BMJ Open

Protocol for an observational cohort study integrating real-world data and microsimulation to assess imaging surveillance strategies in stage I–IIIA NSCLC patients in OneFlorida+

Dejana Braithwaite ,[1,2,3] Shama D Karanth ,[1,2] Jiang Bian ,[4,5] Rafael Meza,[6] Jihyoun Jeon,[7] Martin Tammemagi,[8,9] Meghann Wheeler ,[3] Pianpian Cao,[10] Mindaugas Rackauskas,[11] Pratibha Shrestha,[1] Hyung-Suk (Alex) Yoon,[1,2] Andrea Borondy Kitts,[12] Himanshi Verma,[1] Marissa Carmella Blair,[1] Aokun Chen,[13] Devika Das,[14] Xiangyang Lou,[15] Yonghui Wu,[13] Summer Han,[16] Bruno Hochhegger,[17] Yi Guo ,[13] MK Gould[18]

To cite: Braithwaite D, Karanth SD, Bian J, *et al*. Protocol for an observational cohort study integrating real-world data and microsimulation to assess imaging surveillance strategies in stage I–IIIA NSCLC patients in OneFlorida+. *BMJ Open* 2025;**15**:e101324. doi:10.1136/bmjopen-2025-101324

► Prepublication history for this paper is available online. To view these files, please visit the journal online (https://doi.org/10.1136/bmjopen-2025-101324).

Received 26 February 2025
Accepted 10 September 2025

Check for updates

© Author(s) (or their employer(s)) 2025. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ Group.

For numbered affiliations see end of article.

Correspondence to
Dr Dejana Braithwaite; dbraithwaite@ufl.edu

ABSTRACT

Introduction Although lung cancer remains the leading cause of cancer deaths in the US, recent advances in early detection and treatment have led to improvements in survival. However, there is a considerable risk of recurrence or second primary lung cancer (SPLC) following curative-intent treatment in patients with early-stage non-small cell lung cancer (NSCLC). Professional societies recommend routine surveillance with CT to optimise the detection of potential recurrence and SPLC at a localised stage. However, no definitive evidence demonstrates the effect of imaging surveillance on survival in patients with NSCLC. To close these research gaps, the Advancing Precision Lung Cancer Surveillance and Outcomes in Diverse Populations (PLuS2) study will leverage real-world electronic health records (EHRs) data to evaluate surveillance outcomes among patients with and without guideline-adherent surveillance. The overarching goal of the PLuS2 study is to assess the long-term effectiveness of surveillance strategies in real-world settings.

Methods and analysis PLuS2 is an observational study designed to assemble a cohort of patients with incident pathologically confirmed stage I/II/IIIA NSCLC who have completed curative-intent therapy. Patients undergoing imaging surveillance will be followed from 2012 to 2026 by linking EHRs with tumour registry data in the OneFlorida+ Clinical Research Consortium. Data will be consolidated into a unified repository to achieve three primary aims: (1) Examine the utilisation and determinants of CT imaging surveillance by race/ethnicity and socioeconomic status, (2) Compare clinical endpoints, including recurrence, SPLCs and survival of patients who undergo semiannual versus annual CT imaging and (3) Use the observational data in conjunction with validated microsimulation models to simulate imaging surveillance outcomes within the US population. To our knowledge, this study represents the first attempt to integrate real-world data and microsimulation models to assess the long-term impact and effectiveness of imaging surveillance strategies.

Ethics and dissemination This study involves human participants and was approved by the University of Florida Institutional Review Board (IRB), University of Florida IRB 01, under approval number IRB202300782. The results will be disseminated through publications and presentations at national and international conferences. Safety considerations encompass ensuring the confidentiality of patient information. All disseminated data will be de-identified and summarised.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ Leverages real-world data from a large, diverse clinical research network to evaluate imaging surveillance practices in early-stage non-small cell lung cancer.
⇒ Use natural language processing to extract imaging indications from unstructured electronic health record (EHR) notes, capturing information not available in structured fields.
⇒ Integrates observational data with microsimulation modelling to estimate long-term outcomes under different surveillance strategies.
⇒ Non-randomised surveillance intervals and reliance on EHR documentation may introduce confounding and incomplete capture of recurrence or second primary lung cancers.
⇒ Real-world data may be subject to selection bias and missing information; mitigation strategies include multiple imputations and sensitivity analyses.

INTRODUCTION

Lung cancer ranks as the second most prevalent cancer in both men and women, and it stands as the primary contributor to cancer-related fatalities in the US, accounting for ~25% of all cancer-related deaths.[1] Recent

BMJ Open: first published as 10.1136/bmjopen-2025-101324 on 28 September 2025. Downloaded from http://bmjopen.bmj.com/ on March 5, 2026 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

advances in screening and treatment modalities have led to notable improvements in the prognosis for patients with early-stage non-small cell lung cancer (NSCLC).[2 3] The 5-year survival rate for individuals with early-stage NSCLC has seen a steady upward trajectory, reflecting the efficacy of advanced medical interventions and multi-disciplinary approaches to care.[4] However, despite these positive trends, many patients experience disease recurrence or develop secondary primary lung cancer (SPLC) following curative treatment.[5] Studies suggest that recurrence rates in this patient population range from 30%–50% and 1%–2% yearly risk of SPLC.[6 7] These statistics underscore the potential benefits of surveillance and early detection of cancer recurrence or SPLC in this group.[8]

The National Comprehensive Cancer Network (NCCN) and the European Society for Medical Oncology Clinical Practice Guidelines recommend CT scans every 6 months for the first 2–3 years after curative-intent therapy, followed by annual scans for up to 5 years[9 10] These guidelines are consensus-based rather than evidence-based, as the impact of regular surveillance on long-term outcomes remains unclear.[11] The IFCT-0302 trial compared chest X-ray to CT for surveillance every 6 months for the first 2 years, followed by annual scans until 5 years, and found that while CT detected more recurrences and SPLCs, it did not improve overall survival.[12]

In the US, providers have embraced NCCN guidelines, which have driven the increased utilisation of CT surveillance.[11 13–15] This trend has persisted despite the lack of definitive evidence regarding the optimal frequency of CT surveillance in reducing mortality rates.[16 17] Notably, no clinical trials have demonstrated the efficacy of CT surveillance for NSCLC populations in the US Extrapolating findings from the IFCT-0302 study to an American population is challenging due to the underrepresentation of minority patients.[12] Given that Black patients with localised and loco-regional NSCLC have poorer outcomes than other patients,[4] it is important to address clinical endpoints in that population.

