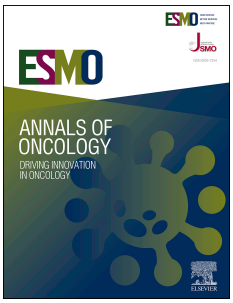
Next-Generation Multicenter Studies: Using Artificial Intelligence to Automatically Process Unstructured Health Records of Patients with Lung Cancer across Multiple Institutions

Mihaela Aldea, L. Zullo, V. Levrat, J. Bennouna, S. Schneider, O. Mercier, E. Mougenot, E. Bergot, C. Dujon, N. Cloarec, et al.

Journal Pre-proof



Next-Generation Multicenter Studies: Using Artificial Intelligence to Automatically Process Unstructured Health Records of Patients with Lung Cancer across Multiple Institutions

M. Aldea, L. Zullo, V. Levrat, J. Bennouna, S. Schneider, O. Mercier, E. Mougenot, E. Bergot, C. Dujon, N. Cloarec, C.Audigier Valette, A. Nuccio, M. Deloger, C. Helissey, S. Simon, A. Carpentier, A. Djarallah, P. Rolland, J.C. Louis, L. Ancillon, B. Vignal, F. Rambaud, P. Tessier, L. Chuttoo, K. Siby, A. Poplu, K. Zarca, S. Michiels, F. Barlesi, F. Le Ouay, B. Besse

PII: S0923-7534(25)06320-3

DOI: https://doi.org/10.1016/j.annonc.2025.12.006

Reference: ANNONC 2034

To appear in: *Annals of Oncology*

Received Date: 23 August 2025

Revised Date: 28 November 2025

Accepted Date: 5 December 2025

Please cite this article as: Aldea M, Zullo L, Levrat V, Bennouna J, Schneider S, Mercier O, Mougenot E, Bergot E, Dujon C, Cloarec N, Valette CA, Nuccio A, Deloger M, Helissey C, Simon S, Carpentier A, Djarallah A, Rolland P, Louis JC, Ancillon L, Vignal B, Rambaud F, Tessier P, Chuttoo L, Siby K, Poplu A, Zarca K, Michiels S, Barlesi F, Le Ouay F, Besse B, Next-Generation Multicenter Studies: Using Artificial Intelligence to Automatically Process Unstructured Health Records of Patients with Lung Cancer across Multiple Institutions, *Annals of Oncology* (2026), doi: https://doi.org/10.1016/j.annonc.2025.12.006.

errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Next-Generation Multicenter Studies: Using Artificial Intelligence to Automatically Process Unstructured Health Records of Patients with Lung Cancer across Multiple Institutions

Population & Method

Multicenter study (10 sites): automated extraction of 31 variables from unstructured health records.
Centrally hosted LLM system.
Trained on 10,016 patients, tested on 311.

Manual
(Traditional abstraction)

AI
(fined-tuned LLM extraction)

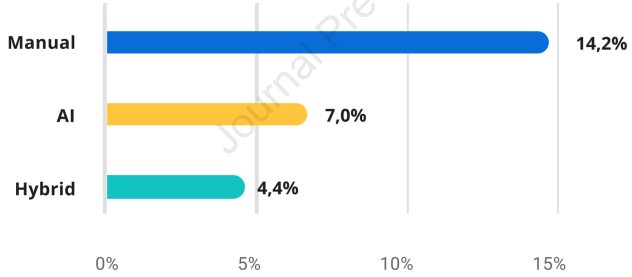
Hybrid
(AI + 30% targeted human review)

Ground truth: concordant data across sources; discrepancies resolved by blinded expert.

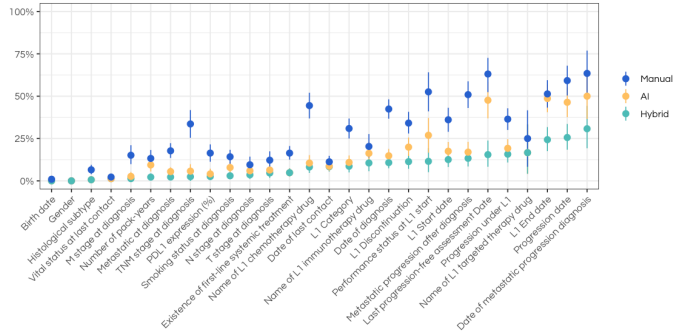
Results

AI use reduced error rates across sites and variables, and inter-center variability: 0.39 (manual), 0.12 (AI).
Time reduction (minutes per patient): 17.5 minutes (manual), 1.7 minutes (AI), 4.6 minutes (hybrid).

Error rates (%)



Error rates per variable (%)



Conclusion

We show that an LLM-based system can accurately extract lung cancer variables from unstructured medical records. It can reduce errors, standardize data across centers, and enable scalable, inclusive multicenter studies.

Next-Generation Multicenter Studies: Using Artificial Intelligence to Automatically Process Unstructured Health Records of Patients with Lung Cancer across Multiple Institutions

Authors:

M. Aldea[1,2], L. Zullo[1], V. Levrat[3], J. Bennouna[4], S. Schneider[5], O. Mercier[6], E. Mougenot[7], E. Bergot[8], C. Dujon[9], N. Cloarec[10], C. Audigier Valette[11], A. Nuccio[1], M. Deloger[1], C. Helissey[12], S. Simon[1], A. Carpentier[13], A. Djarallah[13], P. Rolland[13], J.C. Louis[13], L. Ancillon[13], B. Vignal[13], F. Rambaud[13], P. Tessier[13], L. Chuttoo[13], K. Siby[13], A. Poplu[13], K. Zarca[13], S. Michiels[14], F. Barlesi[1,2], F. Le Ouay[13]*, B. Besse[1,2]*

