1. Title Page

Title	CARE Initiative Pilot Emulation Study: Emulation of a comparative effectiveness study of pembrolizumab and chemotherapy vs. chemotherapy for the first-line treatment of metastatic non-small cell lung cancer
Objective	In alignment with the KEYNOTE-189 randomized controlled trial, the objective is to assess the comparative effectiveness of pembrolizumab, platinum therapy, and pemetrexed vs. platinum therapy and pemetrexed in patients with metastatic non-small cell lung cancer and without EGFR/ALK mutations, and to calibrate results against the KEYNOTE-189 trial.
Protocol version	1.2
Last update date	9/8/2023
Protocol authors	David Merola, PharmD, PhD Aetion, Inc 50 Congress St. Suite 1025 Boston, MA 02109 Jennifer Rider, ScD, MPH Aetion, Inc 50 Congress St. Suite 1025 Boston, MA 02109 Ulka Campbell, PhD Aetion, Inc 5 Penn Plaza 7th floor New York, NY 10001

	David Lenis, PhD Aetion, Inc 5 Penn Plaza 7th floor New York, NY 10001
Sponsor	This work was collectively sponsored by Abbvie, Amgen, AstraZeneca, Bayer, Gilead, Janssen, and Pfizer
Conflict of interest	DM, JR, UC, and DL are shareholders and employees of Aetion, Inc.
Study ID (clinicaltrials.gov)	NCT05908799

Table of contents

2. Abstract

The Coalition to Advance Real-World Evidence through Randomized Controlled Trial Emulation (CARE) Initiative is a program designed to build an empirical evidence base for the use of real-world data in clinical and regulatory decision-making. Using results from randomized clinical trials as a benchmark for causal effect estimates, a series of randomized controlled trial emulations will be conducted using real-world data (RWD) to better understand when observational methods and secondary data collected from routine healthcare practice can provide reliable conclusions on drug effectiveness. This protocol describes the first of several studies and seeks to emulate the KEYNOTE-189 randomized controlled trial.[1] KEYNOTE-189 aimed to assess the comparative effectiveness of pembrolizumab, platinum therapy, and pemetrexed vs. platinum therapy and pemetrexed in patients with metastatic non-small cell lung cancer (NSCLC) and without epidermal growth factor receptor (EGFR)/anaplastic lymphoma kinase (ALK) gene mutations. The KEYNOTE-189 trial provided primary evidence of effectiveness to support regulatory approval of pembrolizumab for this patient population.[2] This emulation study will use data derived from an electronic health record database linked with a cancer registry.

3. Amendments and updates

Version date	Version number	Section of protocol	Amendment or update	Reason
8-10-2023	V1.1	Study Design, Variables, Data Analysis	The treatment contrast of the per-protocol analyses was corrected. The term "cycles" was incorporated into the treatment regimen specification. perf_status variable was corrected to include a time window for the per-protocol analysis	The comparator group regimen incorrectly stated that chemotherapy was to be used as maintenance therapy in portions of the protocol. The description of the treatment contrast was updated to state "cycles" address ambiguity.
9-8-2023	V1.2	Data Analysis	Time-varying covariates were removed from the marginal structural model specified for the per-protocol analysis Cluster-robust variance estimators were added as a possible	Correction of a typo and permitting more flexibility in methods due to high dimensionality of data.

			approach to deriving interval estimates.	

4. Research question and objectives

A summary comparison of the key elements of study design and analysis plan in the KEYNOTE-189 trial vs. this study is available in the appendix.

Table 1. Research question and objective

Objective:	In alignment with the KEYNOTE-189 trial, the objective is to compare the overall survival of patients receiving pembrolizumab, platinum therapy, and pemetrexed vs. platinum therapy and pemetrexed for the treatment of metastatic NSCLC without EGFR/ALK mutations, and to calibrate results against the KEYNOTE-189 trial.
Causal effect of interest:	Average treatment effect
Hypothesis:	Within the population of patients with the tumor type of interest who receive pemetrexed and a platinum therapy, those additionally treated with pembrolizumab will have better survival than those not treated with pembrolizumab. The results will be congruent with the KEYNOTE-189 trial.
Population (mention key inclusion-exclusion criteria):	Patients with metastatic NSCLC who have not previously received systemic therapy for advanced disease and do not have EGFR or ALK mutations.
Exposure:	Initiation of pembrolizumab, pemetrexed, and platinum therapy (cisplatin or carboplatin)
Comparator:	Initiation of pemetrexed and platinum therapy (cisplatin or carboplatin)
Outcome:	Overall survival, defined as the time from treatment initiation to death due to any cause.
Time (when follow-up begins and ends):	Follow-up initiated on the day after exposure regimen ascertainment window and stopped at earliest of: death, loss to follow-up or end of available data
Setting:	Outpatient and inpatient routine care setting
Main measures of effect:	Hazard ratio (intention-to-treat[a]) and 12-month survival proportion

[a] Intention-to-treat is defined by treatment initiation (i.e., agnostic to adherence)

5. Research methods

5.1. Study design

Overview of key design elements of the KEYNOTE-189 trial:

The KEYNOTE-189 trial was a randomized controlled, double-blind trial comparing the efficacy of pembrolizumab vs. placebo when taken in combination with platinum therapy (investigator's choice of cisplatin or carboplatin) and pemetrexed for the first-line treatment of non-squamous non-small cell lung cancer without EGFR or ALK mutations. Patients with *active* brain metastases, on chronic systemic steroids or other immunosuppressive agents, and a history of other primary malignancies were excluded. Follow-up was initiated at treatment randomization (baseline) and stopped at the earliest of death, unacceptable toxic effects, investigator decision, or patient withdrawal of consent. The primary outcome was overall survival, defined as the time from randomization to death due to any cause. An intention-to-treat analysis was performed to estimate the relative hazard of mortality and 12-month survival probability using a Cox proportional hazards model and Kaplan-Meier estimator, respectively. More details on the design elements of the KEYNOTE-189 trial and how these compare to the present real-world evidence emulation study are available in the appendix along with study design diagrams for each of the studies' designs.

Overview of key design elements of the real-world evidence emulation study:

Primary (intention-to-treat) analysis: This is a retrospective cohort study comparing patients with incident use of pembrolizumab, pemetrexed, and platinum therapy vs. pemetrexed and platinum therapy following the first indication of metastatic disease in their electronic health record. The study design is illustrated in Figure 1-A. Patients will be required to have a diagnosis of non-small cell lung cancer and have no evidence of EGFR or ALK mutations prior to treatment initiation. Patients with evidence of brain metastases in the 2 weeks prior to treatment initiation, records in the 90 days prior to treatment initiation indicating use of systemic steroids or other immunosuppressive agents, or a history of other primary malignancies will be excluded. Exposure will be ascertained within a 14-day window of time (i.e., "exposure regimen ascertainment window"), beginning on the day of first record indicating use of a study drug (pembrolizumab, pemetrexed, or platinum therapy), and ending 14-days later ("time zero"). Patients' exposure will be classified based on the receipt of all study drugs in each regimen within the 14-day exposure regimen ascertainment window, which will be incident with respect to metastatic disease; study drug initiation may occur on or after date of metastatic disease. Patients will be required to be alive and at risk of the event throughout the exposure ascertainment window. Confounders will be assessed during the baseline period, defined as all time prior to and including time zero. Follow-up will begin on the day after the 14-day exposure regimen ascertainment window and proceed until the outcome event (death), loss to follow-up (>90-day gap in health record activity), 640 days of follow-up, or administrative end of available data occurs (i.e., last date of record in the dataset). Time-fixed inverse probability weights will be used to adjust for potential confounding. The effect of treatment initiation on the outcome of

interest will be estimated using an intention-to-treat estimator. This primary analysis will be agnostic to treatment adherence and will not account for potential selection bias induced by informative censoring or time-varying confounding. Rationale for this design is described in the next section.