The overarching goal of the Advancing Precision Lung Cancer Surveillance and Outcomes in Diverse Populations (PLuS2) study is to establish a unique data repository to address critical gaps in lung cancer surveillance following curative-intent therapy. The study aims to: (1) Assess the utilisation and determinants of CT imaging surveillance among patients treated with curative intent for early-stage NSCLC, distinguishing between routine (asymptomatic) imaging surveillance versus symptom-directed diagnostic surveillance and stratifying the utilisation patterns by race/ethnicity and socioeconomic status (SES), (2) Compare clinical endpoints, including recurrence, SPLCs and survival, among patients undergoing semi-annual versus annual CT imaging surveillance and (3) Evaluate the comparative effectiveness of imaging surveillance strategies by projecting long-term/lifetime progression and survival outcomes using a previously validated microsimulation model. The conceptual framework for the PLuS2 study is presented in figure 1.

This study will use unstructured and structured data from electronic health records (EHRs), linked with tumour registries, death index, claims and census-tract level data to determine the role of CT surveillance among a diverse population with early-stage NSCLC that received curative-intent therapy.

METHODS AND ANALYSIS

Study design and setting

The PLuS2 study will use real-world data from the OneFlorida+ Clinical Research Consortium.[18] The OneFlorida+ Consortium contains extensive longitudinal patient-level EHR records for approximately ~18 million Floridians with data compiled through various sources, including selected Medicaid and Medicare, cancer registries, vital statistics and EHRs from clinical partners in the Network.[18]

Study population

The target population will include University of Florida (Gainesville and Jacksonville health centres) patients within the OneFlorida+ Consortium who have received a diagnosis of primary pathologically confirmed stage I-IIIA NSCLC and have completed curative-intent therapy and survived for at least 5 months post-therapy initiation. Structured data will be extracted from the EHR (linked with selected claims and tumour registries), while unstructured data will be abstracted from the free-text clinical encounter, radiology and pathology reports. Natural language processing (NLP) pipelines will be developed and validated to ascertain if the indication for imaging was routine surveillance in asymptomatic patients or diagnostic patients with symptoms or signs. Symptoms suggestive of recurrence/SPLC, social determinants of health (SDOH) and behavioural risk factors (eg, smoking, physical activity) will be extracted as detailed in the NLP section below.

Natural language processing to comprehensively characterise the study population

In this study, we will employ NLP methods to extract features from clinical notes, radiology reports and pathology reports. NLP will identify and categorise symptoms such as cough, chest pain, haemoptysis, shortness of breath, unexplained weight loss, recurrent respiratory infections, weakness, fatigue and loss of appetite that prompt symptom-driven CT imaging surveillance.[19] The features extracted from unstructured data will be used to classify CT imaging surveillance scans as either routine/scheduled surveillance or symptom-driven diagnostic orders.[20 21]

Additionally, NLP will extract missing cancer stage information, histological subtypes from pathology reports and SDOH variables from clinical notes.[22 23] The NLP workflow will involve the following steps: (1) Annotate concepts of interest using a dataset of 200 de-identified clinical notes and radiology reports, (2) Generate

BMJ Open: first published as 10.1136/bmjopen-2025-101324 on 28 September 2025. Downloaded from http://bmjopen.bmj.com/ on March 5, 2026 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

