EJMI

DOI: 10.14744/ejmi.266150
EJMI 2026;10(2):150–158

Research Article

Evaluation of AI-Derived Survival Estimates Against Real-World Outcomes in Extensive-Stage Small Cell Lung Cancer

Fatih Atalah, Mahmut Gümüş

Division of Medical Oncology, Department of Internal Medicine, Istanbul Medeniyet University, Istanbul, Türkiye

Abstract

Objectives: Artificial intelligence (AI) is increasingly explored as a prognostic tool in oncology; however, its reliability in real-world settings remains uncertain. We aimed to compare AI-derived survival estimates with observed outcomes in patients with extensive-stage small cell lung cancer (SCLC).

Methods: This retrospective cohort study included 146 patients treated between 2016 and 2025. Baseline demographic, clinical, metastatic, treatment, and laboratory data were entered into a large language model–based system to generate predicted overall survival (OS) and progression-free survival (PFS). Predicted values were compared with observed outcomes. Agreement was assessed using Bland–Altman analysis.

Results: The median observed OS was 10.97 months, whereas the AI-predicted median OS was 13.2 months, indicating an overestimation of 2.2 months. For PFS, the observed and predicted medians were 6.53 and 7.2 months, respectively, corresponding to an overestimation of 0.7 months. Bland–Altman analyses demonstrated systematic positive bias and wide limits of agreement, indicating substantial variability at the individual level, more pronounced for OS.

Conclusion: AI-derived survival estimates based on baseline data tend to overestimate survival and show limited agreement with real-world outcomes in extensive-stage SCLC. These tools may provide general prognostic insight but are not yet suitable for independent clinical decision-making.

Keywords: Small cell lung cancer, artificial intelligence, survival prediction, prognosis, real-world data, Bland–Altman analysis

Cite This Article: *Elpen Kodaz Ç. Hantavirus Infections: A Review of Pathogenesis, Diagnosis, Treatment, and Vaccine Developments. EJMI 2026;10(2):150–158.*

Artificial intelligence has gradually moved from a largely experimental concept to a tool that is increasingly being tested in everyday clinical practice.[1,2] Initial applications focused on image analysis and pattern recognition, but more recent work has expanded into clinical decision support, outcome prediction, and the interpretation of complex medical data.[3,4] Oncology has naturally become one of the leading fields in this transition, given the volume and diversity of information generated during routine care, including clinical variables, imaging, pathology, laboratory findings, and molecular data.[5,6]

Over the past few years, there has been growing interest in whether AI-based systems can support prognostic assessment in cancer.[2,7] Various approaches have been explored, ranging from machine-learning models trained on structured datasets to language-based systems capable of gen-

Address for correspondence: Fatih Atalah, Division of Medical Oncology, Department of Internal Medicine, Istanbul Medeniyet University, Istanbul, Türkiye
Phone: +90 506 350 90 61 **E-mail:** mgumus@superonline.com

Submitted Date: March 25, 2026 **Revision Date:** June 24, 2026 **Accepted Date:** June 30, 2026

erating predictions from clinical narratives or baseline case summaries.[8] In some settings, these tools have shown the ability to distinguish higher-risk from lower-risk patients.[3,9] However, their role in estimating actual survival durations in real-world clinical settings remains less clear.[10,11] Recent work comparing AI-generated survival estimates with observed outcomes has suggested that these systems may not always provide well-calibrated predictions, despite producing outputs that appear clinically plausible.[12]

This distinction is clinically relevant. In daily practice, oncologists are not only interested in whether a patient belongs to a higher- or lower-risk group, but also in how long that patient is realistically expected to live under current treatment conditions. Many existing models are primarily evaluated using discrimination metrics, which reflect ranking performance rather than absolute agreement.[13] As a result, a model may appear statistically robust while still providing survival estimates that are difficult to apply at the bedside.[14] This issue may be even more pronounced for general-purpose AI systems that are not specifically developed or recalibrated for a given tumor type.[2]

Small cell lung cancer represents a setting in which these limitations become particularly important. It remains one of the most aggressive malignancies in medical oncology, characterized by rapid progression, early dissemination, and a high likelihood of relapse.[15] Although the addition of immune checkpoint inhibitors to platinum–etoposide has modestly improved outcomes, survival in extensive-stage disease remains limited.[16,17] In such a context, even relatively small discrepancies between predicted and actual survival may have meaningful clinical implications.

At the same time, our understanding of SCLC biology is evolving. The disease is increasingly recognized as a heterogeneous entity with distinct molecular and transcriptional subtypes, and this classification is beginning to influence translational research.[18] While subtype-driven treatment is not yet part of routine practice, there is growing interest in more individualized strategies, including biomarker-informed use of immunotherapy, antibody–drug conjugates, and other targeted approaches.[19,20] As treatment options expand, identifying patients who are likely to have particularly poor outcomes at an early stage may become increasingly relevant for clinical decision-making.