Secondary (per-protocol) analysis: A secondary analysis will be conducted that will employ the same eligibility criteria as the primary analysis, but with a modified exposure definition and analysis plan. Two static treatment regimens will be compared in the per-protocol population as follows:

1. Exposure: First-line, concurrent use of pembrolizumab, platinum therapy (cisplatin or carboplatin), and pemetrexed every 21 days for 4 cycles, followed by concurrent use of pembrolizumab and pemetrexed every 21 days thereafter until intolerance/toxicity, progression, or mortality.
2. Comparator: First-line, concurrent use of platinum therapy (cisplatin or carboplatin) and pemetrexed every 21 days for 4 cycles, followed by pemetrexed every 21 days thereafter until intolerance/toxicity, progression, or mortality. Upon failure of the original first-line regimen, may use pembrolizumab monotherapy every 21 days.

The study design is illustrated in Figure 1-B. Given that some components of the exposure and comparator treatment regimens are overlapping (i.e., platinum therapy and pemetrexed), a "clone and censor" approach[3] described by Hernan and Robins will be undertaken. In this approach, the analytic cohort will be cloned to create 2 exact copies of every patient, where each "clone" is assigned to one of two static treatment regimens (defined in variables section) and censored upon deviation from their assigned regimen. The justification for using this method is that if a patient experiences an event during follow-up but has a treatment pattern that complies with both exposure strategies of interest (e.g., taking platinum therapy and pemetrexed between days 1-10 of follow-up), the event will be attributed to both groups rather than one group differentially. In alignment with the KEYNOTE-189 trial, cross-over from the non-pembrolizumab group to pembrolizumab monotherapy will be permitted. For this secondary analysis, patients will be required to have adhered to static treatment regimens (i.e., treatment regimens dependent only on treatment history, but not upon other covariates) during follow-up that are detailed in the variables section of this protocol. Time-fixed confounders (assessed at baseline) and time-varying confounders (assessed at each interval of follow-up) will be controlled for using inverse probability of treatment weighting. Potential informative censoring induced by the artificial censoring rules imposed in this analysis will be accounted for using inverse probability of censor weighting. The effect of the treatment on the outcome will be estimated using a per-protocol estimator, where follow-up will begin on the day after treatment initiation and proceed until an outcome event (death), treatment cessation (deviation from regimens described in the variables section of this protocol), loss to follow-up (>90-day gap in health record activity), 640 days of follow-up, or end of available data occurs (i.e., last date of record in the dataset). Rationale for this design is described in the next section.

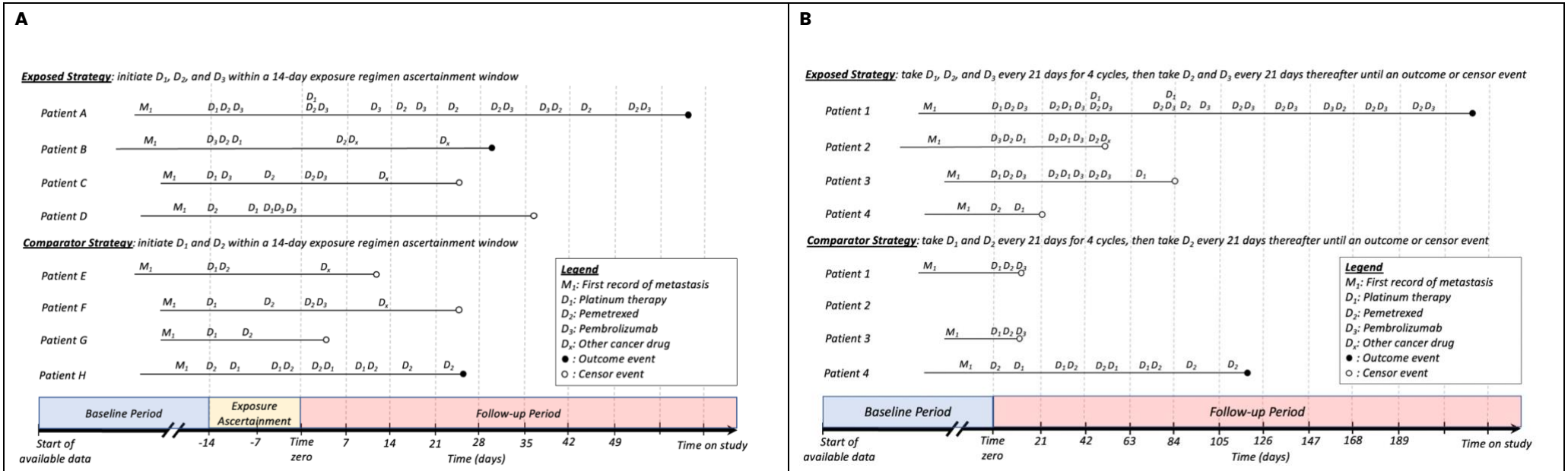

Figure 1. Example patient timelines for the primary (intention-to-treat) and secondary (per-protocol) analysis. **Panel A** Primary (intention-to-treat) analysis. All patients shown are distinct and are classified on the basis of their treatment initiation within a 14-day exposure regimen ascertainment window that is incident with respect to each patient's first record of metastasis. Notably, this analysis is agnostic to adherence and patients are followed up from the end of the exposure regimen ascertainment window until occurrence of an outcome or censor event. Additionally, multiple records for each study drug may occur within the ascertainment window, as observed in Patients D and H. **Panel B** Secondary (per-protocol) analysis. Patients 1-4 are "cloned" and assigned to each treatment strategy upon initiation of a study drug. Patient 1 adheres perfectly to the exposure treatment strategy and experiences the outcome event; therefore, all person-time observations are contributed to the exposed group. Patient 1 also contributes a small amount of person-time to the comparator group, as the patient's treatment history is aligned with the comparator treatment strategy for some follow-up time prior to initiation of D_3. Patient 2 in the "exposed" group is censored upon initiation of a cancer treatment that is not a part of her assigned regimen (D_x) and does not contribute person-time to the comparator group since D_3 is not a component of the comparator regimen. Patient 3 in the "exposed" group is censored at the end of the 21-day interval because that is the point at which the patient's treatment history is no longer compliant with the exposure strategy. Patient 4 in the "comparator" group has a treatment history that is compatible with the comparator strategy and experiences an outcome event but is censored in the "exposed" group due to a lack of "compliance" with the exposed treatment strategy. Notably, if an event occurs during an interval of follow-up time that is compatible with both treatment strategies, the outcome is attributed to both groups.

Rationale for study design choices:

Study design decisions reflect the bias-variance trade-off that is inherent to nearly all epidemiologic investigations, as well as the goal to emulate the KEYNOTE-189 trial results. The primary analysis was constructed to maximize the number of eligible patients for analysis and resultant statistical power, while the secondary analysis was designed to mitigate bias as much as possible at the expense of a greater variance. Both analyses use weighting to obtain marginal treatment effects and censor patients administratively at 21 months (640 days) of follow-up, which may align more with the randomized controlled trial result than conditional treatment effects. The 14-day exposure regimen ascertainment window in the primary analysis was chosen to allow for some variability in treatment patterns that are likely to occur in routine care. Ascertaining exposure in shorter intervals (e.g., on a single day) may be too restrictive and result in few eligible patients, while longer time periods may result in treatment patterns in the study population that are vastly different from the trial. In the per-protocol analysis, a 21-day treatment administration was chosen to align with the dosing schedule prescribed in the KEYNOTE-189 trial.