*shared senior authorship

Affiliations

1. Department of Medical Oncology, Thoracic and Precision Medicine Group, Gustave Roussy, Villejuif, France
2. Paris-Saclay University, Paris, France
3. Department of Pulmonology, Groupe Hospitalier de La Rochelle-Ré-Aunis, La Rochelle, France
4. Department of Medical Oncology, Hôpital Foch, Suresnes, France
5. Department of Pulmonology, Centre Hospitalier de la Côte Basque, Bayonne, France
6. Pulmonology and Lung Transplantation Department, Marie-Lannelongue Hospital, Groupe Hospitalier Paris-Saint-Joseph, Le Plessis-Robinson, France
7. Clinical Research Unit, Centre Hospitalier d'Auxerre, Auxerre, France
8. Department of Pulmonology and Thoracic Oncology, CHU de Caen Normandie, Caen, France
9. Department of Pulmonology, Centre Hospitalier de Versailles André Mignot, Le Chesnay, France
10. Department of Medical Oncology, Centre Hospitalier d'Avignon, Avignon, France
11. Department of Thoracic Oncology, Centre Hospitalier Intercommunal Toulon-La Seyne-sur-Mer, Toulon, France
12. Department of Medical Oncology, Groupe Hospitalier Paris-Saint-Joseph, Paris, France
13. Lifen, Paris, France
14. Office of Biostatistics and Epidemiology, Gustave Roussy, Université Paris-Saclay, Oncostat U1018, Inserm, labeled Ligue Contre le Cancer, Villejuif, France

Corresponding author:

Dr. Mihaela Aldea

Department of Medical Oncology, Thoracic and Precision Medicine Group

Gustave Roussy

114 Rue Edouard Vaillant, 94800 Villejuif, France

Telephone number: +33142114211
Email: mihaela.aldea@gustaveroussy.fr

Short title: Validation of a Next-Generation AI-Assisted Multicenter Study

Manuscript social media information: @mihaela_aldea (Twitter account)

Word count: 4,777 text + (6 figures x 150 words each) = 5,677

Abstract

Background:
Manual abstraction of real-world data (RWD) from unstructured health records (HRs) remains resource-intensive, error-prone, and highly variable across institutions. Large language models (LLMs) offer a scalable alternative, but their performance in multicenter oncology settings is not fully validated.

Patients and Methods:
We conducted a multicenter study within the French Large & Unified Cancer Cohort (LUCC) consortium to compare the accuracy of artificial intelligence (AI)-based data extraction against manual abstraction by clinical research professionals. A fine-tuned LLM was applied to de-identified unstructured HRs in PDF format to extract 31 variables from lung cancer patients across 10 centers. Ground truth was defined as concordant values across sources, with discrepant cases adjudicated by a blinded expert. The primary endpoint was the extraction error rates. Secondary endpoints included per-variable performance, inter-institutional variability, F1-score for multiple-choice variables, added value of hybrid AI–human workflows, and survival analyses.

Results:
Among 10,327 patients with AI-based extraction, 311 were included in the test cohort. Across 8,708 datapoints for 28 variables with only one correct answer, the LLM achieved a 7.0% error rate, outperforming manual abstraction (14.2%, *p<0.001*). The F1-scores of 3 multiple-choice variables were superior (gene alterations 0.97 vs 0.86, comorbidities 0.86 vs 0.76, metastatic sites 0.71 vs 0.69). Inter-institutional variance was lower with AI (0.12% vs 0.39%). A hybrid approach with targeted human review of 30% of low-confidence AI outputs further decreased error rates to 4.4%. Survival analyses based on AI-extracted data closely matched ground truth, with similar median overall and progression-free survival.

Conclusions:
In a multicenter setting, our AI pipeline yielded lower error rates and greater consistency than manual abstraction. These findings support the feasibility of next-generation, AI-enabled multicenter studies to generate high-quality RWD at scale, with potential applicability in prospective clinical trials.

Keywords: artificial intelligence, large language models, electronic health records, real-world data, multicenter, consortium

Highlights:

In a 10-center study, AI extraction from unstructured medical records had a lower error rate than manual abstraction (7% vs 14%).

AI consistently outperformed manual abstraction across 31 clinical variables in lung cancer.

Hybrid AI–human workflow further reduced error rates from 7.0% to 4.4%.

AI extraction lowered inter-institutional variability across participating centers.

Survival outcomes from AI-extracted data closely matched expert-adjudicated ground truth.

Introduction

High-quality data are essential for reliable evidence generation in oncology. Clinical trials remain the cornerstone for assessing efficacy and safety under controlled conditions, while real-world data (RWD) provide complementary insights into routine practice, including outcomes in underrepresented populations and characterization of rare tumor subtypes.[1,2] RWD are increasingly used to inform trial design by providing empirical estimates of event rates, eligibility prevalence, and effect sizes, or to construct synthetic control arms when randomization is not feasible.[3–6]

Multicenter studies are essential for generating large, diverse, and representative datasets, whether in the context of RWD or clinical trials. By capturing heterogeneity across diverse care settings and patient populations, these studies improve external validity and generalizability while reducing center-specific biases. They may also provide sufficient sample sizes to study rare cancers, rare molecular subtypes, and less common clinical scenarios. Examples include Flatiron Health (USA[7]), Unicancer ESME (France[8]), the SACT database (UK)[9], or rare tumor registries involving multiple institutions.[10–12]

However, extracting data faces a fundamental challenge as most health records (HR) are stored in unstructured formats.[13] Although these records contain rich clinical information essential for patient care and research, they are written in natural language with heterogeneous terminology, abbreviations, documentation styles, and quality, particularly across institutions. Currently, RWD generation relies predominantly on manual chart abstraction by research personnel, often without formal medical training and without trial-like monitoring. This approach is labor-intensive and susceptible to human error, with reported error rates ranging from 2.3 to 26.9%, due to mistakes both in data entry and interpretation of the original document.[14] Even in trials, despite standardized collection tools and monitoring, error rates can be high enough to require larger sample sizes to preserve statistical power.[15] In both RWD and trials, data errors can reduce the ability to detect true effects, underscoring the need for scalable approaches to improve accuracy and consistency. Beyond accuracy concerns, manual abstraction depends on local staffing and resources, generating inter-institutional variability and creating barriers to participation for smaller institutions, producing selection bias toward academic centers that limits dataset representativeness.[1]

To address these limitations, there is growing interest in automated extraction using natural language processing (NLP) and LLMs, that may offer a scalable and time-efficient alternative.[16–18] Current state-of-the-art LLMs show promising performance in structuring variables from unstructured notes, pathology, and radiology reports, and in distinguishing clinically relevant from uncertain genomic variants.[18–21] This approach may enhance the representativeness of multicenter cohorts, improve data quality, and reduce the time and resource demands of data collection. However, most LLM studies remain limited to narrowly defined variables or to single-

center settings.[16,18,20–22] Moreover, the most performant LLMs are often proprietary and restricted to public repositories[19,23] to comply with data protection policies. Alternatively, deploying locally open-source models with similar performance is complex,[24] requiring specialized, resource-intensive computing infrastructure, which limits its feasibility in a multicenter setting.