From a practical standpoint, prognostic estimation in extensive-stage SCLC is not only an academic exercise. It directly influences discussions with patients, expectations regarding treatment benefit, and decisions related to treatment intensity, sequencing, and supportive care.[15,21] In this setting, tools that can provide reliable survival estimates at diagnosis could have meaningful clinical value.[10] At the same time, it is equally important to understand the limitations of such tools before they are incorporated into routine practice.[22]

In this study, we aimed to compare AI-predicted survival estimates with observed overall survival and progression-free survival in patients with extensive-stage small cell lung cancer treated in a real-world setting. Rather than focusing solely on model-style performance metrics, we sought to evaluate the level of agreement between predicted and actual outcomes, and to explore whether baseline AI-derived estimates can be considered clinically reliable in this patient population.[12]

Methods

This retrospective cohort study included 146 consecutive patients diagnosed with extensive-stage small cell lung cancer (SCLC) and treated in routine clinical practice between January 2016 and June 2025. Clinical data were obtained through retrospective review of institutional patient files and electronic medical records. Eligible patients had histologically confirmed SCLC with extensive-stage disease at the time of diagnosis and sufficient baseline and follow-up data for survival assessment. Patients with missing key clinical variables or incomplete survival data were excluded. All patients were managed according to standard institutional practice.

Baseline demographic and clinical characteristics were recorded at the time of diagnosis. These included age, sex, and Eastern Cooperative Oncology Group (ECOG) performance status, which was considered a key clinical variable in baseline prognostic assessment. Disease-related variables included the presence and sites of metastases, including liver, brain, bone, and other organ involvement. For patients presenting with brain metastases at diagnosis, information regarding the administration of cranial radiotherapy was also recorded. Baseline laboratory parameters obtained prior to treatment initiation were collected from medical records and included routinely used hematological and biochemical variables.

Treatment-related data were documented for all patients. First-line treatment was categorized as either platinum–etoposide chemotherapy alone or platinum–etoposide combined with immune checkpoint inhibitors. In patients receiving chemoimmunotherapy, carboplatin–etoposide plus atezolizumab was used. Information regarding subsequent lines of therapy was recorded when available. These variables were selected to reflect the information typically available to a medical oncologist at the time of initial treatment decision-making in routine clinical practice.

Overall survival (OS) was defined as the time from the date

of diagnosis to death from any cause or last follow-up. Progression-free survival (PFS) was defined as the time from the initiation of first-line treatment to documented disease progression or death, whichever occurred first. Follow-up duration was calculated from the date of diagnosis to the last clinical contact or death.

For each patient, survival estimates were generated using a large language model–based artificial intelligence system (ChatGPT; OpenAI). The analysis was performed using a GPT-5.2–level model available through the ChatGPT interface at the time of the study. The predictions were generated in January 2026. All patient scenarios were entered sequentially within a single continuous chat session, and each case was evaluated once without repeated queries. The model primarily provided point estimates for survival outcomes; when a range was presented, the midpoint was used for analysis as described below. A standardized prompt structure was applied to all patients to ensure consistency. For each case, a clinical vignette was constructed using baseline information available at the time of diagnosis. The input included age, sex, ECOG performance status, disease stage (extensive-stage SCLC), sites of metastases (including liver, brain, bone, and others), presence of brain metastases and whether cranial radiotherapy was administered, baseline laboratory parameters, and planned first-line treatment (platinum–etoposide or platinum–etoposide plus immunotherapy).

This approach was designed to simulate a real-world baseline prognostic scenario. In this context, the AI model was provided with the same type of clinical information that would typically be available to an oncologist at the time of diagnosis, allowing estimation of expected survival prior to treatment initiation.

A representative example of the prompt used is as follows: "A 68-year-old male with extensive-stage small cell lung cancer, ECOG performance status 2, with liver, bone, and brain metastases. Cranial radiotherapy was administered for brain metastases. Baseline laboratory evaluation showed low hemoglobin (10.9 g/dL), low albumin (3.2 g/dL), and elevated lactate dehydrogenase levels, while other laboratory parameters were within normal limits. The planned first-line treatment is carboplatin–etoposide plus atezolizumab. Based on this information, estimate overall survival and progression-free survival (in months)."

No post-baseline information was provided to the model. Specifically, treatment response, follow-up imaging findings, treatment modifications, toxicities, or subsequent clinical events were not included. The same prompt structure was applied to all patients, and the model was not trained, fine-tuned, or recalibrated using the study dataset. Therefore, the generated outputs represent general-purpose AI-based estimates rather than a disease-specific prediction model. All clinical information entered into the model was fully de-identified, and no direct patient identifiers were used.