In the primary analysis, the use of inverse probability weighting and time-fixed treatment strategies may allow certain patients to contribute information that would have otherwise been discarded during matching procedures or lack of treatment adherence. Furthermore, treatment adherence may be relatively high among patients with lung cancer due to the nature of the disease. Adjustment for informative censoring can introduce additional variability (and reduce precision of) effect estimators, due to the presence of extreme weights. Although the analytic approach in the primary analysis more closely resembles that of the KEYNOTE-189 trial, permitting adjustment for informative censoring and assessing the per-protocol effect (as will be done in the secondary analysis) may give rise to results that more closely align with randomized controlled trials, where relatively high levels of adherence are often observed. At 21 months of follow-up, 33.8% and 17.8% of patients were on the assigned therapy in the pembrolizumab-containing and placebo-containing group of the KEYNOTE-189 trial, respectively.[1] Lastly, time-varying confounding is likely to be present in the real-world oncology population, where clinical condition and performance status can change dramatically over follow-up.

5.2. Setting

Context and rationale for definition of time 0 (and other primary time anchors) for entry to the study population

The criteria defining time zero for the primary (intention-to-treat) and secondary (per-protocol) populations are specified below in Table 2 and illustrated in Figure 1 in the appendix. In the primary population, time zero will be defined by the end of the 14-day exposure regimen ascertainment window (incident with respect to the first record of metastasis) described in the Study Design section. In the secondary population, time zero will be defined by incident receipt of a study drug of interest (also incident with respect to an indication of metastasis in the data source). Time zero will be restricted to dates on or after October 7, 1995 to reflect the timing of availability of published evidence supporting the use of platinum-based therapies in the first-line advanced disease setting.[4–6] Permitting "historical" controls will allow for a greater study sample size and precision of estimates, under the assumption that decisions driving treatment initiation are constant over time. A "washout" window, in which a patient has no records indicating study drug use between

their indicator of metastatic disease to initiation of first study drug of interest, will be applied to the primary and secondary populations to ensure that the exposure regimen identified is each patient's first-line treatment for metastatic disease.

Table 2. Operational Definition of Time 0 (index date) and other primary time anchors

Study population name(s)	Time Anchor Description	Number of entries	Type of entry	Washout window	Code Type[a]	Incident with respect to…
Patients initiating first-line treatment for metastatic disease (primary, intention-to-treat population)	For the exposed and comparator groups, time zero will be defined as the last day of the 14-day exposure regimen ascertainment window. Exposure regimen ascertainment is described in the variables section of this protocol.	Single entry	Incident occurrence of multi-drug exposure sequence	[first date of metastatic disease indicator[b], first date of study treatment of interest administration]	NDC, HCPCS/CPT, RxNorm	Metastatic disease indicator[b]
Patients initiating first-line treatment for metastatic disease (secondary, per-protocol population)	For the exposed and comparator groups, time zero will be defined as the first day after incident occurrence of any study drug (i.e., platinum therapy, pemetrexed, or pembrolizumab).	Single entry	Incident study drug occurrence	[first date of metastatic disease indicator[b], first date of study treatment of interest administration]	NDC, HCPCS/CPT, RxNorm	Metastatic disease indicator[b]

Note: Some study drugs are used in the adjuvant and neoadjuvant settings, as well as the metastatic disease setting. To distinguish the indication for these study drugs, treatments occurring after the first observed metastatic disease indicator will be deemed as treatments for metastatic disease.

[a] See appendix for listing of clinical codes for each study parameter

[b] Indicators of metastatic disease include records indicating stage 4 disease, indicators of metastasis derived from the data vendor, or International Classification of Diseases for Oncology (ICD-O) diagnostic codes indicating metastasis. Information on metastatic disease is drawn from a linked tumor registry and tumor morphology descriptors contained within the health record.

***Context and rationale for study inclusion criteria*:**

Operational definitions for study inclusion criteria are shown in Table 3, and were selected to proxy those of the KEYNOTE-189 trial as closely as possible.[1] Numerous variables in the data source were used to define each criterion in order to harness all available information. For example, treatments were selected using National Drug Code (NDC), Healthcare Common Procedure Coding System (HCPCS)/Current Procedural Terminology (CPT), and RxNorm codes. "Observability" requirements (e.g., continuous activity in health record database) will not be imposed during the baseline period to maximize study sample size and statistical precision of outcome estimates.

Platinum therapy is used in both the adjuvant/neoadjuvant setting, as well as the metastatic setting. Therefore, in defining "No prior systemic treatment for metastatic disease," the date range for assessment (between stage IV diagnosis and the index treatment) was chosen to differentiate between therapies used perioperatively and therapies used in the metastatic setting. We assume that patients are being seen within-network and all relevant healthcare encounters are captured during the baseline period.

Table 3. Operational Definitions of Inclusion Criteria

Criterion	Description	Variable (Data Table Name(s))	Order of application	Assessment window[a]	Code Type(s)[b]
Lung cancer	Diagnosis code indicating cancer of the lung or bronchus	code (Diagnosis) tumor_site_code (Tumor, Tumor Properties)	1	[-all available data, 0]	ICD
Stage IV disease	Evidence of metastasis	metastatic (Tumor) stage_code (Tumor) morphology_code (Tumor, Tumor Properties) code (Diagnosis)	2	[-all available data, 0]	ICD, derived
Non-small cell histology	Lack of evidence of other lung cancer types that are NOT non-small cell lung cancer (e.g., mesothelioma, small cell carcinoma, squamous cell carcinoma, etc)	morphology_code (Tumor) tumor_site_code (Tumor)	3	[-all available data, 0]	ICD
No epidermal growth factor receptor (EGFR) or	Does not have evidence of EGFR or ALK mutations.	code (Genomic)	4	[-all available data, 0]	HGVS

anaplastic lymphoma kinase (ALK) mutations					
No prior systemic treatment for metastatic disease	Patient has no evidence of use of guideline-recommended systemic cancer therapy[c] for metastatic NSCLC	code (Medication Drug, Medication Ingredient, Procedure)	5	ITT analysis: [-Stage IV disease, -15] PP analysis: [-Stage IV disease, -1]	CPT/HCPCS, NDC, RxNorm

Note: Criteria in this table apply to both, the primary and secondary analyses.
CPT = Current Procedural Terminology, HCPCS = Healthcare Common Procedure Coding System, HGVS = Human Genome Variation Society, ICD = International Classification of Diseases, LOINC = Logical Observation Identifiers Names and Codes, NSCLC = non-small cell lung cancer, RxNorm = a standardized nomenclature for clinical drugs
[a] Assessment times shown are inclusive.
[b] See appendix for listing of clinical codes for each study parameter.
[c] Systemic cancer therapies were identified from the National Comprehensive Cancer Network Clinical Practice Guidelines for non-squamous non-small cell lung cancer and include: pembrolizumab, carboplatin, cisplatin, pemetrexed, atezolizumab, paclitaxel, nivolumab, ipilimumab, cemiplimab, tremelimumab, durvalumab, docetaxel, etoposide, gemcitabine, and vinorelbine.

Context and rationale for study exclusion criteria

As with the inclusion criteria, operational definitions for study exclusion criteria were selected to mimic those of the KEYNOTE-189 trial (Table 4).[1] Numerous variables in the data source were used to define each criterion in order to harness all available information.