We propose the concept of "next-generation multicenter studies," which leverage LLMs for automated extraction and structuring of clinical data from unstructured HRs across institutions, using a centralized cloud-based platform and standardized data model to ensure harmonization. To our knowledge, this is the first study to validate an LLM-based approach (lucc-ai-1) for multicenter RWD generation across a comprehensive set of clinical variables (31 variables), with direct head-to-head comparison against manual abstraction, evaluation of inter-institutional consistency, and assessment of confidence-guided hybrid AI-human workflows.

Methods

The LUCC consortium

This study was conducted within the LUCC consortium, launched in June 2023 to establish a large-scale RWD cohort across multiple cancer types through AI-based structuring of HRs. The initial focus was lung cancer (LUCC-Lung cohort). The consortium includes university hospitals, comprehensive cancer centers, private clinics, and community hospitals across France, ensuring representation of diverse care settings and patient populations. In this consortium, institutions retained scientific control and data ownership, while industry (Lifen) provided the AI expertise, pipeline and a centralized data hub.

Participating centers exported health records in PDF format directly from their clinical systems without specific ordering requirements. No systematic audit of completeness against full EHR contents was performed. Our approach does not require interoperability with institutional EHR systems, making it independent of the underlying clinical information systems used by participating institutions. Health records ranged from completely unstructured free-text narratives to semi-structured formats with section headers (e.g., "Comorbidities", "Molecular profile").

As of July 2025, LUCC-Lung comprised 10,327 adults with histologically confirmed lung cancer, irrespective of stage or treatment, with at least one medical report available since February 2021.

Patient Selection for the validation study

We randomly selected 311 patients from the LUCC Lung cohort, stratified by center. Additional inclusion criteria were: (1) availability of five or more French-language medical records including hospitalization, consultation reports or tumor board summaries, and (2) absence of prior oncological history or concomitant cancers. Patients with concomitant cancers were excluded to reduce clinical and documentation complexity for both manual and AI abstraction and to first validate model performance in standard single-cancer situations. This cohort served as our test set and was kept isolated from the training process to ensure an unbiased evaluation. Document type distributions across the test cohort are detailed in Supplementary Table S4.

Variable Selection

Thirty-one variables relevant to patient characterization and outcome analyses (Supplementary S1) were selected by Gustave Roussy, whose variable definitions and codebook were provided to Lifen, the AI provider. Variables were selected to enable comprehensive patient characterization, treatment pattern analysis, and survival outcome assessment and included demographic data, risk factors, comorbidities, molecular data (molecular status of *EGFR, ALK, ROS1, RET, KRAS, BRAF, HER2, MET* and PD-L1 expression), first-line systemic treatment details, and follow-up information.

Twenty-eight variables had a standard format with only one correct answer (date, number, boolean, or category) including the option to mark information as unavailable ("null" value). Both AI and manual extraction were instructed to indicate unavailable information rather than infer values.

Three variables were collected as multiple-choice fields: genomic alterations (8 genes), comorbidities (11 comorbidities), and metastatic sites at first-line treatment initiation (22 sites) (Supplementary S1). For these variables, each option could be marked as present or absent. For operational reasons and because the distinction was rarely explicit in our material, we did not differentiate between "unavailable" (testing not performed or not documented) and "negative" (tested and negative). Both cases were recorded as negative values.

Comparison Groups

Three groups were evaluated using identical de-identified documents and processing rules (Supplementary S2). All groups accessed the same source documents without direct access to HRs and were supported by identical automatic coherence constraints enforcing variable formats, restricting entries to allowed values, and ensuring compliance with conditional appearance rules (Supplementary S1). For each included patient (with ≥5 eligible medical records), all available medical reports were provided to every group, without any upper limit on the number of documents.

Group 1 (Manual): Each center designated clinical research professionals with experience in clinical data abstraction, to manually abstract data through an electronic case report form (eCRF) platform, with written instructions. Only initial entries were analyzed, without data-management queries or corrections. Abstractors had access to basic document viewing without advanced search tools.
Group 2 (AI Model): Data were extracted automatically using *lucc-ai-1*, which provided predictions with confidence scores and supporting text excerpts.
Group 3 (Hybrid): AI-extracted values underwent targeted human review. The model provided predicted values with associated confidence scores, and two human experts (distinct from Group 1) reviewed predictions in ascending order of AI confidence, up to 30% of extractions. Reviewers accessed an enhanced interface with: (1) AI predictions and confidence scores, (2) supporting excerpts highlighted in source documents, (3) keyword search functionality across all patient documents, and (4) options to accept, modify, or reject predictions (Supplementary Figure S3).

Ground Truth Definition

Values from each group were compared for each data point. A data point was defined as one variable value for one patient**.** When values were identical across groups, the shared value constituted the ground truth. In all cases of discrepancies, a senior clinical research professional under the supervision of medical oncologists (M.A., L.Z.), blinded to the extraction method (manual or AI) and not involved in the manual data abstraction group, adjudicated the correct entry based solely on source documents. The adjudicated value could differ from all initial assessments.

Endpoints

Primary Endpoint

The primary endpoint was the difference in error rates for data abstraction from medical records across institutions, comparing an AI-based approach with manual abstraction by trained clinical research professionals.

For the 28 variables with standard formats and an additional option to mark information as unavailable, performance was evaluated using two support sets: (i) complete support (all ground-truth data points, including null values) and (ii) common non-null support (excluding null values across all groups).

Error rates were defined as the proportion of discrepancies relative to each support set. Any deviation from the ground truth was counted as incorrect. For example, "December 2022" or "December 13, 2022" were incorrect if the correct date was "December 12, 2022", "Stage III" was

incorrect if "Stage IIIA" was required and "PD-L1 expression = 0%" was incorrect if testing was not performed and the ground truth was "null".

Secondary Endpoints

Secondary endpoints included the estimation of per-variable performance metrics and assessment of inter-institutional variability; computation of F1-scores for multiple-choice variables for both AI-based extraction and manual abstraction; quantification of the added value of hybrid AI–human validation workflows; and comparison of survival curves across all three approaches.