The AI system provided survival estimates in months for both overall survival and progression-free survival. In cases where survival estimates were expressed as a range, the midpoint of the interval was used as the final value for analysis. When a single numerical estimate was provided, that value was recorded directly. Predicted OS and PFS values were documented for each patient and used for comparison with observed outcomes.

The study was approved by the local ethics committee (No: 2025/0042; Date: 03.07.2025) and was conducted in accordance with the Declaration of Helsinki. The study was reported in line with STROBE recommendations for observational studies.[23]

Statistical analysis

Descriptive statistics were used to summarize baseline characteristics. Continuous variables were presented as mean ± standard deviation or median (interquartile range), depending on distribution, while categorical variables were expressed as frequencies and percentages.

Observed OS and PFS were estimated using the Kaplan–Meier method, and median survival times with corresponding confidence intervals were calculated for the overall cohort.

Agreement between AI-predicted and observed survival outcomes was evaluated using Bland–Altman analysis to assess both systematic bias (mean difference between predicted and observed values) and individual-level variability (limits of agreement). Bland–Altman analysis was performed in patients with complete event data (i.e., those who experienced death or progression), while censored observations were excluded from this analysis.

All statistical analyses were conducted using IBM SPSS Statistics (version 26.0; IBM Corp., Armonk, NY, USA). Graphical analyses, including Kaplan–Meier survival curves and Bland–Altman plots, were generated using GraphPad Prism (version 9.0; GraphPad Software, San Diego, CA, USA).

Results

A total of 146 patients with extensive-stage small cell lung cancer were included in the analysis. The median age was 65.2 years (range, 42.3–82.4), and the majority of patients were male (61.6%). Most patients were current or former smokers (95.9%). ECOG performance status was 0–1 in 93.2% of patients, while 6.8% had ECOG ≥2. The median

follow-up duration for the entire cohort was 14.8 months (95% CI: 13.2–16.4). First-line treatment consisted of platinum–etoposide chemotherapy alone in 118 patients (80.8%) and platinum–etoposide combined with immunotherapy in 28 patients (20.2%).

At diagnosis, liver metastases were present in 42.5% of patients, brain metastases in 19.2%, and bone metastases in 53.4%. Multiple metastatic sites (≥2) were observed in 38.4% of patients.

Baseline laboratory parameters showed a median white blood cell count of 8.8×10^9/L, platelet count of 307×10^9/L, albumin level of 3.69 g/dL, lactate dehydrogenase (LDH) level of 299 U/L, and C-reactive protein (CRP) level of 17 mg/L.

During follow-up, disease progression occurred in 89% of patients. At the time of last follow-up, 73.3% of patients had died, while 26.7% were alive. Detailed baseline characteristics are presented in Table 1.

The median observed overall survival (OS), estimated using the Kaplan–Meier method, was 10.97 months. The corresponding AI-predicted median OS was 13.2 months, resulting in a positive prediction bias of +2.2 months.

For progression-free survival (PFS), the observed median PFS was 6.53 months, while the AI-predicted median PFS was 7.2 months, corresponding to a prediction bias of +0.7 months.

These findings suggest a tendency toward overestimation of survival outcomes by the AI-based model. A detailed comparison of observed and predicted survival values is provided in Table 2.

Kaplan–Meier analysis demonstrated that AI-predicted survival curves were generally shifted toward longer survival compared with observed outcomes.

For overall survival (Fig. 1), the AI-predicted curve showed higher survival probabilities in the early and intermediate time periods, with a steeper decline observed after approximately 12–15 months. In contrast, the observed survival curve demonstrated a more gradual decline and a longer tail, reflecting a subset of patients with prolonged survival.

For progression-free survival (Fig. 2), the difference between observed and AI-predicted curves was less pronounced but still evident, with AI predictions slightly overestimating PFS across most time points.

Table 1. Baseline demographic and clinicopathological features of patients with extensive-stage small cell lung cancer

Characteristic	n (%) or median (range)
Age, years	65.2 (42.3–82.4)
Sex	
Male	90 (61.6)
Female	56 (38.4)
Smoking status	
Current or former smoker	140 (95.9)
Never smoker	6 (4.1)
ECOG performance status	
0–1	136 (93.2)
≥2	10 (6.8)
Disease stage	
Extensive-stage	146 (100)
Metastatic involvement at diagnosis	
Liver metastasis	62 (42.5)
Brain metastasis	28 (19.2)
Bone metastasis	78 (53.4)
≥2 metastatic sites	56 (38.4)
Baseline laboratory parameters	
White blood cell count, $\times10^9$/L	8.8 (3.9–24.9)
Platelet count, $\times10^9$/L	307 (105–487)
Albumin, g/dL	3.69 (2.88–4.2)
LDH, U/L	299.00 (101.00–2031.00)
CRP, mg/L	17.00 (1.00–344.00)
Treatment regimen	
Chemotherapy	118 (80.8)
Chemotherapy + Immunotherapy	28 (20.2)
Disease progression	
Yes	130 (89)
No / censored	16 (11)
Exitus status at last follow-up	
Alive	39 (26.7)
Deceased	107 (73.3)

ECOG: Eastern Cooperative Oncology Group; LDH: Lactate dehydrogenase; CRP: C-reactive protein.

