative study of human vs. AI-based approaches. *J. Clin. Oncol.* **42**, 1523–1523 (2024).
31. Liu, W., Bahig, H. & Palma, D. A. Oligometastases: emerging evidence. *J. Clin. Oncol.* **40**, 4250–4260 (2022).
32. Primakov, S. P. et al. Automated detection and segmentation of non-small cell lung cancer computed tomography images. *Nat. Commun.* **13**, 3423 (2022).
33. Cai, J. et al. Deep Lesion Tracker: Monitoring Lesions in 4D Longitudinal Imaging Studies. In *Proc. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 15154–15164 (IEEE Computer Society, 2021).
34. Lu, Z. et al. Small language models: survey, measurements, and insights. *arXiv* https://doi.org/10.48550/arXiv.2409.15790 (2024). Preprint (2024).
35. Boehm, K. M. et al. Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. *Nat. Cancer* **3**, 723–733 (2022).
36. Kim, H. et al. A rapid assessment tool for systemic treatment outcomes in colorectal cancer with deep bidirectional transformers. *J. Clin. Oncol.* **42**, e15567–e15567 (2024).
37. Cha, Y. J. et al. Clinicopathological Characteristics of NRG1 Fusion-Positive Solid Tumors in Korean Patients. *Cancer Res Treat.* **55**, 1087–1095 (2023).
38. Li, M. M. et al. Standards and guidelines for the interpretation and reporting of sequence variants in cancer: a joint consensus recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. *J. Mol. Diagn.* **19**, 4–23 (2017).
39. Buchner, J. A. et al. Development and external validation of an MRI-based neural network for brain metastasis segmentation in the AURORA multicenter study. *Radiother. Oncol.* **178**, 109425 (2023).
40. Cassinelli Petersen, G. et al. Real-time PACS-integrated longitudinal brain metastasis tracking tool provides comprehensive assessment of treatment response to radiosurgery. *Neuro-Oncol. Adv.* **4**, vdac116 (2022).
41. Kim, J. et al. A study on user satisfaction regarding the Clinical Decision Support System (CDSS) for medication. *Health. Inf. Res* **18**, 35–43 (2012).

Acknowledgements

This study was funded by the Big Data Center at the National Cancer Center of Korea (grant number: 2020-data-we08). This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: RS-2023-KH136094).The funder played no role in study design, data collection, data analysis, data interpretation, or writing of this manuscript. Portions of the content of this paper were presented at the 2023 CARO-COMP Joint Scientific Meeting (September 22, 2023, Montreal, Canada) and the Practical Big Data Workshop 2023 (May 19, 2023, Ann Arbor, MI).

Author contributions

J.S.C., J.S.K., and S.J.S. designed the research. E.S.B., H.K. and J.E.C. collected the data. S.J.S. verified the raw data. E.S.B., J.S.K., and S.J.S. developed E.T.L. and C.D.S.S. J.S.C., H.K., E.S.B,. J.E.C., J.S.L., J.S.K., and S.J.S. analyzed the results. J.S.C. wrote the manuscript, H.K., E.S.B., and S.J.S. critically revised the manuscript, and all authors provided feedback. All authors had full access to all the data in the study and read and approved the final manuscript.

Competing interests

The following pending patent applications are related to this manuscript: (1) Yonsei University has filed a domestic (Korea) patent application with the application number 10-2021-0003683, covering aspects of the extract-transform-loading (ETL) system discussed in this research. The inventors are S.J.S., E.S.B., and J.E.C. (2) Yonsei University has also filed an international patent application under the Patent Cooperation Treaty (PCT) with the application number PCT/KR2021/007295, covering aspects of the longitudinal tumor tracking using CT imaging system discussed in this research. The inventors are S.J.S., J.S.K., J.S.L., J.S.C., and others. Both patent applications are currently in the filed status as of July 30, 2024. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01508-2.

Correspondence and requests for materials should be addressed to Jin Sung Kim or Sang Joon Shin.