Table 4. Operational Definitions of Exclusion Criteria

Criterion	Description	Variable (Data Table Name(s))	Order of application	Assessment window[a]	Code Type(s)[b]	Applied to study populations:
Squamous cell tumor morphology	The patient has evidence of squamous cell lung cancer morphology	morphology_code (Tumor)	1	[-all available data, 0]	ICD, derived	Primary and secondary

Biologic cancer therapy for metastatic disease	Evidence of biologic therapy[c] for metastatic NSCLC (including ALK- and EGFR-directed therapies)	code (Medication Drug, Medication Ingredient, Procedure)	2	ITT analysis: [-all available data, 0]	CPT/HCPCS, NDC, RxNorm	Primary and secondary
Prior non-lung malignancy	Has evidence of primary malignancies beyond the lung tissue, except basal cell carcinoma of the skin, superficial bladder cancer, squamous cell carcinoma of the skin, in situ cervical cancer, or other in situ cancers	tumor_site_code (Tumor)	3	[-all available data, 0]	ICD	Primary and secondary
CNS metastases	Has no diagnosis codes of central nervous system (CNS) metastases and/or carcinomatous meningitis in the 2 weeks prior to treatment initiation	code (Diagnosis) tumor_site_code (Tumor)	4	ITT analysis: [-28, -15] PP analysis: [-14, -1]	ICD	Primary and secondary
Treatment with disease modifying agents, corticosteroids, or immunosuppressive drugs	Has a record of systemic treatments[d] for autoimmune disease in the 2 years prior to the index date	code (Medication Drug, Medication Ingredient, Procedure)	5	[-730, 0]	CPT/HCPCS NDC, RxNorm	Primary and secondary
Chronic corticosteroids	Has at least two records 60 days apart indicating use of corticosteroid agents[e] in the 3 months prior to the index date	code (Medication Drug, Medication Ingredient, Procedure)	6	[-90, 0]	CPT/HCPCS NDC, RxNorm	Primary and secondary
Prior treatment with PD-L1 or PD-L2 agent or antibody targeting other immuno-regulatory receptors or mechanisms.	Has evidence of use of one of the following agents prior to the index date: pembrolizumab, nivolumab, cemiplimab, atezolizumab, durvalumab	code (Medication Drug, Medication Ingredient, Procedure)	7	ITT analysis: [-all available data, -15] PP analysis: [-all available data, -1]	CPT/HCPCS NDC, RxNorm	Primary and secondary

Mortality at baseline	Patients experiencing mortality during the exposure regimen ascertainment window or during the assessed baseline period. Note: Both, death during the baseline period and death during the exposure regimen ascertainment window will be evaluated separately. It is possible for a patient to have a recorded mortality event during the baseline period because mortality is assumed to occur on the first day of the month and year of death recorded in the database. Exact dates of mortality are not recorded in the data to preserve patient privacy.	month_year_death (Patient Demographic)	8	ITT analysis: [-14, 0] PP analysis: [-all available data, 0]	Derived	Primary and secondary

CPT = Current Procedural Terminology, HCPCS = Healthcare Common Procedure Coding System, HGVS = Human Genome Variation Society, ICD = International Classification of Diseases, LOINC = Logical Observation Identifiers Names and Codes, NSCLC = non-small cell lung cancer, RxNorm = a standardized nomenclature for clinical drugs

[a] Assessment times shown are inclusive.

[b] See appendix for listing of clinical codes for each study parameter.

[c] Biologic therapies, including those targeting EGFR and ALK mutations, identified from the National Comprehensive Cancer Network Clinical Practice Guidelines for non-squamous non-small cell lung cancer: erlotinib, afatinib, gefitinib, dacomitinib, crizotibib, and cetuximab.

[d] Systemic treatment(s) for autoimmune disease were defined based on investigators' substantive knowledge as follows: methotrexate, sulfasalazine, hydroxychloroquine, leflunomide, azathioprine, etanercept, adalimumab, infliximab, certolizumab, golimumab, anakinra, tocilizumab, sarilumab, abatacept, rituximab, tofacitinib, barivitinib, upadacitinib, mycophenolate.

[e] Corticosteroid drugs were defined using investigators' substantive knowledge and include the following: prednisone, prednisolone, methylprednisolone, dexamethasone, hydrocortisone.

5.3. Variables

Context and rationale for exposure(s) of interest

Primary analysis (intention-to-treat) exposure definition: The two treatment strategies that will be compared in the intention-to-treat population will be as follows:

1. Exposure: Initiation of first line pembrolizumab, platinum therapy, and pemetrexed within a 14-day exposure assessment window
2. Comparator: Initiation of first line platinum therapy and pemetrexed within a 14-day exposure assessment window

A patient will be classified as having "initiated" each multi-drug regimen if they had an incident order, dispensation, and/or administration for all component study drugs of that regimen within a 14-day exposure assessment window. For both treatment groups, the exposure regimen ascertainment window will be incident with respect to the first record of metastatic disease to distinguish between therapies that are being used in the neoadjuvant or adjuvant setting versus those being used as first-line therapy for metastatic disease. All available information in the data source will be used to ascertain exposure status, including procedure codes (i.e., HCPCS/CPT "J" codes), as well as NDC and RxNorm codes available from the medication-related data tables. This exposure definition is illustrated in Figure 1-A in the Study Design section.

Secondary analysis (per-protocol) exposure definition: The two static treatment regimens that will be compared in the per-protocol population are as follows:

1. Exposure: First-line, concurrent use of pembrolizumab, platinum therapy (cisplatin or carboplatin), and pemetrexed every 21 days for 4 cycles, followed by concurrent use of pembrolizumab and pemetrexed every 21 days thereafter until intolerance/toxicity, progression, or mortality.
2. Comparator: First-line, concurrent use of platinum therapy (cisplatin or carboplatin) and pemetrexed every 21 days for 4 cycles, followed by pemetrexed every 21 days thereafter until intolerance/toxicity, progression, or mortality. Upon failure of the original first-line regimen, may use pembrolizumab monotherapy every 21 days.

These real-world data treatment regimen definitions were chosen as a compromise between 1) the heterogeneity of adherence patterns observed in routine clinical practice, and 2) the goal of emulating the KEYNOTE-189 trial as closely as possible. By permitting patients to take treatments any time within a 21-day gap more patients may be eligible for classification into an exposure of interest without compromising the general adherence pattern imposed in the KEYNOTE-189 trial. Discontinuation of treatments or treatment switch will be assumed to be due to intolerance/toxicity or progression events.

A patient will be classified as having "initiated" each multi-drug regimen if they had an incident order, dispensation, and/or administration for any component study drugs of that regimen. As with the intention-to-treat exposure definition, the "incident" order, dispensation, and/or administration will be with respect to the first record of metastatic disease. Also similar to the primary analysis, all available information in the data source will be used to

determine exposure status, including procedure codes (i.e., HCPCS/CPT "J" codes), as well as NDC and RxNorm codes available from the medication-related data tables. Patients with evidence of non-study drugs between the initial metastatic disease indicator and incident study drug use will be excluded from the analysis by design to ensure the study population being captured is receiving first-line treatment. Refer to Figure 1-B in the Study Design section for several illustrations of the static treatment regimens.