Table 2. Comparison of observed and AI-predicted survival outcomes in extensive-stage small cell lung cancer

Outcome	Observed median (months)	AI-predicted median (months)	Prediction bias (months)
Overall Survival (OS)	10.97	13.2	+2.2
Progression-Free Survival (PFS)	6.53	7.2	+0.7

OS: Overall survival; PFS: Progression-free survival. Observed median survival values were estimated using the Kaplan–Meier method. AI-predicted medians were generated using a baseline-only simulated prognostic model without incorporation of treatment response or follow-up data. Prediction bias was calculated as the difference between AI-predicted and observed median survival.

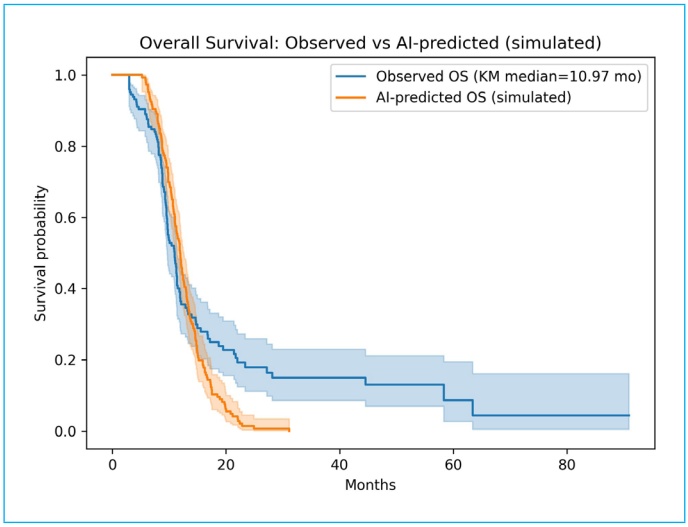

Figure 1. Kaplan–Meier curves for overall survival (OS): observed versus AI-predicted survival. The observed OS curve (blue) is compared with the AI-predicted OS curve (orange). Shaded areas represent 95% confidence intervals. The AI-predicted curve demonstrates a general tendency toward overestimation of survival compared with observed outcomes.

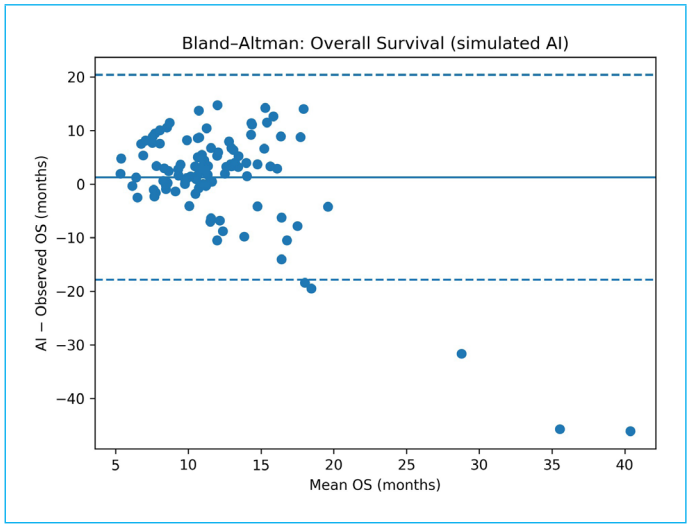

Figure 3. Bland–Altman plot for agreement between AI-predicted and observed overall survival. The plot shows the difference between AI-predicted and observed OS against their mean values. The solid line represents the mean difference, while the dashed lines indicate the limits of agreement. The analysis demonstrates systematic overestimation and considerable inter-individual variability.

Agreement between AI-predicted and observed overall survival was evaluated using Bland–Altman analysis (Fig. 3). The analysis demonstrated a positive mean difference, indicating systematic overestimation of survival by the AI model.

There was also considerable variability at the individual patient level, as reflected by the wide limits of agreement. While most patients clustered around the mean difference, several outliers were observed, particularly at higher mean survival values, suggesting reduced agreement in patients with longer survival.