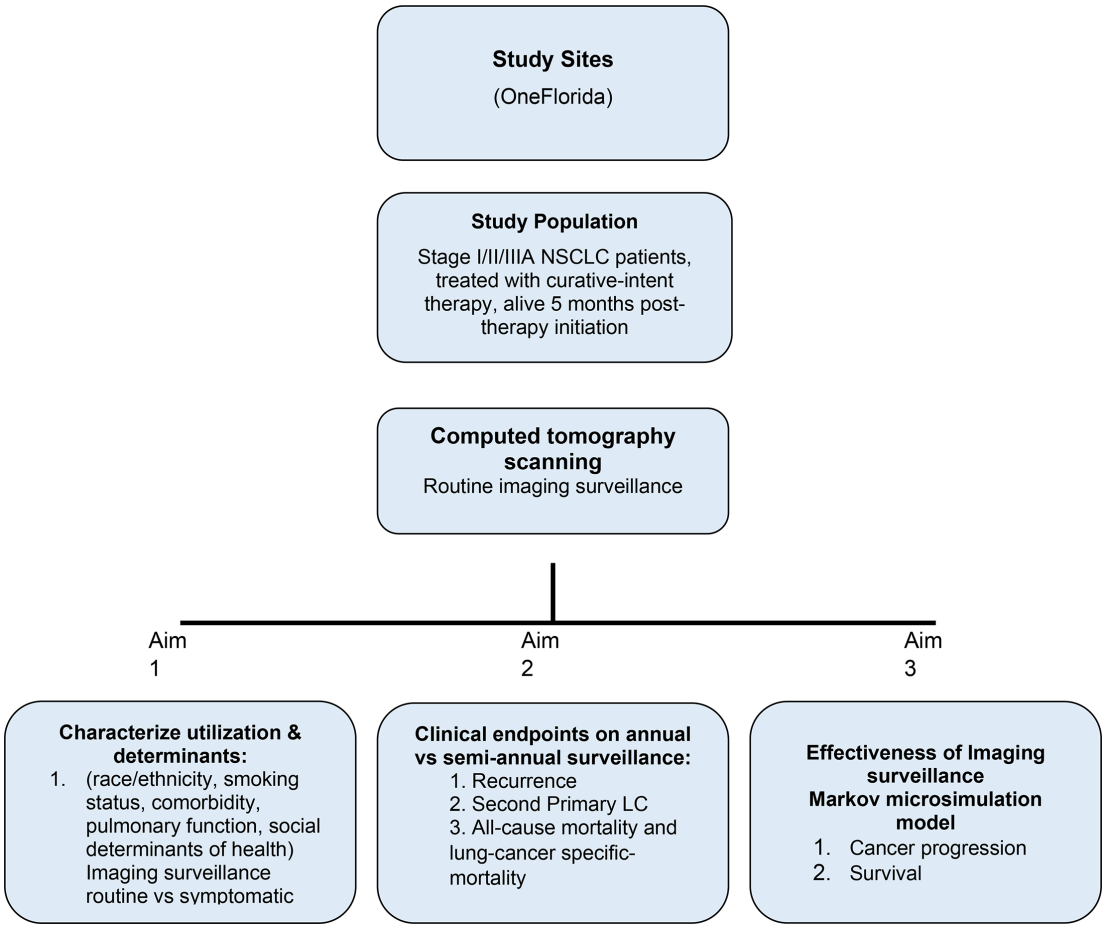


Figure 1 PLuS2 design overview. LC, lung cancer; NSCLC, non-small cell lung cancer.

pre-annotation results using a large language model (LLM), followed by corrections from human annotators, (3) Develop and evaluate NLP models by instruction-tuning the LLM for concept extraction tasks and (4) Extract the identified concepts from the broader dataset.

We will use Meta-Llama-3-70B-Instruct as the LLM for pre-annotation and concept extraction tasks.[24] Model performance will be evaluated using standard metrics, including accuracy, precision, recall and F1-score,[25] calculated on a held-out validation set comprising 200 manually annotated clinical notes and radiology reports. Inter-rater reliability for human annotations will be assessed using Cohen's Kappa statistic to ensure consistency and quality of labelled data. All NLP workflows will be implemented using reproducible pipelines, and annotated datasets and model code will be made available on reasonable request to support transparency and replication.

Comorbidity

Comorbidity will be assessed using the Charlson Comorbidity Index[20 21 26] and/or the Elixhauser comorbidity score,[22 27] both extracted from EHR data.[28 29] This selected method will be used for its proven effectiveness in summarising comorbidities accurately.[30] To construct comorbidity burden, we will use International Classification of Diseases-10 (ICD-10) and Current Procedural Terminology (CPT) codes from EHR data spanning the 12 months prior to NSCLC treatment or imaging, applying NLP techniques as needed.[20–22 31] Patients will be stratified based on their health status into three categories: those in optimal health (Charlson/Elixhauser score of 0), those in moderate health (score of 1) and those in poor health (score of 2 or higher).[32] Comorbidity status will be regarded as a dynamic variable, varying over time and dependent on the recorded health status at each imaging surveillance time-point.

Pulmonary function

History of Chronic Obstructive Pulmonary Disease (COPD) will be identified through the EHR data and categorised as yes versus no.[33 34] Additionally, COPD will be classified using the updated global initiative for obstructive lung disease classification, comprising categories A, B, C and D based on spirometry indicators.[35 36] Additionally, we will evaluate forced expiratory volume in 1 s (FEV1) and ejection fraction for patients with heart failure. Impaired pulmonary function will be defined as FEV1/forced vital capacity <70%.[34]

Functional limitations

Functional limitation indicators will be identified using claims and EHR data using the previously validated Segal Index.[37–39] This inclusion will aid in determining whether the lack of adherence to recommended imaging

BMJ Open: first published as 10.1136/bmjopen-2025-101324 on 28 September 2025. Downloaded from http://bmjopen.bmj.com/ on March 5, 2026 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

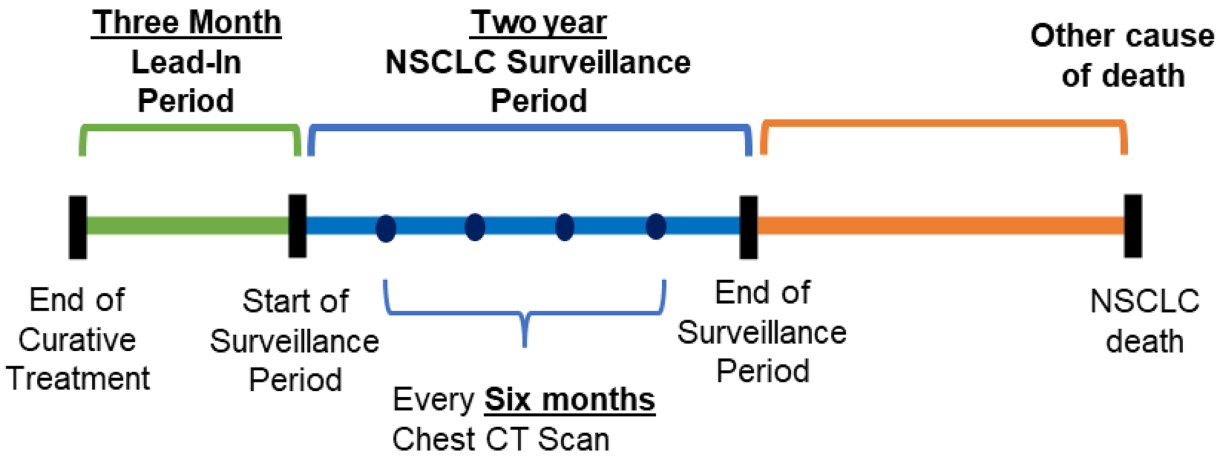


Figure 2 PLuS2 conceptual model guideline adherence is based on the National Comprehensive Cancer Network's published guidelines. The 'Lead-in Period' refers to the non-surveillance period following the end of curative treatment. NSCLC, non-small cell lung cancer.

procedures can be attributed to functional deficits. Mirroring the approach used for comorbidity assessment, we will evaluate functional limitations over the 12 months preceding imaging surveillance. The Segal Index is claims-based, with frailty scores ranging from 0–1, with higher scores indicating greater functional impairment.[37 38] For this study, we will use a cut-off of 0.2 to classify individuals as frail, identifying those with significant functional limitations that may influence adherence to imaging procedures.