Table 5. Operational Definitions of Exposure

Exposure group name(s)	Details	Exposure Regimen Ascertainment Window[a]	Variable (Data Table Name(s))	Code Type[b]	Applied to study populations:	Incident with respect to…
Exposure group (intention-to-treat analysis): Pembrolizumab and chemotherapy	A time-fixed treatment, defined as follows: 1) Incident use of pembrolizumab, platinum therapy (cisplatin or carboplatin), or pemetrexed 2) Occurrence of remaining drugs not captured in 1) within a 14-day window following 1)	[-14,0]	code (Medication Drug, Medication Ingredient, Procedure)	CPT/HCPCS NDC, RxNorm	Primary	Metastatic disease indicator[c]
Comparator group (intention-to-treat analysis): Chemotherapy	1) Incident use of platinum therapy (cisplatin or carboplatin) or pemetrexed 2) Occurrence of remaining drug not captured in 1) within a 14-day window following 1)	[-14,0]	code (Medication Drug, Medication Ingredient, Procedure)	CPT/HCPCS NDC, RxNorm	Primary	Metastatic disease indicator[c]
Exposure group (per-protocol analysis): Pembrolizumab and chemotherapy	A static treatment regime, defined as follows: Concurrent use of pembrolizumab, pemetrexed, and platinum therapy (cisplatin or carboplatin) every 21 days for 4 cycles, followed by concurrent use of pembrolizumab and pemetrexed every 21 days	N/A	code (Medication Drug, Medication Ingredient, Procedure)	CPT/HCPCS NDC, RxNorm	Secondary	Metastatic disease indicator[c]

Comparator group (per-protocol analysis): Chemotherapy	A static treatment regime, defined as follows: Concurrent use of pemetrexed and platinum therapy (cisplatin or carboplatin) every 21 days for 4 cycles, followed by use of pemetrexed every 21 days. May cross over to pembrolizumab monotherapy.	N/A	code (Medication Drug, Medication Ingredient, Procedure)	CPT/HCPCS NDC, RxNorm	Secondary	Metastatic disease indicator[c]

[a] Assessment times shown are inclusive.

[b] See appendix for listing of clinical codes for each study parameter.

[c] Indicators of metastatic disease include records indicating stage 4 disease, indicators of metastasis derived from the data vendor, or ICD-O diagnostic codes indicating metastasis. Information on metastatic disease is drawn from a linked tumor registry and tumor morphology descriptors contained within the health record.

Context and rationale for outcome(s) of interest

The outcome of interest is overall survival, which will be estimated as the probability of survival beyond 12 months from time zero and as a marginal relative hazard of mortality comparing the exposure vs. comparator groups. Month and year of mortality are available for patients in the data source; however, exact dates of death are not available to preserve patient privacy. Date of death will be assumed to have occurred on the first day of the month of death given in the data source.

Table 6. Operational Definitions of Outcome

Outcome name	Details	Assessment window	Type of outcome	Variable (Data Table Name(s))	Applied to study populations:
12-month survival	Probability of surviving at least 12 months from time zero	[1, end of follow-up]	Binary	month_year_death (Patient Demographic)	Primary and secondary
Hazard ratio for mortality	Relative hazard of all-cause mortality	[1, end of follow-up]	Time-to-event	month_year_death (Patient Demographic)	Primary and secondary

Context and rationale for follow-up

Primary analysis (intention-to-treat): The effect of *treatment initiation* on the outcome of interest will be estimated (i.e., "intention-to-treat effect"), where follow-up will begin on the day after a 14-day exposure regimen ascertainment window and proceed until the outcome event (death), loss to follow-up (>90-day gap in health record activity), 640 days after the index date, or end of available data occurs. "Health record activity" will be defined by lab results, treatments, and vitals records. See Figure 1-A for an illustration of follow-up in the primary intention-to-treat analysis. Administrative censoring at 21 months (640 days) will be imposed to align with the duration of the KEYNOTE-189 trial at the time initial results were published.

Secondary analysis (per-protocol): The effect of the *treatment* on the outcome will be estimated (i.e., "per-protocol effect"), where follow-up will begin on the day after initiation of a study drug in each regimen of interest and proceed until an outcome event (death), treatment cessation (deviation from regimens described in the variables section of this protocol), loss to follow-up (>90-day gap in health record activity), 640 days after the index date, or end of available data occurs. "Health record activity" will also be defined by lab results, treatments, and vitals records (similar to the primary analysis). In aligning with the KEYNOTE-189 trial, cross-over from the non-pembrolizumab group to pembrolizumab monotherapy will be permitted. Administrative censoring at 21 months (640 days) will be imposed to align with the duration of the KEYNOTE-189 trial at the time initial results were published. See Figure 1-B for illustrations of follow-up in the secondary per-protocol analysis.

Table 7. Operational Definitions of Follow-Up

Primary analysis (intention-to-treat)

Follow-up start[a]	Day 1	
Follow-up end[b]	**Select all that apply**	**Specify**
Date of outcome	Yes	Mortality
End of observation in data	Yes	>90-day gap in any health record activity (lab results, treatments, and vitals records) OR End of data
Day X following index date	Yes	Day 640 (End of study period)
End of study period	No	
End of exposure	No	
Date of add to/switch from exposure	No	
Other date	No	

Secondary analysis (per-protocol)

Follow-up start[a]	Day 1	
Follow-up end[b]	**Select all that apply**	**Specify**
Date of outcome	Yes	Mortality
End of observation in data	Yes	>90-day gap in any health record activity (lab results, treatments, and vitals records)
Day X following index date	Yes	Day 640
End of study period	Yes	End of data
Date of treatment change or cessation	Yes	Patients will be censored upon deviation of their "assigned" static treatment regime described in the Exposure section
Other date	No	

[a] Follow-up begins at time shown relative to 'time zero' (see Time Zero for more information).
[b] Follow-up ends at the first occurrence of any of the selected criteria that end follow-up.

Context and rationale for covariates

Confounding variables were chosen on the basis of the investigators' substantive knowledge and the backdoor criterion.[7]

The Charlson comorbidity index has been shown to be a prognosticator of survival in patients with NSCLC, particularly those with no EGFR or ALK mutations.[8,9]

Table 8. Operational Definitions of Covariates

Characteristic	Details	Type of variable	Assessment window	Time-varying[a]	Variable (Data Table Name(s))	Code Type[b]	Applied to study populations:
Age	Years between year of index date and year of birth	Continuous	[0, 0]	No	year_of_birth (Patient Demographics)	Derived	Primary, Secondary
Sex	Patient's sex	Categorical	[0, 0]	No	sex (Patient Demographics)	Derived	Primary, Secondary
BMI	Patient's BMI	Continuous	[0, 0]	Yes	code (Vitals)	LOINC	Primary, Secondary
Marital status	Patient's marital status (married, single, unknown)	Categorical	[0, 0]	No	marital_status (Patient Demographics)	Derived	Primary, Secondary
Race	Patient's race (American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, White, Unknown)	Categorical	[0, 0]	No	race (Patient Demographics)	Derived	Primary, Secondary
Ethnicity	Patient's ethnicity (Hispanic or Latino, Not Hispanic or Latino, Unknown)	Categorical	[0, 0]	No	ethnicity (Patient Demographics)	Derived	Primary, Secondary
Region	Patient's regional location in the United States (US) or outside the US.	Categorical	[0, 0]	No	patient_regional_location (Patient Demographics)	Derived	Primary, Secondary
Diagnosis date	Date of original primary cancer diagnosis (or, if unavailable, date	Date	[-all available data, 0]	No	diagnosis_date (Tumor)	Derived	Primary, Secondary