A similar analysis was performed for progression-free survival (Fig. 4). Bland–Altman evaluation for PFS also demonstrated variability between predicted and observed values, with a distribution of differences around the mean and sev-

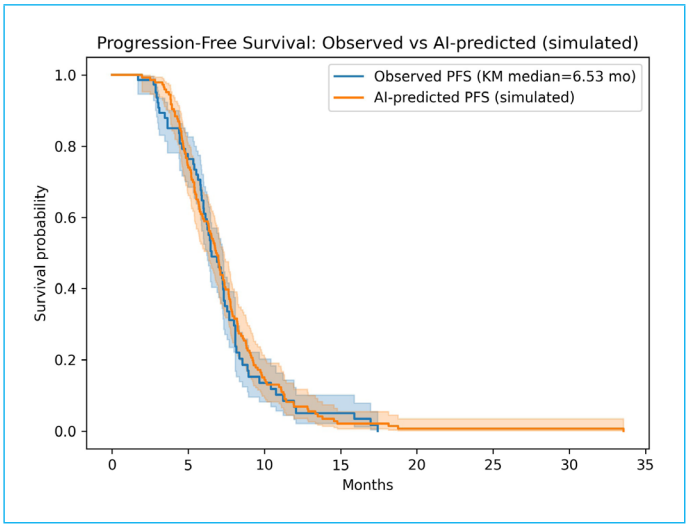

Figure 2. Kaplan–Meier curves for progression-free survival (PFS): observed versus AI-predicted survival. The observed PFS curve (blue) is compared with the AI-predicted PFS curve (orange). Shaded areas represent 95% confidence intervals. AI predictions show a modest but consistent overestimation of PFS.

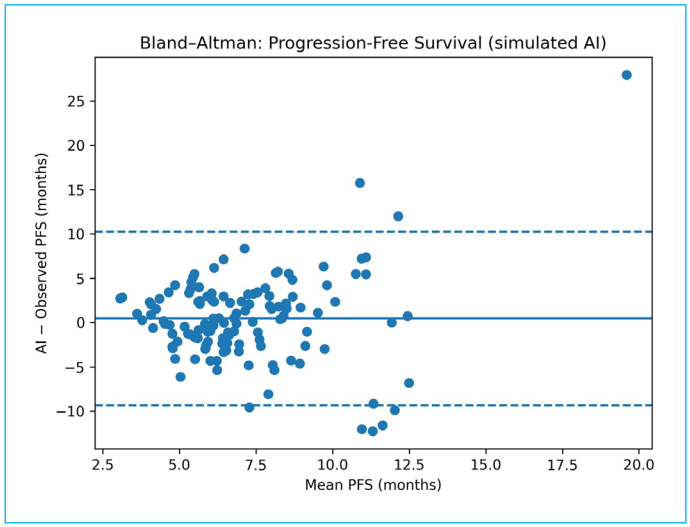

Figure 4. Bland–Altman plot for agreement between AI-predicted and observed progression-free survival (PFS). The plot shows the difference between AI-predicted and observed PFS against their mean values. The solid line represents the mean difference, while the dashed lines indicate the limits of agreement. The distribution demonstrates variability in prediction accuracy with several outliers, indicating imperfect agreement between predicted and observed PFS.

eral notable outliers. Although the average difference appeared smaller compared with overall survival, variability at the individual level remained evident.

Discussion

In this retrospective real-world cohort of patients with extensive-stage small cell lung cancer, we observed that AI-derived survival estimates consistently exceeded the survival outcomes documented in clinical practice. The magnitude of this discrepancy was more pronounced for overall survival, with a median difference of 2.2 months, whereas the corresponding difference for progression-free survival was more modest at 0.7 months. Beyond this average tendency toward overestimation, agreement analysis demonstrated considerable variability at the individual patient level, indicating that prediction error was not uniform across the cohort. Notably, while the AI-based model appeared to approximate short-term outcomes in some patients, its performance was less consistent in those with longer survival durations. Taken together, these findings suggest that baseline AI-generated survival estimates may provide a general directional signal but lack the level of calibration required for reliable patient-level prognostic use in extensive-stage SCLC.

The use of artificial intelligence for prognostic estimation in oncology has expanded rapidly in recent years, driven by the increasing availability of large clinical datasets and advances in both machine learning and natural language processing.[2,7] Earlier approaches largely relied on structured data and traditional machine-learning algorithms, whereas more recent work has explored the use of large language models to generate clinically interpretable predictions from baseline patient summaries.[8,22] Several studies have reported that AI-based systems can distinguish between higher- and lower-risk patients across different tumor types, including lung cancer, hepatocellular carcinoma, and breast cancer.[3,9] However, the ability of these systems to provide accurate, patient-level survival estimates in real-world settings remains uncertain.[10,11] In particular, emerging evidence suggests that although AI-generated outputs may appear clinically plausible, their calibration against observed outcomes is often suboptimal, raising concerns about their direct applicability in routine practice.[13]