The three variables originally collected as multiple-choice fields with no differentiation between "unavailable" and "absent", were transformed into sets of binary variables, one for each possible value. This transformation enabled the calculation of micro F1-score for each option, which is clinically relevant since partial identification (e.g., correctly extracting some molecular alterations while missing others) still provides useful information. For example, the "genes" variable was decomposed into eight binary variables, each corresponding to the presence or absence of a specific gene alteration.

Processing duration was measured and compared across three groups.

To assess the clinical relevance of extraction accuracy, overall survival (OS) and real-world progression-free survival (PFS) curves were generated from AI and manual abstraction versus the ground truth. Survival was calculated from diagnosis (OS, overall population) or start of first-line treatment (OS and PFS, treated subgroup) to the event: death or last follow-up for OS, and progression, death, or the last imaging report without documented progression for PFS.

Sample size calculations were based on a non-inferiority design comparing AI accuracy to manual data abstraction. Assuming a human baseline accuracy of 95% and the same accuracy by AI under the alternative hypothesis, a non-inferiority margin of 10%, a one-sided alpha level of 0.025, 300 patients would provide 98% power for a patient-level data point. Statistical analyses were performed under direct academic oversight (S.M.).

Ethical Considerations

All living participants were individually informed through an information note developed in collaboration with a patient association, written in clear and accessible language. The note outlined the study objectives, data processing procedures, and participants' right to refuse the use of their data. All study data were de-identified, with access restricted to authorized personnel; every access event was logged and traceable. Information was encrypted during both

transmission and storage, and hosted by Lifen, a provider certified to store and process health data in compliance with the French law. The study complies with the General Data Protection Regulation,[32] the French Data Protection Act,[33] and the French Public Health Code,[34] and was approved by the LUCC Scientific and Ethics Committee (Supplementary S5).

AI Methodology

We developed *lucc-ai-1*, an AI pipeline to transform unstructured clinical documents into structured data by combining preprocessing, fine-tuned domain-specific LLMs, and patient-level reconciliation algorithms (Fig. 1). Several open-source LLMs were evaluated, including Mistral-7B, Mistral-24B, Llama-3.1-8B, Granite-3.0 8B, Ministral 8B, Hermes-3 8B, Falcon-3 7B, MedGemma-27B. On the training set, Mistral-24B showed the best balance of performance and computational cost (Supplementary S6-1). Fine-tuning was performed on a single NVIDIA H100 GPU (80 GB).

HR preprocessing: Participating centers provided unstructured HR data in PDF format. The initial step was extraction of medical text, using a pipeline that combined open-source and proprietary tools for optical character recognition, template removal, reading order detection, and de-identification of personally identifiable information. This process produced a consistent, de-identified text corpus for downstream analysis.

Document-level predictions: the number of documents per patient frequently exceeded the input limits of contemporary open-source LLMs, even after preprocessing. To address this, we fine-tuned an LLM to operate at the document level, predicting values for each individual document.

We compared two training paradigms: (1) Single Document: for each variable, only the annotated document, specification, and target excerpt were supplied; (2) All Documents: up to five records were supplied irrespective of containing the target, with specification and target excerpt. The first approach suited simple variables answerable from a single record, the second addressed cross-document reasoning (e.g., treatments) but was computationally heavier.

Prediction confidence was quantified using token-level probability distributions.

Iterative human-in-the-loop training: Model training proceeded iteratively to minimize annotation errors and reduce manual effort. We began with a task-untrained base model producing predictions across the corpus. Clinical research professionals corrected a subset to create an initial labeled set. The updated model was reapplied, generating predictions for review, prioritized by low confidence. This train–predict–review loop progressively refined labels and improved model performance (Supplementary Fig. S6-2).

As of July 2025, the development corpus comprised 10,016 patient records split into a training set of 9,316 patients and an internal validation set of 700 patients used for model selection. Critically, the 311 patients included in the present study constituted a separate, held-out test set that remained fully blinded throughout model development.

Patient-level reconciliations: To derive a single prediction per patient and variable, we aggregated document-level outputs. Invalid predictions, including those with unrecognized values, incorrect supporting context, or low model confidence, were systematically excluded. For most variables, we used one of two heuristics: (1) selecting the most frequent value, weighted by confidence, or (2) choosing the value from the most recent document. We also designed a few customized reconciliation strategies for four specific variables, based on their unique characteristics. This process yielded a final patient-level prediction and corresponding confidence which could directly be compared with human annotations.

Results

Patient cohort and data characteristics

Performance was evaluated in 311 patients across 10 participating institutions representing diverse care settings (university hospitals, comprehensive cancer centers, private clinics, and community hospitals), with institutional contributions ranging from 9 to 73 patients (Supplementary S7). A total of 5505 source documents were used (Supplementary S4). Manual abstraction was performed by clinical research professionals with experience in clinical data abstraction (median 5 years [range 1-17]; median 4 years [range 0-8] for lung cancer specifically) (Supplementary S8)

Data structuring was performed on 9,641 data points across 31 variables, including 28 variables with standard formats and an additional option to mark information as unavailable ("null") and 3 multiple-choice variables allowing partial identification. Patients' characteristics (per ground truth) are depicted in Supplementary S9. A median of 12 documents (IQR 8–20) per patient were processed by the AI pipeline. Manual chart abstraction required a median of 17.5 minutes per patient (IQR 9.7 - 32.2). AI processing time was negligible once the model was trained, requiring 9 hours of compute time when distributed across three H100 GPU machines to process 311 patients, corresponding to approximately 1.7 minutes per patient.

Overall patient-level data extraction performance

For the full support analysis on the 28 variables with the "null" option, 8,708 patient-level data points were verified against the ground truth. Manual abstraction yielded an error rate of 14.2%, with 1,240 incorrect data points, while AI extraction achieved a lower error rate of 7.0%, with 611 incorrect data points (*p<0.001*). In the common non-null support, 5,512 data points were evaluated. Manual abstraction showed a 22.5% error rate (1,240 errors), whereas AI extraction demonstrated an 11.1% error rate (611 errors) (*p<0.001*).

Results are further detailed in Fig. 2, using the following classification: true positive (both set and ground truth were identical and non-null): 52.1% for AI vs. 47.7% for manual in the complete support; 82.4% vs. 75.3% in the non-null support; true negative (both were identical and null): 40.8% vs. 38.1% in the complete support; 6.5% vs. 2.2% in the non-null support; false positive (the set was non-null and the ground truth was null):1.5% vs. 4.2% in the complete support; 2.4% vs. 6.7% in the non-null support; false negative (the set was null and the ground truth was non-null):2.7% vs. 5.78% in the complete support; 4.3% vs. 9.01% in the non-null support; and discrepant value (both were non-null but differed): 2.8% vs. 4.3% in the complete support; 4.4% vs. 6.8% in the non-null support.