Characteristic	Details	Type of variable	Assessment window	Time-varying[a]	Variable (Data Table Name(s))	Code Type[b]	Applied to study populations:
	of incident metastatic disease indicator)						
Performance status	ECOG performance status (or equivalent value of ECOG that relates to Karnofsky performance status LOINC Code for ECOG: 89247-1 TNX Code for ECOG: 2002 LOINC Code for Karnofsky: 89243-0 Categorize as follows: • Set to 0 if ECOG=0 or Karnofsky=[90, 100] • Set to 1 if ECOG=1 or Karnofsky=[70, 80] • Set to 2 if ECOG=2 or Karnofsky=[50, 60] • Set to 3 if ECOG=3 or Karnofsky=[30, 40] • Set to 4 if ECOG=4 or Karnofsky=[10, 20] Citation: Oken MM, Creech RH, Tormey DC, Horton J, Davis TE, McFadden ET, Carbone PP. Toxicity and response criteria of the Eastern Cooperative Oncology Group. Am J Clin Oncol. 1982 Dec;5(6):649-55. PMID: 7165009.	Categorical	ITT analysis: [-all available data, -15] most recent value PP analysis: [-all available data, 0]	Yes	code (Lab Results)	LOINC, Derived	Primary, Secondary
Treatment history	Prior systemic treatments See appendix for treatment types	Categorical	ITT analysis: [-diagnosis date, -first date indicating	Yes	code (Medication Drug, Medication Ingredient, Procedure)	RxNorm, NDC, CPT/HCPCS	Primary, Secondary

Characteristic	Details	Type of variable	Assessment window	Time-varying[a]	Variable (Data Table Name(s))	Code Type[b]	Applied to study populations:
			metastatic disease]				
Neoadjuvant or Adjuvant Therapy	Occurrence of at least 1 prior systemic treatment defined in 'Treatment history' variable	Binary	N/A	No			
Corticosteroid use	Frequency of steroid use (count of records indicating corticosteroid use 30 days prior to treatment initiation)	Continuous	ITT analysis: [-45, -15] PP analysis: [-30, -1]	Yes	code (Medication Drug, Medication Ingredient, Procedure)	RxNorm, NDC, CPT/HCPCS	Primary, Secondary
PD-L1 expression/results	PD-L1 test Report. Categorize as follows: • Set to 0 if <1%, • Set to 1 if ≥1% • Set to 2 if 1-49% • Set to 3 if ≥50% • otherwise set to "NA" <1%, ¸≥1%, 1-49%, ≥50% *Assess up to 90 days after index date. Justification: PD-L1 test results may be delayed following treatment initiation and PD-L1 status is not assumed to be a mediator of the effect of treatment on the outcome.*	Categorical	[-all available data, 90] Take latest value	No	code (Lab Result)	LOINC	Primary, Secondary
Smoking status	Current or former; never	Categorical	[-all available data, 0] most recent value	No	code (Diagnosis)	ICD	Primary, Secondary

Characteristic	Details	Type of variable	Assessment window	Time-varying[a]	Variable (Data Table Name(s))	Code Type[b]	Applied to study populations:
Brain metastases	History of brain metastasis	Binary	ITT analysis: [-diagnosis date, -15] PP analysis: [-diagnosis date, 0]	Yes	code (Diagnosis) tumor_site_code (Tumor, Tumor Properties)	ICD	Primary, Secondary
Thoracic radiotherapy	History of thoracic radiotherapy for non-metastatic disease	Binary	ITT analysis: [-diagnosis date, -15] PP analysis: [-diagnosis date, 0]	No	code (Procedure)	ICD, HCPCS/CPT	Primary, Secondary
Lung resection	History of lung resection	Binary	ITT analysis: [-diagnosis date, -15] PP analysis: [-diagnosis date, 0]	No	code (Diagnosis)	ICD, HCPCS/CPT	Primary, Secondary
Estimated glomerular filtration rate (eGFR)	eGFR derived from data vendor	Continuous	ITT analysis: [-all available data, -15] PP analysis: [-all available data, 0] Most recent record	Yes	code (Lab Result)	LOINC, Derived	Primary, Secondary

Characteristic	Details	Type of variable	Assessment window	Time-varying[a]	Variable (Data Table Name(s))	Code Type[b]	Applied to study populations:
Venous thromboembolism	History of deep vein thrombosis or pulmonary embolism	Binary	ITT analysis: [-all available data, -15] PP analysis: [-diagnosis date, 0] most recent value	Yes	code (Diagnosis)	ICD, HCPCS/CPT	Primary, Secondary
Time since metastatic indicator	Time between metastatic disease indicator and 'time zero'	Continuous	ITT Analysis: [-first date indicating metastatic disease, 0] PP analysis: -first date indicating metastatic disease, 0]	No	metastatic (Tumor) stage_code (Tumor) morphology_code (Tumor, Tumor Properties) code (Diagnosis)	ICD, derived	Primary, Secondary
Vitals frequency	Frequency of vitals records in year prior (proxy for in-network engagement)	Continuous	ITT analysis: [-380, -15] PP analysis: [-365, 0]	Yes	code (Vitals Signs) date (Vitals Signs)	LOINC	Primary, Secondary
Charlson comorbidity index	Charlson comorbidity index	Categorical	ITT analysis: [-380, -15] PP analysis: [-365, 0] most recent value	Yes	code (Diagnosis)	ICD	Primary, Secondary

Characteristic	Details	Type of variable	Assessment window	Time-varying[a]	Variable (Data Table Name(s))	Code Type[b]	Applied to study populations:
Histology (adenocarcinoma)	Tumor histology (adenocarcinoma) Set equal to 1 if relevant codes in appendix occur anytime prior to or on 90 days post-index date. Otherwise set to 0.	Binary	[-all available data, 90]	No	morphology_code (Tumor)	ICD-O	Primary, Secondary

[a] Time-varying covariates that will be considered for inclusion in statistical models for weight estimation in the secondary ("per-protocol") analysis. Variables will be assessed in the "assessment window" stated in the table and updated at every discrete time interval of follow-up.

[b] See appendix for listing of clinical codes for each study parameter

5.4. Data analysis

Context and rationale for analysis plan

Table 9. Primary, secondary, and subgroup analysis specification

A. Primary analysis (intention-to-treat)

Hypothesis:	The relative hazard of mortality is not equal among patients who initiated pembrolizumab and chemotherapy vs. chemotherapy alone.
Exposure contrast:	Patients initiating pembrolizumab, pemetrexed, and platinum therapy (cisplatin or carboplatin) vs. patients initiating pemetrexed and platinum therapy (cisplatin or carboplatin) within a 14-day exposure-assessment window following the first occurrence of metastatic disease
Outcome:	Marginal hazard ratio for mortality, standardized to the empirical distribution of baseline confounders of the entire study population.
Analytic software:	R Statistical Software (v4.1.2; R Core Team 2021) Packages: tidyverse (v2.0.0), mice (v3.15.0), survival (v3.1-8)
Model(s):	$h(t, L_0) = h_0(t) * e^{\beta_1 A_0 + \beta_2^T L_0}$