Adherence to surveillance imaging

Using ICD-9/ICD-10 procedure codes and CPT-4 codes, we will determine the intervals of surveillance receipt based on published guidelines by the NCCN.[13] Guideline adherence will be defined as receipt of a chest CT scan every 6 months (range: 5–7 months) during the first 2 years, followed by an annual chest CT scan (range: 11–13 months) thereafter (figure 2), starting from the date of surgery or stereotactic body radiation therapy. For cases of non-adherence to imaging surveillance, we will evaluate the interval from the completion of curative-intent treatment to the point of deviation from CT surveillance guidelines. Secondary outcomes for Aim 1 include distinguishing between routine (asymptomatic) surveillance and symptom-directed diagnostic imaging (eg, worsening shortness of breath, worsening cough, new chest pain, weight loss or new chest X-ray abnormality). NLP models will be employed to determine whether each CT scan represents routine surveillance or a symptom-directed diagnostic scan. We will also examine patient-level (eg, race/ethnicity, smoking status, comorbidities, pulmonary function, SDOH) and system-level factors (eg, rurality, practice type) for their impact on surveillance adherence.

Additional outcomes: recurrence and mortality

Other outcomes include recurrence, SPLC, all-cause mortality and lung cancer-specific mortality. Recurrence will be identified through EHR using validated algorithms.[23] More specifically, NLP will be used to extract relevant information from unstructured records, such as radiology reports and physician notes, to indicate recurrence and SPLC.[23] This process will align with the two-phase approach in the study by Hassett *et al*,[23] where Phase 1 involves detecting recurrence and Phase 2 focuses on determining recurrence timing. We will extract information, including SPLC tumour stage at diagnosis, pathohistological type, adjuvant therapy received post-curative-intent therapy, and the presence or absence of metastases, to enhance the accuracy and timing of recurrence detection.

To evaluate all-cause and lung cancer-specific mortality, data from the cancer registry and the national death index will be used to obtain vital status, underlying causes of death and age at death. Additional outcomes include the rate of false positives, complications following imaging surveillance and the receipt of treatment among patients who develop recurrence or SPLC.

Microsimulation model

Aim 3 will involve the development of a discrete-state Markov microsimulation model. This model will integrate two essential components: the natural progression of cancer and a surveillance model. Together, these components will enhance our understanding of cancer progression and survival following the initial diagnosis. The Markov microsimulation model will enable us to estimate several key outcomes, including the number of cancer recurrences, SPLC-specific deaths that are prevented, and the life years gained under both semiannual and annual surveillance schedules. Model inputs, which detail the parameters and variables used in the simulation, are presented in table 1.

Potential confounders or effect modifiers

Patient-level factors such as age, race/ethnicity, sex, health insurance status and smoking status (including pack years smoked) will be extracted from EHR using previously validated methods.[33 40] Additionally, area-level measures of SES, measured at the census block level,

BMJ Open: first published as 10.1136/bmjopen-2025-101324 on 28 September 2025. Downloaded from http://bmjopen.bmj.com/ on March 5, 2026 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Table 1 Microsimulation model inputs and potential estimation approaches

Model inputs	Estimation approach
Surveillance schedule	Patient or group-specific schedules
Proportion of stage I/II/IIIA NSCLC patients who would not develop recurrence/SPLC during lifetime	Proportion of progression-free stage I/II/IIIA NSCLC patients during 5-year follow-up
Proportion of stage I/II/IIIA NSCLC patients who develop recurrence during lifetime	Proportion of stage I/II/IIIA NSCLC patients with recurrence during 5-year follow-up
Proportion of stage I/II/IIIA NSCLC patients who would develop SPLC during their lifetime	Proportion of stage I/II/IIIA NSCLC patients who develop SPLC during 5-year follow-up
Average time to detectability	Calibrate to progression-free survival
Average time to recurrence of becoming symptomatic	Calibrate to the rate of symptomatic detection of SPLC
False positive rate of a CT surveillance scan	Estimated from observed data
Complication rate after CT surveillance scan	Estimated from observed data
Excess mortality from treatment by comorbidity level	Mortality rate from treatment toxicity
Mortality rate from recurrence or SPLC detected by symptoms	Mortality rate from recurrence or SPLC detected by symptoms
Mortality rate from recurrence or SPLC detected by surveillance	Mortality rate from recurrence or SPLC detected by surveillance
Mortality rate from other causes by comorbidity level	Mortality rate from other causes

NSCLC, non-small cell lung cancer; SPLC, second primary lung cancer.

include the diversity index score (reflecting racial and ethnic diversity in a geographic area), median household income, average annual health insurance and public transportation expenditures, the proportion of adults with a college degree, and the proportion of adults with internet access.[41] Tumour registry-derived data will be obtained for all patients, including month and year of diagnosis, cancer site, histological type, stage at diagnosis, survival time, vital status and cancer treatment.

Patient and public involvement

The study does not involve patient or public participation in its design and execution.

Sample size and power analysis

Based on our initial data analysis conducted within the OneFlorida+ Consortium from 2012 to 2022, we anticipate a total sample size of approximately 1729 distinct patients diagnosed with stage I-IIIA NSCLC for the years 2012 to 2026. Within our cohort, we expect to observe recurrence rates ranging from 11%–22%, SPLCs occurring in 7%–11% of cases and deaths accounting for 28%–40% of cases.

Current guidelines recommend semi-annual surveillance for the first 2 years after curative-intent therapy, followed by annual surveillance for the next 3 years. In this study, we aim to compare this standard approach with 'lesser' surveillance strategies, including annual surveillance during the first 2 years and no surveillance.