Inter-institution variability

Error rates differed markedly by center for manual abstraction (range 11.5–23.4%), reflecting variability in expertise and local practices. AI extraction error rates were consistently lower across all centers, ranging from 5.8% to 13.2% (Fig. 3A). The weighted sample variance in error rates across institutions decreased from 0.39% for manual abstraction to 0.12% for AI extraction. Manual abstraction time correlated positively with error rate (*r =0.662*), suggesting longer review did not improve accuracy (Fig. 3B).

Variable-specific performance

For all 31 variables, AI extraction demonstrated lower error rates than with manual abstraction (Fig. 4 A-B). For birthdate, gender, and vital status, error rates were below 3% for both AI and manual abstraction, with AI achieving 0% for birthdate and gender. Histological type, M stage, and PD-L1 expression had <5% error with AI but >5% with manual (Fig. 4).

Variables with error rates below 10% in AI extraction included date of last contact, metastatic status at diagnosis, TNM stage, N stage, T stage, smoking status, number of pack-years, and presence of first-line systemic treatment. Manual abstraction, by contrast, exceeded 10% error for all of these except N stage.

Five variables had error rates above 20% in AI extraction: performance status at first-line start, first-line end date, first-line progression date, date of last progression-free assessment, and date

of metastatic progression diagnosis. In comparison, manual abstraction showed error rates above 20% for 15 variables.

Variables showing the largest absolute difference in error rates between AI and manual abstraction (>20%) included metastatic progression after diagnosis, name of first-line chemotherapy drug, TNM stage at diagnosis, date of diagnosis, performance status at first-line start, and first-line treatment category.

The micro-F1 score for genomic alterations was 0.96 for AI versus 0.86 for manual; for comorbidities, 0.86 versus 0.75; and for metastatic sites, 0.70 versus 0.68, respectively (Fig. 4B).

Hybrid AI–human workflow

The distribution of AI prediction confidence is shown in Fig. 5A. Thirty percent of predictions had confidence scores below 83%, 20% were below 71%, and 10% were below 50%. A targeted hybrid workflow, in which a human reviewer assessed 30% of the lowest-confidence AI predictions, reduced the error rate from 7.0% to 4.4% across the complete support and from 11.2% to 7% across the common non-null support (Fig. 5B). The hybrid approach reduced the error rate to below 5% for clinical variables including number of pack-years, smoking status at diagnosis, TNM stage, metastatic status at diagnosis, N stage, and T stage (Fig. 5C). Additionally, variables with more than a twofold reduction in error included all of the above, as well as last progression-free assessment date, performance status at first-line treatment initiation, and first-line treatment end date.

This hybrid approach required a median of 4.6 minutes per patient (IQR 3.6–6.1) for human review and validation, including AI processing time, representing a 73.7% decrease compared with manual abstraction.

Identified factors impacting the error rate

Error propagation in dependent variables: A total of 14 variables were conditionally dependent on others. For example, progression date is contingent on the accurate identification of progression status, and the first-line (L1) treatment category depends on detecting the presence of an L1 treatment. In these cases, errors in parent variables can propagate to dependents. As the number of dependencies increases, cumulative error rates tend to rise, a pattern observed in both AI and manual abstraction. For variables with no dependencies (14/28), error rates were 4.9% (AI) vs. 14.4% (manual); with one dependency (7/28), 16.3% vs. 34.5%; with two (7/28), 32.7% vs. 47.1%.

Errors in date-format variables: Eight variables used date formats, where only exact matches were accepted. This strict criterion resulted in elevated error rates: 19.7% (range 0–50%) for AI and 30.8% (0.9–63.1%) for manual, particularly from minor day-level discrepancies. For discrepant values, we observed median differences of 63 days [IQR : 25-206 days] for AI method (96 discrepant values) and 62 days [IQR : 24-170 days] for manual method (145 discrepant values)(Supplementary table S10).

With a more lenient approach (±15 days and ±30 days), error rates decreased to 17.1% then to 15.9% for AI and to 25.4% then to 23.6% for manual.

Error in identifying systemic treatment: Manual abstractors misclassified systemic treatment initiated for early-stage disease as first-line metastatic therapy in 7 cases. When restricting the analysis to patients with stage IV disease at diagnosis (N=148), error rates were 17.1% with manual abstraction, 12.1% with AI extraction and 6.7% with the hybrid method. For the drug name errors, one clinical research professional did not identify the commercial name of pemetrexed as corresponding to pemetrexed and therefore did not select pemetrexed as a treatment. AI extraction yielded lower error rates for most variables, except for two: the last date of received treatment and the date of the last imaging showing absence of progression (Supplementary Figure S11). The hybrid method further reduced error rates across all variables, with the greatest improvement observed for these two dates (Supplementary Figure S11).

Impact on relevant biomarkers and survival analyses

Genomic biomarkers relevant to lung cancer identified by AI closely matched the ground truth. Compared to manual abstraction, AI produced fewer false negatives (5 vs. 13) and false positives (2 vs. 16) (Fig. 6A, confusion matrices shown in Supplementary S12).

When applied to survival end points, the AI-extracted data produced Kaplan–Meier curves that closely matched the ground truth (Fig. 6B-D). For patients receiving first-line chemoimmunotherapy in the metastatic setting, the median OS was 20.1 (95% CI: 14.7 - non-reached (NR)) months for the ground truth, 20.1 (95% CI: 16.9 - NR) months for AI extraction, and 23 (95% CI: 19.7 - NR) months for manual abstraction (Fig. 6C). The PFS was 8.1 (95% CI 6 - 16.9) months for the ground truth, 8.1 (95% CI 5.5 - 18.5) months for AI extraction, and 10.1 (95% CI 8 - 13.8) months for manual abstraction (Fig. 6D).