	where $h(t, L_0)$ is the hazard of mortality at discrete time interval 't,' conditional on the vector of potential confounders, "L_0," assessed at baseline; $h_0(t)$ is the baseline hazard at discrete time interval 't'; and A_0 is an indicator for treatment initiation, coded as "1" and "0" for the exposure and comparator group, respectively. The proportional hazards assumption will be checked through visual inspection of plots containing scaled Schoenfeld residuals vs. time for each covariate. The Breslow method of handling ties will be used.
Confounding adjustment method	
	Stabilized, time-fixed, inverse probability of treatment weights will be used to adjust for confounding. Individual-level weights will be estimated by the following formula: $$SW^A = \frac{f(A_0)}{f(A_0 \mid L_0)}$$ where A_0 is the treatment that the patient received and L_0 is a vector of baseline confounders. The quantity in the denominator will be estimated using a logistic regression model with A_0 as the dependent variable and the L_0 vector as the independent variables and the quantity in the numerator will be estimated with a logistic regression model containing A_0 as the dependent and sole variable in the model. The model will be fit to a person level dataset (i.e., one row per patient, with each row representing a patient's complete experience during baseline and follow-up). All potential confounder variables will be considered for inclusion in the weight estimation (Table 8). However, as it is not possible to predict the quantity of missing values and sparseness of the data at the time of writing this protocol, the precise functional form of the final regression model will be determined at the time of analysis. The most flexible functional form will be chosen to reduce the possibility of bias due to modelling assumptions. To verify that weights were estimated correctly, the mean weight will be checked to be approximately equal to 1. The empirical distribution of weights will be used to determine whether there are any extreme values, suggesting positivity violations. In the presence of suspected random positivity violations (e.g., due to sparse data), modelling assumptions will be altered to smooth over "zero" or "near-zero" quantities within levels of the exposure and covariates. If this approach is unsuccessful, weight trimming may be considered whereby patients with weights in the top and bottom 1st percentile will be removed from the analysis. Structural positivity violations should not be possible, given our eligibility criteria and treatment strategies.
Missing data methods	

	The data source is comprised of data drawn from over 50 healthcare networks, some of which may contribute certain variables to varying degrees. This is one plausible reason for missing values, and while identifying information on contributing healthcare networks is not available in the database, it can be proxied by patient regional location (and all other measured variables in analysis). This theory is supported by dependence of missing indicators for various variables (bmi, marital status, etc.) on region. Therefore, the data are believed to be Missing at Random (MAR). To address missing values that are MAR, multiple imputation with chained equations (fully conditional specification) can be employed. This approach is appropriate to address data that are MAR and permits the flexibility needed to impute different variable types (continuous, categorical, etc.). A series of conditional models will be specified to estimate the parameters of a joint distribution from which the imputed values will be drawn. Predictive mean matching, ordered logistic regression, and multinomial logistic models will be considered to impute continuous, ordinal, and nominal variables that contain missing values, respectively. The final imputation models and variables to be imputed will be determined by the extent of missingness and structural relationship of the study variables of interest. Inference regarding the final effect estimates will account for the imputation procedure. Please refer to the appendix for more information on missing data.
Subgroup Analyses	
	Subgroup analyses will be conducted similarly to the KEYNOTE-189 trial[1] to facilitate comparison: 1. Age (<65, ≥65) 2. Sex (male, female) 3. ECOG performance Status (0, 1, 2+) 4. Smoking status (current or former, never) 5. Brain metastases at baseline (yes, no) 6. PD-L1 tumor proportion score (<1%, ≥1%, 1-49%, ≥50%) 7. Platinum-based drug (carboplatin, cisplatin)

Hypothesis:	The 12-month survival probability is not equal among patients who initiated pembrolizumab and chemotherapy vs. chemotherapy alone.

Exposure contrast:	Patients initiating pembrolizumab, pemetrexed, and platinum therapy (cisplatin or carboplatin) vs. patients initiating pemetrexed and platinum therapy (cisplatin or carboplatin) within a 14-day exposure-assessment window following the first occurrence of metastatic disease.
Outcome:	Marginal 12-month survival probability, standardized to the empirical distribution of baseline confounders of the entire study population.
Analytic software:	R Statistical Software (v4.1.2; R Core Team 2021) Packages: tidyverse (v2.0.0), mice (v3.15.0), survival (v3.1-8), survminer (v0.4.9), boot (v1.3-28)
Model(s):	A Kaplan-Meier estimator, weighted by time-fixed inverse probability of treatment weights will be used to estimate the 12-month survival probability.[11] The weighted survival probability $S_a(t)$ for treatment group 'A=a' at time 't' will be as follows: $$S_a(t) = \prod_t 1 - \frac{d_{ta}}{r_{ta}}$$ where $d_{ta} = w_{it} \cdot Y_{it} \cdot I(A_{it} = a)$ denotes the weighted number of mortality events and $r_{ta} = \sum_{i=1}^{N} w_{it} \cdot I(A_{it} = a)$ denotes the weighted risk set. A non-parametric bootstrap or cluster-robust standard errors will be used to derive 95% confidence intervals around 12-month survival probability estimates.
Confounding adjustment method	
	Inverse probability of treatment weights will be estimated to control for potential confounding using the same approach outlined above (see intention-to-treat mortality hazard ratio).
Missing data methods	
	The process of accounting for missing values in this analysis will be the same as outlined above (see intention-to-treat mortality hazard ratio).
Subgroup Analyses	List all subgroups
	Provided PD-L1 tumor proportion score is available in the data source, overall survival will be estimated within subgroups of patients with a PD-L1 tumor proportion score of <1%, 1-49%, and ≥50% to facilitate comparison with the KEYNOTE-189 trial.[1]

B. Secondary analysis (per-protocol)

Hypothesis:	The relative hazard of mortality is not equal among patients who adhered to the static treatment regimen of pembrolizumab and chemotherapy vs. chemotherapy alone, as illustrated and specified in the study design and variables section of this protocol.

Exposure contrast:	Concurrent use of pembrolizumab, pemetrexed, and platinum therapy (cisplatin or carboplatin) every 21 days for 4 cycles, followed by concurrent use of pembrolizumab and pemetrexed every 21 days Vs. Concurrent use of pemetrexed and platinum therapy (cisplatin or carboplatin) every 21 days for 4 cycles, followed by use of pemetrexed every 21 days. May cross over to pembrolizumab monotherapy.
Outcome:	Marginal hazard ratio for mortality, standardized to the empirical distribution of baseline and time-varying confounders of the entire study population.
Analytic software:	R Statistical Software (v4.1.2; R Core Team 2021) Packages: tidyverse (v2.0.0), mice (v3.15.0), survey (v4.1-1)
Model(s):	A weighted, pooled-over-time marginal structural logistic regression model will be used to estimate the per-protocol mortality hazard ratio as follows: $$logit[\Pr(D_t = 1 \mid A, L_0, \overline{D}_{t-1} = 0, C_{t-1} = 0)] = \alpha_{0,t} + \beta_1 A + \beta_2^T L_0$$ where D_t denotes a death event at time 't', 'A' is a time-updated binary indicator of adherence to the static treatment regime specified in the study design and variables section of this protocol (taking on value of 0 for the comparator group and 1 for exposed group at discrete time points 't'), L_0 denotes a time-fixed vector of confounders assessed at baseline, C_{t-1} is an indicator of censoring at time 't-1' (e.g., due to non-adherence to regime of interest, loss to follow-up, or administrative end of data), and $\alpha_{0,t}$ is a time-varying intercept (estimated as a polynomial function of 't'). Provided the probability of morality is less than 15% in each discrete time interval, the treatment coefficient of the pooled logistic model will approximate a relative hazard. As it is not possible to predict the quantity of missing values and sparseness of the data at the time of writing this protocol, the precise functional form of the final regression model will be determined at the time of analysis. The most flexible functional form will be chosen to reduce the possibility of residual bias arising from modelling assumptions.
Confounding adjustment method	
	To account for potential confounding bias, stabilized time-varying inverse probability of treatment weights will be applied to the study sample to generate a pseudo-population, in which the treatment at time 't' is independent of confounder history. The weights will take on the following form:

$$SW^A = \prod_{t=0}^{T} \frac{f(A_t|\bar{A}_{t-1})}{f(A_t|\bar{A}_{t-1}, L_0, \bar{L}_v)}$$

where A_t denotes time-varying treatment status that the patient actually received at discrete time 't' (taking on values 0 for adherence to the comparator regimen and 1 for adherence to the exposed regimen), L_0 denotes a vector of time-fixed baseline confounders, and L_v denotes a vector of time-varying confounders. The overbar is used to represent treatment history.