Our findings are consistent with emerging reports evaluating AI-based survival estimation in real-world oncology cohorts.[2] In a recent study examining patients with hepatocellular carcinoma, AI-generated survival predictions were similarly found to exceed observed outcomes, with a clear tendency toward overestimation despite the use of clinically relevant baseline inputs.[12] Notably, that study also highlighted variability in prediction accuracy across patient subgroups, suggesting that apparent agreement at the population level may mask substantial individual-level discrepancies.[12] A comparable pattern was observed in our cohort, where the mean difference between predicted and observed survival was accompanied by wide limits of agreement, indicating inconsistent performance across patients. While the underlying disease context differs, the convergence of these findings supports the notion that baseline-only AI models may systematically overestimate survival when applied to heterogeneous real-world populations.[13] In the setting of extensive-stage SCLC, where disease trajectories are often more aggressive and less predictable than in many solid tumors, such discrepancies may become even more clinically relevant, particularly when prognostic estimates are used to inform treatment planning or patient counseling.[15]

Importantly, the pattern observed in our study is not unique to AI-based systems and has been reported across a wide range of conventional prognostic models in oncology.[14] Several externally validated nomograms and machine-learning–based tools have demonstrated good discriminatory performance while simultaneously showing limited calibration when applied to independent patient cohorts.[13] For example, survival prediction models developed in sarcoma, renal cell carcinoma, and breast cancer have been shown to overestimate survival in certain populations, whereas in other settings, underestimation has also been reported depending on cohort characteristics and disease stage.[24,25] This suggests that even well-established models may lose accuracy when transferred outside the populations in which they were developed. In this context, our findings align with a broader methodological issue rather than representing a limitation specific to large language models.[10] However, unlike disease-specific models that can be recalibrated using updated datasets, general-purpose AI systems lack built-in mechanisms for adjustment, which may further amplify discrepancies when applied to heterogeneous real-world cohorts such as patients with extensive-stage SCLC.[2]

Several factors may explain the tendency toward overestimation observed in our cohort. First, the use of baseline-only inputs inherently limits the model's ability to account for dynamic clinical events that strongly influence survival in SCLC, such as early disease progression, treatment-related toxicity, or rapid decline in performance status.[26] These factors often emerge shortly after treatment initiation and can substantially shorten survival, yet they are not captured in a static, diagnosis-based assessment. Second, large language models do not operate as calibrated prognostic tools but rather generate estimates based

on probabilistic associations derived from broad training data.[2,8] As a result, the outputs may reflect an "average" expected trajectory rather than the more variable and often less favorable course observed in real-world patients. Third, differences between clinical trial populations and routine practice may further contribute to this discrepancy. Patients in real-world settings frequently present with greater comorbidity burden, more advanced disease, and less consistent treatment delivery, all of which can negatively affect outcomes.[26] In a disease such as extensive-stage SCLC, where clinical deterioration can occur rapidly, even small deviations from idealized conditions may translate into meaningful differences in survival, thereby contributing to the observed gap between predicted and actual outcomes.[15]

The clinical implications of these findings become more apparent when considered in the context of extensive-stage SCLC. Despite advances with the addition of immune checkpoint inhibitors to first-line chemotherapy, outcomes remain poor and highly variable, and prognostic assessment at diagnosis continues to be challenging.[16,17] At the same time, SCLC is increasingly recognized as a biologically heterogeneous disease, with distinct molecular and transcriptional subtypes that may differ in treatment sensitivity and clinical behavior.[18] Although these classifications have not yet been fully integrated into routine clinical decision-making, they are beginning to inform the development of novel therapeutic strategies, including subtype-directed approaches and targeted agents such as antibody–drug conjugates.[19,20] In this evolving landscape, early identification of patients who are likely to have particularly unfavorable outcomes may become increasingly relevant. From this perspective, AI-based tools hold theoretical appeal, as they offer the possibility of integrating multiple baseline variables into a single prognostic estimate.[2] However, our findings suggest that, in their current form, such tools may not yet provide the level of precision required to guide individualized treatment decisions in SCLC, particularly when used without dynamic clinical input.[10]

An additional observation in our study was the difference in prediction accuracy between overall survival and progression-free survival. While both endpoints were overestimated by the AI model, the magnitude of discrepancy was notably greater for overall survival. This difference may reflect the inherently distinct nature of these endpoints. Progression-free survival is more directly related to tumor biology and initial treatment response, and therefore may be more predictable from baseline disease characteristics. In contrast, overall survival is influenced by a broader range of factors, including subsequent lines of therapy, treatment tolerability, supportive care, and intercurrent clinical events.[26] These post-baseline variables are not captured in a static input model and may introduce additional variability that reduces prediction accuracy.[13] The larger divergence observed in OS predictions in our cohort is consistent with this interpretation and highlights the limitations of relying solely on baseline data when attempting to estimate long-term outcomes in patients with advanced cancer.