Discussion

Here, we report the validation study of an LLM pipeline, *lucc-ai-1*, automated end-to-end, from patient identification and notification to de-identification, preprocessing, and AI-based extraction of lung cancer RWD, on a centralized platform supporting multicenter participation without local resources. As data capture was PDF-based, our model operated without interoperability issues, making it agile and scalable. Moreover, we used fine-tuned open-source models, allowing compliant processing of real patient data. The comparison of AI data extraction with manual abstraction by clinical research professionals across ten French institutions showed a lower overall error rate (7.0% vs. 14.2%) across all variables and centers, consistent with other studies reporting high LLM accuracy in data extraction and summarization, sometimes exceeding medical experts.[18–22,25,26] It also confirms that manual chart review is labor-intensive and error-prone, similar to other studies with reported error rates for manual abstraction.[15,27,28] Recent work by Jee et al. used BERT-based classifiers trained separately for each variable, but at the cost of flexibility, since each new variable requires a separate model. Their single-center curated dataset also limited robustness to heterogeneous document formats and did not address patient-level aggregation, making direct comparison with our approach challenging.[17]

The risk of hallucinations is recognized as a critical limitation of LLMs when processing long unstructured HRs.[29] While the term “hallucination” is commonly used to describe false outputs from generative LLMs, our setting involves structured prediction with predefined value sets, allowing clearer classification of outputs as correct or incorrect. To mitigate pitfalls, we required predictions to be supported by excerpts and iteratively corrected annotation errors. Consequently, AI extraction showed lower false negatives (2.7% vs. 5.7%) and false positives (1.5% vs. 4.2%) than manual abstraction.

Our review of discrepancies between manual annotation, AI output, and the ground truth revealed several systematic differences in data collection behavior. One notable source of error for the manual abstraction could be the “first reported information” bias. When reviewing patient files sequentially, clinical research professionals often complete eCRFs as soon as they encounter relevant information. However, they may not revisit a variable if more precise or conflicting information appears in later documents (e.g., mistaking surgery for diagnosis date). In contrast, AI processed all documents, reducing this bias. Another discrepancy was strict rule compliance: the AI was explicitly trained to follow all rules (Supplementary S2), with any deviation counted as incorrect.

The hybrid review experiment, where 30% of low-confidence AI predictions were reviewed by a human, showed that error rates decreased further from 7.0% to 4.4%, indicating that limited human oversight can enhance data quality without a full re-abstraction. In RWD studies, re-abstraction is rarely performed while in clinical trials data quality is ensured through electronic

data capture systems combined with source data verification, often applied selectively within risk-based monitoring frameworks.[30] Our results suggest that AI-guided review could complement these approaches by focusing oversight on the most error-prone data points. Unlike random review, where gains scale with volume, targeting the lowest 10% confidence predictions yielded the largest error reduction, with diminishing gains thereafter. This confidence-guided strategy may thus offer more efficient review allocation, though further refinement is needed to optimize its implementation.

Concerning the multicenter component, this is the first demonstration of AI-based automatic structuring across the largest multicenter consortium to date, involving 10 institutions of varying case volumes, EHR (Electronic Health Records) systems, and practice settings. *lucc-ai-1* outperformed manual abstraction across all centers and reduced inter-institutional variability, despite heterogeneous data sources and practices. While proactive monitoring is standard in clinical trials, it is usually absent in RWD studies. Proposed alternatives to reduce errors have included early data downloads and analyses of deviations, written protocol clarifications, and refresher training sessions.[31] Nonetheless, inter-institutional variability persisted in this study despite standardized written instructions for manual abstraction. Our findings suggest that AI-based extraction can improve data homogeneity by reducing reliance on site-specific human input, inherently influenced by training, experience, and workflows.

Limitations of the study include acceptance of concordant values without adjudication, which may underestimate absolute error rates. We consider this impact limited, as concordance was established across three complementary methods applied to the same source documents, each with modest error rates; further adjudication would add another comparison layer with diminishing returns. Additionally, only initial manual entries were analyzed without data-management queries or corrections, which may overestimate manual error rates compared with a fully monitored trial setting but reflects typical multicenter RWD practice. Another major constraint is the need for high-quality labeled data; this study was feasible because thousands of annotated examples per variable were created via manual review of AI predictions. While we are developing transfer learning strategies for zero-shot extraction, these capabilities are not yet implemented in our current framework. Performance varies with variable complexity and may not generalize to variables with limited or no training data. This poses challenges when introducing new variables or modifying extraction rules, each requiring annotation effort (Supplementary S6-2). Also, extracting longitudinal variables (e.g., treatments, progression events) proved challenging. Our current document-level approach lacks the capacity to reconcile multiple valid instances of the same variable across time. As a result, we limited extraction to first-line treatments. Although we explored incorporating temporal information, the required reconciliation heuristics were too complex for the current implementation. Finally, the study was

limited to French-language records from patients with lung cancer, and generalizability to other languages, cancer types, or healthcare settings remains untested.

Future work will focus on extracting time-dependent variables, defining more atomic document-level data elements that can be aggregated deterministically into patient-level variables, and improving confidence scores for human-in-the-loop validation. Expanding beyond the French language to support additional languages is important for broader international applicability. Although LLMs offer multilingual capabilities, their performance must be validated in diverse linguistic contexts. Similarly, applying this approach beyond lung cancer to other malignancies and disease areas warrants further investigation. Beyond retrospective RWD extraction, *lucc-ai-1* is undergoing prospective validation in a pragmatic randomized phase III trial (PULSE, NCT05692999), to test suitability for trial use.

Conclusion

In conclusion, our findings validate *lucc-ai-1*, a centrally hosted LLM-based framework, for automated extraction of lung cancer research variables from unstructured medical narratives. Compared to manual abstraction, the model achieved lower error rates and greater consistency across centers. By reducing workload and standardizing data capture, this approach enables a "next-generation" multicenter study model that is scalable, less dependent on local research infrastructure, and more inclusive of smaller hospitals, thereby improving the representativeness of real-world oncology data.

Acknowledgements: English editing was performed with ChatGPT-4o and -5. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication

Funding:
This work was supported by Lifen. No grant number is applicable.

Conflicts of interest:

Mihaela Aldea – research funding (institution) from Amgen, AstraZeneca, Sandoz, Owkin.
Lodovica Zullo - no conflicts of interest.
Virginie Levrat - no conflicts of interest.
Jaafar Bennouna - reports serving on advisory councils or committees for AstraZeneca, Daiichi Sankyo, MSD, Johnson & Johnson; and honoraria from AstraZeneca, Bayer, Bristol Myers

Squibb, Merck, MSD, Daiichi, Servier, Ipsen, Pierre Fabre Oncologie, Regeneron, Johnson & Johnson.