The available pilot data, in combination with published estimates, indicate that 45.6% of patients undergo semi-annual CT surveillance, while 15.8% opt for annual CT surveillance. Our power and sample size considerations focused on examining the minimum ORs detectable based on the current sample size, focusing on the variable of receiving any chest CT scan during follow-up and using race/ethnicity as a primary predictor of interest. For example, data from OneFlorida+ showed that 59.7% of Whites, 13.4% of Hispanics, 13.3% of Blacks, 1.3% of Asians and 8.0% of individuals from other racial/ethnic backgrounds received at least one surveillance chest CT scan. Given this distribution and using Whites as the reference group, a target sample size of 1729 will have 80% power to detect ORs of ≤0.66 (or ≥1.51) for NH-Blacks and ≤0.66 (or ≥1.50) for Hispanics, respectively, for receipt of surveillance chest CT at a type I error rate of 0.05, which is within the plausible range of CT surveillance effect sizes reported in the existing literature.[16 42–45] Importantly, subgroup analyses stratified by race/ethnicity or SES will be conducted using the Florida cohort, which provides sufficient diversity and sample size to support these comparisons.

Our study is positioned to detect significant differences in recurrence and SPLC rates by surveillance interval, aligning with the primary hypothesis of Aim 2, which posits that semi-annual lung cancer surveillance for asymptomatic recurrence leads to more favourable outcomes compared with annual surveillance. These outcomes encompass earlier detection of recurrence or SPLC, higher diagnosis rates and extended overall survival, compared with annual surveillance starting 2 years after curative-intent therapy. Based on the guidelines, we will compare semi-annual surveillance to annual and no surveillance in the first 2 years and then assess annual surveillance against no surveillance in the following 3 years.

BMJ Open: first published as 10.1136/bmjopen-2025-101324 on 28 September 2025. Downloaded from http://bmjopen.bmj.com/ on March 5, 2026 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Analytical plan

Aim 1 will evaluate CT imaging surveillance utilisation patterns and their drivers in patients with stage I-IIIA NSCLC. Descriptive statistics will summarise all variables for the study cohort overall and by race/ethnicity, comorbidity burden, functional limitations and impaired pulmonary function. Aggregate and interval-specific utilisation rates will be calculated, and NLP will classify CT scans as routine surveillance or symptom-driven. Adherence to guidelines will be assessed, and differences by predictors will be tested using χ^2, t-tests or Analysis Of Variance (ANOVA). Mixed-effects logistic regression will examine associations between race/ethnicity, SES and repeated CT surveillance.[46]

The second aim will evaluate recurrence and SPLC rates, time to recurrence or SPLC, and survival. Cumulative risk of each outcome will be estimated using the Kaplan-Meier method,[47] and differences in outcomes between the two surveillance strategies will be measured using the log-rank test. Logistic regression will be used to compare the rates of recurrence and SPLC by surveillance strategies, and Cox proportional hazard models will be used to estimate the hazard ratios for recurrence, SPLC or mortality. For lung cancer-specific mortality, deaths from other causes will be treated as a competing risk. Specifically, we will use the Fine-Grey sub-distribution hazard model to estimate the cumulative incidence of lung cancer-specific mortality while accounting for competing risks.[48] If the proportional hazards assumption for Cox regression is violated, we will consider alternative approaches such as stratified Cox models or flexible parametric survival models to ensure robust estimation.[49]

If the proportional hazards assumption is violated for time-varying covariates such as comorbidities and functional status, we will use extended Cox models to incorporate time-dependent covariates.[50] Additionally, we will consider marginal structural models with inverse probability weighting to address potential time-dependent confounding, particularly when covariates may be influenced by prior surveillance exposure.[51]

Subgroup analyses will be performed to explore whether the primary associations between imaging surveillance strategy and lung cancer outcomes are modified by sociodemographic (race/ethnicity and SES) or clinical characteristics (comorbidity burden, functional limitations, pulmonary function). Interaction terms will be created between the surveillance strategy and the stratification variables, and the Wald test will be used to determine if any interactions are significant ($p<0.05$).[52] The findings from these analyses will serve as inputs for our microsimulation models in Aim 3.

The third aim will focus on evaluating the comparative effectiveness of various lung cancer surveillance strategies, considering their implications for long-term outcomes in a diverse population of stage I-IIIA NSCLC patients. To achieve this aim, a discrete state Markov microsimulation model will be developed to model cancer progression and survival. We will model the time to recurrence or SPLC and excess mortality due to treatment, death from disease, recurrence, SPLC or death from other causes (figure 3). Assuming a percentage of patients with stage I-IIIA NSCLC will eventually get a recurrence, SPLC or cure, based on the observed proportions of recurrence or SPLC from the Aim 2 data, we will model the time to detectability or the time to recurrence or SPLC becoming symptomatic, assuming exponential or other time distributions, calibrating these to observed data from Aim 2. Using these calibrated distributions, the model will then be used to simulate the time to detectability and time to symptoms of recurrence/SPLC. This will allow for distinction between four surveillance scenarios: (1) Recurrence/SPLC becomes symptomatic before surveillance starts,