From a clinical perspective, these findings suggest that AI-derived survival estimates should be interpreted with caution when used in routine oncology practice.[22] While such tools may offer a rapid and structured way to synthesize baseline clinical information, their outputs should not be considered definitive or used in isolation for decision-making. In particular, the tendency toward overestimation observed in our study may lead to overly optimistic expectations if not interpreted within the appropriate clinical context. Instead, AI-based predictions may be better positioned as adjunctive tools that complement, rather than replace, clinical judgment.[27] They may provide a general sense of disease trajectory or facilitate discussions in situations where prognostic uncertainty is high, but their limitations—especially in terms of calibration and individual-level variability—must be clearly recognized.[13] As AI applications continue to evolve, integrating dynamic clinical data, treatment response, and longitudinal follow-up into prediction models may improve their reliability and clinical relevance.[2]

Strengths and Limitations

This study has several strengths. First, the analysis was conducted in a relatively homogeneous cohort consisting exclusively of patients with extensive-stage SCLC, a setting in which prognostic estimation is both clinically relevant and challenging. Second, the sample size was adequate for a real-world study in this disease context, and all patients had complete baseline clinical, laboratory, and outcome data, enabling consistent comparison between predicted and observed survival. Third, the use of a standardized prompt structure ensured that all patients were evaluated under comparable conditions, minimizing variability related to input differences. In addition, the follow-up duration was sufficient to capture meaningful survival outcomes in a malignancy characterized by short median survival.

However, several limitations should also be acknowledged. The retrospective design introduces the potential for selection bias and unmeasured confounding. Furthermore, AI-based predictions were generated using baseline data only and did not incorporate dynamic clinical variables, which likely contributed to the observed discrepancies. We also did not perform subgroup analyses according to treatment modality, such as separating patients who received chemotherapy alone from those treated with chemoimmunotherapy, and

it is possible that prediction performance may differ across these groups. Although a standardized prompt was used, outputs from large language models may still be influenced by contextual variability that cannot be fully controlled. Finally, while our approach reflects a realistic clinical scenario, it does not account for potential improvements achievable through model recalibration or the incorporation of longitudinal data. Future studies integrating dynamic clinical inputs and treatment response variables may help enhance the accuracy and clinical applicability of AI-based prognostic tools.

Conclusion

AI-based survival estimates generated from baseline clinical information tend to overestimate both overall and progression-free survival in patients with extensive-stage SCLC, with a more pronounced discrepancy observed for overall survival. Although these models may capture general prognostic trends, their limited calibration and substantial variability at the individual level restrict their direct use in clinical decision-making. These findings highlight the need for careful validation of AI-based tools in disease-specific, real-world settings. Future approaches incorporating dynamic clinical data and treatment response may improve the accuracy and clinical utility of AI-driven prognostic models in oncology.

Disclosures

Ethics Committee Approval: Institution granting ethics approval: Istanbul Medeniyet University Ethics Committee; Approval Number: 2025/0042; Approval Date: 03.07.2025

Informed Consent: This study was conducted retrospectively using de-identified patient data. Informed consent was waived by the ethics committee.

Author Contributions: Concept – FA, MG; Design – FA, MG; Supervision – MG, FA; Fundings – None; Materials – FA; Data Collection and/or Processing – FA; Analysis and/or Interpretation – FA; Literature Review – FA; Writing – FA, MG; Critical Review – MG, FA.

Conflict of Interest: The authors declare that there are no conflicts of interest related to this study.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Use of AI for Writing Assistance: Artificial intelligence tools (ChatGPT, OpenAI) were used for language editing and proofreading of the manuscript. The AI system was also used as a study tool to generate survival predictions for analysis. The authors take full responsibility for the content of the manuscript.