Sophie Schneider - no conflicts of interest.

Olaf Mercier - no conflicts of interest.

Emmanuelle Mougenot - no conflicts of interest.

Emmanuel Bergot - no conflicts of interest.

Cécile Dujon - no conflicts of interest.

Nicolas Cloarec - no conflicts of interest.

Clarisse Audigier Valette - no conflicts of interest.

Antonio Nuccio - no conflicts of interest.

Marc Deloger - no conflicts of interest.

Solenne Simon - no conflicts of interest.

Carole Helissey - reports receiving honoraria from Janssen, Astellas Pharma, Sanofi, AstraZeneca, Bayer, Ipsen, and Roche; and receiving travel, accommodations, or expenses from Janssen, outside of the current work.

Alexandre Carpentier - Employee of Lifen.

Azeddine Djarallah - Employee of Lifen.

Pierre Rolland - Employee of Lifen.

Jean-Charles Louis - Employee of Lifen.

Lou Ancillon - Employee of Lifen.

Benjamin Vignal - Employee of Lifen.

Florent Rambaud - Employee of Lifen.

Pierre Tessier - Employee of Lifen.

Lisa Chuttoo - Employee of Lifen.

Khadija Siby - Employee of Lifen.

Aliette Poplu - Employee of Lifen.

Kevin Zarca - Employee of Lifen.

Stefan Michiels - reports outside the scope of the submitted work : fees for Scientific Committee Study member : Roche, for Data and safety monitoring member of clinical trials: IQVIA, Kedrion, Servier, Yuhan

Fabrice Barlesi - reports Consulting or Advisory Role: Roche/Genentech (Inst), Novartis (Inst), Bristol Myers Squibb (Inst), AstraZeneca/MedImmune (Inst), Boehringer Ingelheim (Inst), Lilly (Inst), Merck Serono (Inst), MSD Oncology (Inst), Takeda (Inst), Bayer (Inst), Amgen (Inst), Eisai Europe (Inst), Sanofi (Inst), Mirati Therapeutics (Inst). Research Funding: Roche/Genentech (Inst), AstraZeneca/MedImmune (Inst), Bristol Myers Squibb (Inst), Pierre Fabre (Inst), AbbVie (Inst), Amgen (Inst), Bayer (Inst), Boehringer Ingelheim (Inst), Eisai (Inst), Lilly (Inst), Ipsen (Inst), Innate Pharma (Inst), Novartis (Inst), Merck Serono (Inst), MSD Oncology (Inst), Pfizer (Inst), Sanofi/Aventis (Inst), Takeda (Inst). Travel, Accommodations, Expenses: Roche/Genentech.

Franck Le Ouay - CEO at Lifen.

Benjamin Besse – advisory board honoraria (institution) from AbbVie, BioNTech SE, Bristol Myers Squibb, Chugai Pharmaceutical, CureVac AG, Daiichi Sankyo, F. Hoffmann-La Roche Ltd, PharmaMar, Regeneron, Sanofi-Aventis, Turning Point Therapeutics; consulting (institution) for AbbVie, Eli Lilly, Ellipses Pharma Ltd, F. Hoffmann-La Roche Ltd, Genmab, Immunocore, J&J, MSD, OSE Immunotherapeutics, Owkin, Taiho Oncology; steering committee participation (institution) for AstraZeneca, BeiGene, Genmab A/S, GlaxoSmithKline, J&J, MSD, OSE Immunotherapeutics, PharmaMar, Roche-Genentech, Sanofi, and Takeda; speaker fees (institution) from AbbVie, AstraZeneca, Chugai Pharmaceutical, Daiichi Sankyo, Hedera Dx, J&J, MSD, Roche, Sanofi-Aventis, and Springer Healthcare Ltd

References

1. Maio D, Perrone M, Conte F. Real-World Evidence in Oncology: Opportunities and Limitations. *Oncologist*. 2020;25:e746–e752.

2. Mavroeidis L, Napolitano A, Huang P, Jones RL. Real-world evidence for ultra rare cancers. *Rare Tumors*. February 15, 2024;16:20363613241234207.

3. Baumfeld Andre E, Reynolds R, Caubel P, Azoulay L, Dreyer NA. Trial designs using real-world data: The changing landscape of the regulatory approval process. *Pharmacoepidemiol Drug Saf*. October 2020;29(10):1201–1212.

4. Dagenais S, Russo L, Madsen A, Webster J, Becnel L. Use of real-world evidence to drive drug development strategy and inform clinical trial design. *Clin Pharmacol Ther*. January 2022;111(1):77–89.

5. Yap TA, Jacobs I, Baumfeld Andre E, Lee LJ, Beaupre D, Azoulay L. Application of real-world data to external control groups in oncology clinical trial drug development. *Front Oncol*. 2021;11:695936.

6. Shortreed SM, Rutter CM, Cook AJ, Simon GE. Improving pragmatic clinical trial design using real-world data. *Clin Trials*. June 2019;16(3):273–282.

7. Evidence Solutions [Internet]. [cited August 7, 2025]. Available at: https://flatiron.com/real-world-evidence

8. ESME [Internet]. Unicancer. 2021 [cited August 7, 2025]. Available at: https://www.unicancer.fr/en/programs/esme/

9. Bright CJ, Lawton S, Benson S, Bomb M, Dodwell D, Henson KE, et al. Data resource profile: The systemic anti-cancer therapy (SACT) dataset. *Int J Epidemiol*. February 1, 2020;49(1):15–15l.

10. Aldea M, Marinello A, Duruisseaux M, Zrafi W, Conci N, Massa G, et al. RET-MAP: An international multicenter study on clinicobiologic features and treatment response in patients with lung cancer harboring a RET fusion. *J Thorac Oncol*. May 2023;18(5):576–586.

11. Drilon A, Duruisseaux M, Han J-Y, Ito M, Falcon C, Yang S-R, et al. Clinicopathologic features and response to therapy of NRG1 fusion-driven lung cancers: The eNRGy1 global multicenter registry. *J Clin Oncol*. September 1, 2021;39(25):2791–2802.

12. Lange S, Bleckmann A, Kasenda B. 122P Molecular testing, treatment, and response of patients with advanced solid tumors harboring an NTRK gene fusion: Second interim results of the REALTRK registry. *ESMO Open*. 2023;8.