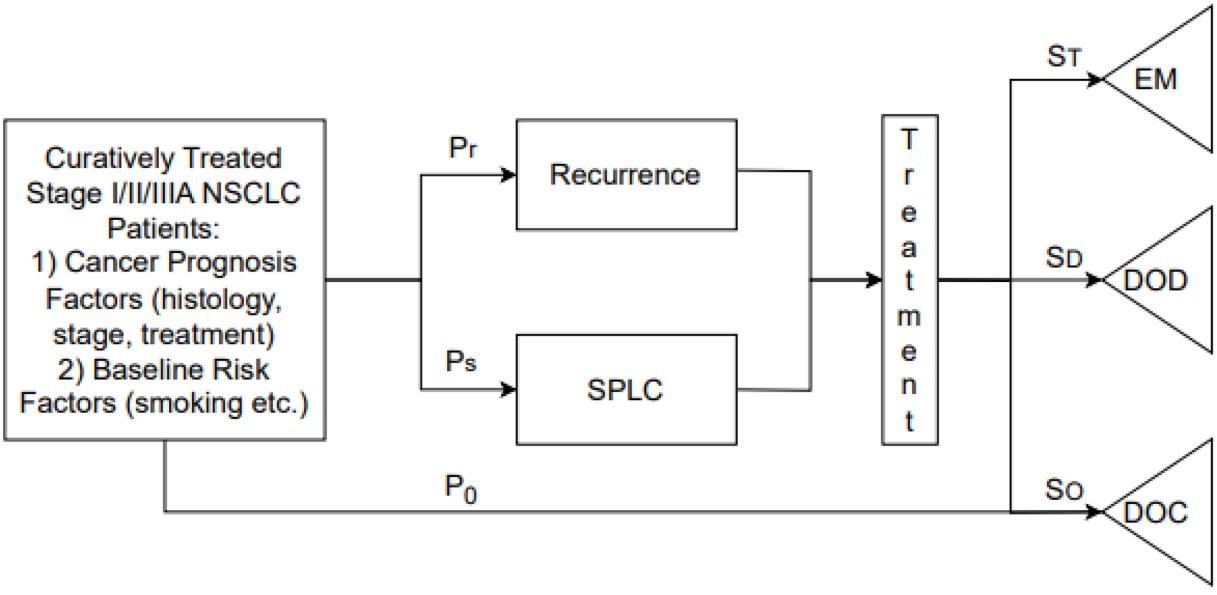


Figure 3 Microsimulation model overview. DOC, death due to other causes; DOD, death due to lung cancer; EM, excess mortality; NSCLC, non-small cell lung cancer; SPLC, second primary lung cancer.

BMJ Open: first published as 10.1136/bmjopen-2025-101324 on 28 September 2025. Downloaded from http://bmjopen.bmj.com/ on March 5, 2026 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

(2) Recurrence/SPLC becomes symptomatic during a surveillance interval, (3) Recurrence/SPLC is detected prior to symptoms at the surveillance and (4) Recurrence/SPLC becomes symptomatic and detectable after all surveillance is complete. Model inputs will include clinical prognostic factors, relevant covariates and demographic characteristics. More details on microsimulation modelling are provided in table 1. All model parameters will be estimated using data obtained from Aims 1 and 2 or calibrated using observed data from the least squares error or maximum likelihood estimation approach. Findings will then be extrapolated to the general population based on data from the surveillance, epidemiology and end results programme. The model will be used to project long-term progression and survival outcomes after lung cancer diagnosis under different surveillance schedules. This will be accomplished by varying the periodicity and timing of the CT surveillance scans and assessing alternative surveillance interventions based on input from the study team and the findings of Aims 1 and 2. Finally, a sensitivity analysis in which we will vary the uptake rates of surveillance CTs will be performed to evaluate its impact on long-term population-level outcomes.

DISCUSSION

The overarching goal of the PLuS2 study is to leverage real-world data and microsimulation modelling to bridge a significant gap in NSCLC surveillance within clinical practice and policy. While the IFCT-0302 trial did not demonstrate improved survival for CT imaging surveillance compared with chest X-ray, its limited power and lower recurrence rates than expected may have contributed to inconclusive findings.[12 53] Moreover, although randomised clinical trials are the gold standard for evaluating interventions, their strict inclusion criteria often limit generalisability.[54]

This study protocol outlines a robust observational design leveraging real-world data from a large, diverse clinical research network. By integrating microsimulation modelling with EHRs and tumour registry data, the study enables a comprehensive evaluation of imaging surveillance strategies in early-stage NSCLC. A key methodological strength is the use of NLP to extract imaging indications from unstructured clinical notes, addressing a major gap in current surveillance research.

However, several limitations should be acknowledged. The non-randomised nature of surveillance intervals may introduce confounding by indication, potentially affecting comparisons between semi-annual and annual imaging strategies. Additionally, reliance on EHR documentation may result in incomplete capture of recurrence events or SPLCs, which could impact outcome assessments.

Real-world data also present challenges such as selection bias and missing information. For example, restricting the cohort to patients who completed curative-intent therapy may limit generalisability and introduce bias. Furthermore, variables like smoking history may be incompletely documented. To mitigate these issues, we will apply multiple imputation techniques for missing data and conduct sensitivity analyses to assess the robustness of our findings under different assumptions.

ETHICS AND DISSEMINATION

This study involves human participants and was approved by the University of Florida Institutional Review Board (IRB), University of Florida IRB 01, under approval number IRB202300782. The Data Coordinating Centre (DCC) at the University of Florida will oversee establishing and maintaining a system to collect common data elements from all participating institutions within the OneFlorida+ network. Furthermore, the DCC will ensure that all data elements are standardised into a format that is compliant with the Health Insurance Portability and Accountability Act regulations.[18]

Findings from the PLuS2 study will be disseminated through peer-reviewed publications and presentations at national and international conferences. The first manuscript will describe imaging surveillance utilisation patterns and determinants among a diverse patient cohort. Subsequent publications will focus on comparative outcomes and the effectiveness of surveillance strategies, with the goal of informing future clinical guidelines and policy decisions.

Author affiliations
[1]Department of Surgery, University of Florida, Gainesville, Florida, USA
[2]Department of Cancer Centre, UF Health, Gainesville, Florida, USA
[3]Department of Epidemiology, University of Florida, Gainesville, Florida, USA
[4]Regenstrief Institute Inc, Indianapolis, Indiana, USA
[5]Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, Indiana, USA
[6]University of British Columbia, Vancouver, British Columbia, Canada
[7]Department of Epidemiology, University of Michigan, Ann Arbor, Michigan, USA
[8]Ontario Health, Toronto, Ontario, Canada
[9]Department of Health Sciences, Brock University, St. Catharines, Ontario, Canada
[10]Department of Public Health, Purdue University College of Health and Human Services, West Lafayette, Indiana, USA
[11]Department of Surgery, University of Florida College of Medicine, Gainesville, Florida, USA
[12]Rescue Lung Society, Jupiter, Florida, USA
[13]Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, Florida, USA
[14]Department of Medicine, University of Florida College of Medicine, Gainesville, Florida, USA
[15]Department of Biostatistics, University of Florida College of Medicine and College of Public Health and Health Professions, Gainesville, Florida, USA
[16]Stanford University School of Medicine, Stanford, California, USA
[17]Department of Radiology, University of Florida College of Medicine, Gainesville, Florida, USA
[18]Department of Health Systems Science, Kaiser Permanente Bernard J Tyson School of Medicine, Pasadena, California, USA

Acknowledgements We acknowledge the assistance of Shea Herlihy and Maisey Ratcliffe in providing editorial support.