Peer-review: Externally peer reviewed

References

1. Topol EJ. High-performance medicine: The convergence of human and artificial intelligence. Nat Med 2019;25(1):44–56. [CrossRef]
2. Kolla L, Parikh RB. Uses and limitations of artificial intelligence for oncology. Cancer 2024;130(12):2101–7. [CrossRef]
3. Huang D, Li Z, Jiang T, Yang C, Li N. Artificial intelligence in lung cancer: Current applications, future perspectives, and challenges. Front Oncol 2024;14:1486310. [CrossRef]
4. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer 2018;18(8):500–10. [CrossRef]
5. Roche JJ, Seyedshahi F, Rakovic K, Thu AW, Le Quesne J, Blyth KG. Current and future applications of artificial intelligence in lung cancer and mesothelioma. Thorax 2025;80(12):957–65. [CrossRef]
6. Bi WL, Hosny A, Schabath MB, Giger ML, Birkbak NJ, Mehrtash A, et al. Artificial intelligence in cancer imaging: Clinical challenges and applications. CA Cancer J Clin 2019;69(2):127–57. [CrossRef]
7. Mehan N, Desinghe TD, Saha A. Development and evaluation of large-language models (LLMs) for oncology: A scoping review. PLOS Digit Health 2025;4(8):e0000980.
8. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023;620(7972):172–80. [CrossRef]
9. Yuan L, An L, Zhu Y, Duan C, Kong W, Jiang P, et al. Machine learning in diagnosis and prognosis of lung cancer by PET-CT. Cancer Manag Res 2024;16:361–75. [CrossRef]
10. Collins GS, Chester-Jones M, Gerry S, Ma J, Matos J, Sehjal J, et al. Clinical prediction models using machine learning in oncology: Challenges and recommendations. BMJ Oncol 2025;4(1):e000914. [CrossRef]
11. Smiley A, Villarreal-Zegarra D, Reategui-Rivera CM, Escobar-Agreda S, Finkelstein J. Methodological and reporting quality of machine learning studies on cancer diagnosis, treatment, and prognosis. Front Oncol 2025;15:1555247. [CrossRef]
12. Kavak EE, Erdat EC, Altundağ Derin Z, Dilli İ, Kubilay Tolunay P, Öksüzoğlu B, et al. Comparison of AI chatbot predicted and realworld survival outcomes in hepatocellular carcinoma. Sci Rep 2025;15(1):21728. [CrossRef]
13. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW; Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: The Achilles heel of predictive analytics. BMC Med 2019;17(1):230. [CrossRef]
14. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: A framework for traditional and novel measures. Epidemiology 2010;21(1):128–38. [CrossRef]
15. Kim SY, Park HS, Chiang AC. Small cell lung cancer: A review. JAMA 2025;333(21):1906–17. [CrossRef]
16. Horn L, Mansfield AS, Szczęsna A, Havel L, Krzakowski M, Hochmair MJ, et al. First-line atezolizumab plus chemother-

apy in extensive-stage small-cell lung cancer. N Engl J Med 2018;379(23):2220–9. [CrossRef]
17. Paz-Ares L, Dvorkin M, Chen Y, Reinmuth N, Hotta K, Trukhin D, et al. Durvalumab plus platinum-etoposide versus platinum-etoposide in first-line treatment of extensive-stage small-cell lung cancer (CASPIAN): A randomised, controlled, open-label, phase 3 trial. Lancet 2019;394(10212):1929–39. [CrossRef]
18. Gay CM, Stewart CA, Park EM, Diao L, Groves SM, Heeke S, et al. Patterns of transcription factor programs and immune pathway activation define four major subtypes of SCLC with distinct therapeutic vulnerabilities. Cancer Cell 2021;39(3):346–60.e7. [CrossRef]
19. Sen T, Takahashi N, Chakraborty S, Takebe N, Nassar AH, Karim NA, et al. Emerging advances in defining the molecular and therapeutic landscape of small-cell lung cancer. Nat Rev Clin Oncol 2024;21(8):610–27. [CrossRef]
20. Zhang J, Zeng X, Guo Q, Sheng Z, Chen Y, Wan S, et al. Small cell lung cancer: Emerging subtypes, signaling pathways, and therapeutic vulnerabilities. Exp Hematol Oncol 2024;13(1):78. [CrossRef]
21. Mennecier B, Khalifa J, Descourt R, Greillier L, Naltet C, Falchero L, et al. Real-life clinical management patterns in extensive-stage small cell lung cancer across France: A multi-method study. BMC Cancer 2024;24(1):421. [CrossRef]
22. Wong EYT, Verlingue L, Aldea M, Franzoi MA, Umeton R, Halabi S, et al. ESMO guidance on the use of large language models in clinical practice (ELCAP). Ann Oncol 2025;36(12):1447–57. [CrossRef]
23. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. Ann Intern Med 2007;147(8):573–7. [CrossRef]
24. Utsumi T, Ueda T, Fukasawa S, Komaru A, Sazuka T, Kawamura K, et al. Prognostic models for renal cell carcinoma recurrence: External validation in a Japanese population. Int J Urol 2011;18(9):667–71. [CrossRef]
25. Bongers MER, Thio QCBS, Karhade AV, Stor ML, Raskin KA, Lozano Calderon SA, et al. Does the SORG algorithm predict 5-year survival in patients with chondrosarcoma? An external validation. Clin Orthop Relat Res 2019;477(10):2296–303. [CrossRef]
26. Booth CM, Tannock IF. Randomised controlled trials and population-based observational research: Partners in the evolution of medical evidence. Br J Cancer 2014;110(3):551–5. [CrossRef]
27. Parikh RB, Obermeyer Z, Navathe AS. Regulation of predictive analytics in medicine. Science 2019;363(6429):810–2. [CrossRef]