13. Sedlakova J, Daniore P, Horn Wintsch A, Wolf M, Stanikic M, Haag C, et al. Challenges and best practices for digital unstructured data enrichment in health research: A systematic narrative review. *PLOS Digit Health*. October 2023;2(10):e0000347.

14. Goldberg SI, Niemierko A, Turchin A. Analysis of data errors in clinical research databases. *AMIA Annu Symp Proc*. November 6, 2008;242–246.

15. Garza MY, Williams T, Ounpraseuth S, Hu Z, Lee J, Snowden J, et al. Error rates of data processing methods in clinical research: A systematic review and meta-analysis of manuscripts identified through PubMed. *Int J Med Inform*. March 2025;195(105749):105749.

16. Gauthier M-P, Law JH, Le LW, Li JJN, Zahir S, Nirmalakumar S, et al. Automating access to real-world evidence. *JTO Clin Res Rep*. June 2022;3(6):100340.

17. Jee J, Fong C, Pichotta K, Tran TN, Luthra A, Waters M, et al. Automated real-world data integration improves cancer outcome prediction. *Nature*. December 2024;636(8043):728–736.

18. Sun VH, Heemelaar JC, Hadzic I, Raghu VK, Wu C-Y, Zubiri L, et al. Enhancing precision in detecting severe immune-related adverse events: Comparative analysis of large language models and International Classification of Disease codes in patient records. *J Clin Oncol*. December 10, 2024;42(35):4134–4144.

19. Lin K-H, Kao T-H, Wang L-C, Kuo C-T, Chen PC-H, Chu Y-C, et al. Benchmarking large language models GPT-4o, llama 3.1, and qwen 2.5 for cancer genetic variant classification. *NPJ Precis Oncol*. May 15, 2025;9(1):141.

20. Rajaganapathy S, Chowdhury S, Buchner V, He Z, Jiang X, Yang P, et al. Synoptic reporting by summarizing cancer pathology reports using Large Language Models. *medRxiv*. May 9, 2024; Available at: http://dx.doi.org/10.1101/2024.04.26.24306452

21. Grothey B, Odenkirchen J, Brkic A, Schömig-Markiefka B, Quaas A, Büttner R, et al. Comprehensive testing of large language models for extraction of structured data in pathology. *Commun Med (Lond)*. March 31, 2025;5(1):96.

22. Shahid F, Hsu M-H, Chang Y-C, Jian W-S. Using generative AI to extract structured information from free text pathology reports. *J Med Syst*. March 13, 2025;49(1):36.

23. Freyer O, Wiest IC, Kather JN, Gilbert S. A future role for health applications of large language models depends on regulators enforcing safety standards. *Lancet Digit Health*. September 2024;6(9):e662–e672.

24. Umeton R, Kwok A, Maurya R. GPT-4 in a Cancer Center - Institute-Wide Deployment Challenges and Lessons Learned. *NEJM AI*. 2024;1.

25. Van Veen D, Van Uden C, Blankemeier L, Delbrouck J-B, Aali A, Bluethgen C, et al. Adapted

large language models can outperform medical experts in clinical text summarization. *Nat Med*. April 2024;30(4):1134–1142.

26. Asgari E, Montaña-Brown N, Dubois M, Khalil S, Balloch J, Yeung JA, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *NPJ Digit Med*. May 13, 2025;8(1):274.

27. Feng JE, Anoushiravani AA, Tesoriero PJ, Ani L, Meftah M, Schwarzkopf R, et al. Transcription error rates in retrospective chart reviews. *Orthopedics*. September 1, 2020;43(5):e404–e408.

28. Garza MY, Williams TB, Ounpraseuth S, Hu Z, Lee J, Snowden J, et al. Comparing Medical Record Abstraction (MRA) error rates in an observational study to pooled rates identified in the data quality literature. *BMC Med Res Methodol*. December 18, 2024;24(1):304.

29. Adams L, Busch F, Han T, Excoffier J-B, Ortala M, Löser A, et al. LongHealth: A question answering benchmark with long clinical documents. *J Healthc Inform Res*. September 2025;9(3):280–296.

30. Adams A, Adelfio A, Barnes B, Berlien R, Branco D, Coogan A, et al. Risk-based monitoring in clinical trials: 2021 update. *Ther Innov Regul Sci*. May 2023;57(3):529–537.

31. Guthrie LB, Oken E, Sterne JAC, Gillman MW, Patel R, Vilchuck K, et al. Ongoing monitoring of data clustering in multicenter studies. *BMC Med Res Methodol*. March 13, 2012;12(1):29.

Figure Legends:

Fig. 1. Overview of the AI pipeline, *lucc-ai-1*, and workflow for real-world data structuring. Schematic representation of the end-to-end pipeline lucc-ai-1 used for automated data extraction. Multicenter clinical documents were pre-processed and annotated using a structured data dictionary. A large language model (LLM) was fine-tuned to extract patient-level values from unstructured text. Predictions were reconciled across documents, and outputs were compared against manually collected values by clinical research professionals. Discrepancies were adjudicated to establish a ground truth reference set.

Fig. 2. Overall performance of AI and manual extraction compared to the ground truth. **A.** Breakdown for each dataset vs ground truth with full support; **B.** Breakdown for each dataset vs ground truth with common non-null support.

Fig.3. A. Inter-institution variability in error rates for AI vs manual abstraction ordered by increasing AI error rate in the common non-null support. **B.** Average document volume and manual abstraction time per patient across institutions.

Fig. 4. A. Variable-specific error rates for AI and manual abstraction in the common non-null support, with confidence intervals. **B.** Micro F1-scores for categorical variables extracted by AI vs manual abstraction, with confidence intervals.

Fig. 5. A. Distribution of predictions by confidence level, for which 30% of predictions were manually reviewed from the low-confidence prediction values; **B.** Error rates of hybrid workflow according to percentage of data points manually reviewed on the full and common non-null support (28 variables). **C.** Error rate of AI and manual compared to the hybrid approach per variable.

Fig. 6. A. Frequency of genomic alterations derived from AI, manual and ground truth datasets. **B.** Overall survival of the entire population derived from AI, human, and ground truth datasets. **C.** Overall survival derived from AI, human, and ground truth datasets, for patients with chemoimmunotherapy as first-line treatment**. D.** Progression-free survival derived from AI, human, and ground truth datasets, for patients with chemoimmunotherapy as first-line treatment. mo: months