Contributors DB, MKG, SDK and JJ conceptualised and led the study. All authors contributed to the study design and protocol development and have reviewed and approved the final version of the manuscript. All included authors provided written informed consent. DB is the guarantor of this work.

Funding The study is funded by the National Cancer Institute (R01CA284646).

BMJ Open: first published as 10.1136/bmjopen-2025-101324 on 28 September 2025. Downloaded from http://bmjopen.bmj.com/ on March 5, 2026 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Competing interests All authors have no competing interests to declare.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

ORCID iDs
Dejana Braithwaite http://orcid.org/0000-0001-8376-5903
Shama D Karanth http://orcid.org/0000-0001-5371-6908
Jiang Bian http://orcid.org/0000-0002-2238-5429
Meghann Wheeler http://orcid.org/0000-0003-2658-7738
Yi Guo http://orcid.org/0000-0003-0587-4105

REFERENCES

1 Sheikhbahaei S, Verde F, Hales RK, *et al*. Imaging in Therapy Response Assessment and Surveillance of Lung Cancer: Evidenced-based Review With Focus on the Utility of ^{18}F-FDG PET/CT. *Clin Lung Cancer* 2020;21:485–97.
2 Araghi M, Mannani R, Heidarnejad Maleki A, *et al*. Recent advances in non-small cell lung cancer targeted therapy; an update review. *Cancer Cell Int* 2023;23:162.
3 Institute NC. Cancer stat facts: lung and bronchus cancer. SEER. n.d. Available: https://seer.cancer.gov/statfacts/html/lungb.html
4 Siegel RL, Miller KD, Wagle NS, *et al*. Cancer statistics, 2023. *CA Cancer J Clin* 2023;73:17–48.
5 Sheikhbahaei S, Mena E, Yanamadala A, *et al*. The Value of FDG PET/CT in Treatment Response Assessment, Follow-Up, and Surveillance of Lung Cancer. *AJR Am J Roentgenol* 2017;208:420–33.
6 Ng C, Pircher A, Augustin F, *et al*. Evidence-based follow-up in lung cancer? *Memo* 2020;13:73–7.
7 Philip B, Jain A, Ramesh P, *et al*. Follow up and surveillance post lung cancer surgery: a narrative review. *Video-Assist Thorac Surg* 2023;8:7.
8 Stirling RG, Chau C, Shareh A, *et al*. Effect of Follow-Up Surveillance After Curative-Intent Treatment of NSCLC on Detection of New and Recurrent Disease, Retreatment, and Survival: A Systematic Review and Meta-Analysis. *J Thorac Oncol* 2021;16:784–97.
9 National Comprehensive Cancer Network, Inc. NCCN Clinical Practice Guidelines in Oncology: non-small cell lung cancer. Version 3. 2020.
10 Vansteenkiste J, Crinò L, Dooms C, *et al*. 2nd ESMO Consensus Conference on Lung Cancer: early-stage non-small-cell lung cancer consensus on diagnosis, treatment and follow-up. *Ann Oncol* 2014;25:1462–74.
11 Schneider BJ, Ismaila N, Aerts J, *et al*. Lung Cancer Surveillance After Definitive Curative-Intent Therapy: ASCO Guideline. *JCO* 2020;38:753–66.
12 Westeel V, Foucher P, Scherpereel A, *et al*. Chest CT scan plus x-ray versus chest x-ray for the follow-up of completely resected non-small-cell lung cancer (IFCT-0302): a multicentre, open-label, randomised, phase 3 trial. *Lancet Oncol* 2022;23:1180–8.
13 NCCN Clinical Practice Guidelines in Oncology, Non-Small Cell Lung Cancer. Version 3. 2019.
14 Jaklitsch MT, Jacobson FL, Austin JHM, *et al*. The American Association for Thoracic Surgery guidelines for lung cancer screening using low-dose computed tomography scans for lung cancer survivors and other high-risk groups. *J Thorac Cardiovasc Surg* 2012;144:33–8.
15 Colt HG, Murgu SD, Korst RJ, *et al*. Follow-up and surveillance of the patient with lung cancer after curative-intent therapy: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest* 2013;143:e437S–e454S.
16 McMurry TL, Stukenborg GJ, Kessler LG, *et al*. More Frequent Surveillance Following Lung Cancer Resection Is Not Associated With Improved Survival: A Nationally Representative Cohort Study. *Ann Surg* 2018;268:632–9.
17 Mettler FA Jr, Bhargavan M, Faulkner K, *et al*. Radiologic and nuclear medicine studies in the United States and worldwide: frequency, radiation dose, and comparison with other radiation sources--1950-2007. *Radiology* 2009;253:520–31.
18 Hogan WR, Shenkman EA, Robinson T, *et al*. The OneFlorida Data Trust: a centralized, translational research data infrastructure of statewide scope. *J Am Med Inform Assoc* 2022;29:686–93.
19 Braithwaite D, Karanth SD, Slatore CG, *et al*. Personalised Lung Cancer Screening (PLuS) study to assess the importance of coexisting chronic conditions to clinical practice and policy: protocol for a multicentre observational study. *BMJ Open* 2022;12:e064142.
20 Yang X, Bian J, Gong Y, *et al*. MADEx: A System for Detecting Medications, Adverse Drug Events, and Their Relations from Clinical Notes. *Drug Saf* 2019;42:123–33.
21 Wu Y, Yang X, Bian J, *et al*. Combine Factual Medical Knowledge and Distributed Word Representation to Improve Clinical Named Entity Recognition. *AMIA Annu Symp Proc* 2018;2018:1110–7.
22 Sauer BC, Jones BE, Globe G, *et al*. Performance of a Natural Language Processing (NLP) Tool to Extract Pulmonary Function Test (PFT) Reports from Structured and Semistructured Veteran Affairs (VA) Data. *EGEMS (Wash DC)* 2016;4:1217.
23 Hassett MJ, Uno H, Cronin AM, *et al*. Detecting Lung and Colorectal Cancer Recurrence Using Structured Clinical/Administrative Data to Enable Outcomes Research and Population Health Management. *Med Care* 2017;55:e88–98.
24 Hugging Face Meta. Available: https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct [Accessed 15 Aug 2025].
25 Varoquaux G, Colliot O. Evaluating machine learning models and their diagnostic value. In: Colliot O, ed. *Machine Learning for Brain Disorders*. Springer US, 2023: 601–30.
26 Erb CT, Su KW, Soulos PR, *et al*. Surveillance Practice Patterns after Curative Intent Therapy for Stage I Non-Small-Cell Lung Cancer in the Medicare Population. *Lung Cancer (Auckl)* 2016;99:200–7.
27 Tammemägi MC, Ten Haaf K, Toumazis I, *et al*. Development and Validation of a Multivariable Lung Cancer Risk Prediction Model That Includes Low-Dose Computed Tomography Screening Results: A Secondary Analysis of Data From the National Lung Screening Trial. *JAMA Netw Open* 2019;2:e190204.
28 Gould MK, Munoz-Plaza CE, Hahn EE, *et al*. Comorbidity Profiles and Their Effect on Treatment Selection and Survival among Patients with Lung Cancer. *Annals ATS* 2017;14:1571–80.
29 Braithwaite D, Gould MK. Is Lung Cancer Screening Reaching the People Who Are Most Likely to Benefit? *JAMA Netw Open* 2022;5:e2243171.
30 Ettinger DS, Wood DE, Aisner DL, *et al*. Non-Small Cell Lung Cancer, Version 3.2022, NCCN Clinical Practice Guidelines in Oncology. *J Natl Compr Canc Netw* 2022;20:497–530.
31 Jeon J, Holford TR, Levy DT, *et al*. Smoking and Lung Cancer Mortality in the United States From 2015 to 2065: A Comparative Modeling Approach. *Ann Intern Med* 2018;169:684–93.
32 Shivade C, Raghavan P, Fosler-Lussier E, *et al*. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc* 2014;21:221–30.
33 Gould MK, Sakoda LC, Ritzwoller DP, *et al*. Monitoring Lung Cancer Screening Use and Outcomes at Four Cancer Research Network Sites. *Annals ATS* 2017;14:1827–35.
34 Lamprecht B, Schirnhofer L, Kaiser B, *et al*. Subjects with Discordant Airways Obstruction: Lost between Spirometric Definitions of COPD. *Pulm Med* 2011;2011:780215.
35 Mannino DM, Buist AS. Global burden of COPD: risk factors, prevalence, and future trends. *Lancet* 2007;370:765–73.
36 Rabe KF, Hurd S, Anzueto A, *et al*. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *Am J Respir Crit Care Med* 2007;176:532–55.
37 Segal JB, Chang H-Y, Du Y, *et al*. Development of a Claims-based Frailty Indicator Anchored to a Well-established Frailty Phenotype. *Med Care* 2017;55:716–22.
38 Segal JB, Huang J, Roth DL, *et al*. External validation of the claims-based frailty index in the national health and aging trends study cohort. *Am J Epidemiol* 2017;186:745–7.
39 Newton KM, Peissig PL, Kho AN, *et al*. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. *J Am Med Inform Assoc* 2013;20:e147–54.
40 Yang X, Yang H, Lyu T, *et al*. A natural language processing tool to extract quantitative smoking status from clinical narratives. IEEE International Conference on Healthcare Informatics; 2020

BMJ Open: first published as 10.1136/bmjopen-2025-101324 on 28 September 2025. Downloaded from http://bmjopen.bmj.com/ on March 5, 2026 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

41 Krieger N. Overcoming the absence of socioeconomic data in medical records: validation and application of a census-based methodology. *Am J Public Health* 1992;82:703–10.
42 Backhus LM, Farjah F, Liang C-K, *et al*. Imaging surveillance and survival for surgically resected non-small-cell lung cancer. *J Surg Res* 2016;200:171–6.
43 Malhotra J, Rotter D, Jabbour SK, *et al*. Receipt of recommended surveillance with imaging in elderly survivors of early stage non-small cell lung cancer. *Lung Cancer (Auckl)* 2018;125:205–11.
44 Karzijn R, Alberts L, Kelder JC, *et al*. Post-treatment Surveillance for Stage I and II Non-small Cell Lung Cancer: Impact on Clinical Outcome. *Anticancer Res* 2016;36:5413–8.
45 Westeel V, Choma D, Clément F, *et al*. Relevance of an intensive postoperative follow-up after surgery for non-small cell lung cancer. *Ann Thorac Surg* 2000;70:1185–90.
46 Ozechowski TJ, Turner CW, Hops H. Mixed-effects logistic regression for estimating transitional probabilities in sequentially coded observational data. *Psychol Methods* 2007;12:317–35.
47 Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. *J Am Stat Assoc* 1958;53:457–81.
48 Zhang X, Zhang M-J, Fine J. A proportional hazards regression model for the subdistribution with right-censored and left-truncated competing risks data. *Stat Med* 2011;30:1933–51.
49 Royston P, Parmar MKB. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Stat Med* 2002;21:2175–97.
50 Therneau TM, Grambsch PM. Modeling survival data: extending the cox model. In: *Statistics for biology and health*. New York, NY: Springer, 2000. Available: http://link.springer.com/10.1007/978-1-4757-3294-8
51 Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology (Sunnyvale)* 2000;11:550–60.
52 Marshall SW. Power for tests of interaction: effect of raising the Type I error rate. *Epidemiol Perspect Innov* 2007;4:4.
53 Westeel V, Barlesi F, Foucher P, *et al*. Results of the phase III IFCT-0302 trial assessing minimal versus CT-scan-based follow-up for completely resected non-small cell lung cancer (NSCLC). *Ann Oncol* 2017;28:v452.
54 He Z, Tang X, Yang X, *et al*. Clinical Trial Generalizability Assessment in the Big Data Era: A Review. *Clin Transl Sci* 2020;13:675–84.

BMJ Open: first published as 10.1136/bmjopen-2025-101324 on 28 September 2025. Downloaded from http://bmjopen.bmj.com/ on March 5, 2026 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.