The conditional probability mass functions that comprise the numerator and denominator of SW^A will be estimated using a pooled logistic regression model that includes treatment status at time 't', treatment history, covariate history (baseline and time-varying), and a time-varying intercept (using functional of time using splines). The model will be fit to a person-time level dataset (i.e., multiple rows per patient, each row representing a patient's experience during a discrete time interval). Each person-time contribution to the weight will be calculated by taking the cumulative product of predicted probabilities from the model over each person-time interval from t=0 through event or censoring, "T," where T = min(event, censoring).

As illustrated in the study design and variables section of this protocol, patients will be censored when they do not adhere to the treatment regimens of interest. To account for potential informative censoring induced by this artificial censoring rule, inverse probability of censoring SW^C weights will be estimated and applied jointly with SW^A to the study sample. The censoring weights will take on the following form:

$$SW^C = \prod_{t=0}^{T} \frac{f(C_t = 0|\bar{A}_{t-1}, C_{t-1} = 0)}{f(C_t = 0|\bar{A}_{t-1}, C_{t-1} = 0, \bar{L}_v)}$$

Intuitively, censoring weights weight patients who remain in the analysis, and have the same treatment and covariate history as those who were censored, more heavily to "compensate" for those that were censored. Therefore, the mean weight over follow-up time should be 1. This will be verified graphically to check for model misspecification.

Using a similar approach to treatment weight estimation, the conditional probability mass functions that comprise the numerator and denominator of SW^C will be estimated using a pooled logistic regression model that includes censoring status at time 't', treatment history, covariate history (time-varying), and a time-varying

	intercept (using functional of time using splines) for the denominator, and the same variables (except for covariate history) for the numerator. The final study sample will be weighted by the product of the two weights at each person-time interval as follows: $$SW^{A,C} = SW^{A} \cdot SW^{C}$$
Missing data methods	
	The data source is comprised of data drawn from over 50 healthcare networks, some of which may contribute certain variables to varying degrees. This is one plausible reason for missing values, and while identifying information on contributing healthcare networks is not available in the database, it can be proxied by patient regional location (and all other measured variables in analysis). This theory is supported by dependence of missing indicators for various variables (bmi, marital status, etc.) on region. Therefore, the data are believed to be Missing at Random (MAR). To address missing values that are MAR, multiple imputation with chained equations (fully conditional specification) can be employed. This approach is appropriate to address data that are MAR and permits the flexibility needed to impute different variable types (continuous, categorical, etc.). A series of conditional models will be specified to estimate the parameters of a joint distribution from which the imputed values will be drawn. Predictive mean matching, ordered logistic regression, and multinomial logistic models will be considered to impute continuous, ordinal, and nominal variables that contain missing values, respectively. The final imputation models and variables to be imputed will be determined by the extent of missingness and structural relationship of the study variables of interest. Inference regarding the final effect estimates will account for the imputation procedure. Please refer to the appendix for more information on missing data.
Subgroup Analyses	List all subgroups
	Subgroup analyses will be conducted similarly to the KEYNOTE-189 trial[1] to facilitate comparison: 1. Age (<65, ≥65) 2. Sex (male, female) 3. ECOG performance Status (0, 1, 2+) 4. Smoking status (current or former, never) 5. Brain metastases at baseline (yes, no)

	6. PD-L1 tumor proportion score (<1%, ≥1%, 1-49%, ≥50%) 7. Platinum-based drug (carboplatin, cisplatin)

Hypothesis:	The 12-month survival probability is not equal among patients who adhered to the static treatment regimen of pembrolizumab and chemotherapy vs. chemotherapy alone, as illustrated and specified in the study design and variables section of this protocol.
Exposure contrast:	Concurrent use of pembrolizumab, pemetrexed, and platinum therapy (cisplatin or carboplatin) every 21 days for 4 cycles, followed by concurrent use of pembrolizumab and pemetrexed every 21 days Vs. Concurrent use of pemetrexed and platinum therapy (cisplatin or carboplatin) every 21 days for 4 cycles, followed by use of pemetrexed every 21 days. May cross over to pembrolizumab monotherapy.
Outcome:	Marginal 12-month survival probability, standardized to the empirical distribution of baseline and time-varying confounders of the entire study population.
Analytic software:	R Statistical Software (v4.1.2; R Core Team 2021) Packages: tidyverse (v2.0.0), mice (v3.15.0), survival (v3.1-8), survminer (v0.4.9), boot (v1.3-28)
Model(s):	A Kaplan-Meier estimator, weighted by time-varying inverse probability of treatment and censoring weights will be used to estimate the 12-month survival probability.[11] The weighted survival probability $S_a(t)$ for treatment group 'A=a' at time 't' will be as follows: $$S_a(t) = \prod_t 1 - \frac{d_{ta}}{r_{ta}}$$ where $d_{ta} = w_{it} \cdot Y_{it} \cdot I(A_{it} = a)$ denotes the time-specific weighted number of Y_{it} mortality events and $r_{ta} = \sum_{i=1}^{N} w_{it} \cdot I(A_{it} = a)$ denotes the weighted risk set. A non-parametric bootstrap or cluster-robust standard errors will be used to derived 95% confidence intervals around 12-month survival probability estimates.
Confounding adjustment method	
	Time-varying inverse probability of treatment and censoring weights will be estimated to control for potential confounding using the same approach outlined above (see " $SW^{A,C}$ " estimated for estimation of the per-protocol mortality hazard ratio).
Missing data methods	

	The process of accounting for missing values in this analysis will be the same as that outlined above in the per-protocol mortality hazard ratio section on missing data methods.
Subgroup Analyses	List all subgroups
	Provided PD-L1 tumor proportion score is available in the data source, overall survival will be estimated within subgroups of patients with a PD-L1 tumor proportion score of <1%, 1-49%, and ≥50% to facilitate comparison with the KEYNOTE-189 trial.[1]

Table 10. Sensitivity and exploratory analyses – rationale, strengths and limitations

	What is being varied? How?	Why? (What do you expect to learn?)	Strengths of the sensitivity analysis compared to the primary	Limitations of the sensitivity analysis compared to the primary
		Sensitivity Analyses		
Contemporaneous cohort	The study population will be restricted to subjects with a time of cohort entry ("index date") of June 6, 2017 or later.	Pembrolizumab first received accelerated approval for use as a first-line agent in NSCLC on this date. Furthermore, the use of historical controls may introduce bias if time of cohort entry is a confounder. This analysis tests the implicit assumption of the main analysis that time calendar time is not a confounder.	A contemporaneous cohort will allow for a more closely matched cohort across exposure groups with respect to calendar time, potentially minimizing the risk of confounding by calendar time.	Because this analysis excludes subjects, the sample size may be substantially lower by design. Consequently, precision of outcome estimates may be reduced.
Treatment administrations only	Treatment will only be defined using procedure codes indicating medication administration (i.e., HCPCS/CPT codes)	Medication orders, administrations, and dispensations are not clearly distinguished in the data source. For this reason, the primary analysis implicitly assumes all medication records were tantamount to administrations.	Medication administrations are of particular interest, as they directly indicate a patient's exposure to the drug (as opposed to medication orders or dispensations)	HCPCS/CPT codes may not always appear in patient records when a drug is administered, leading to potential measurement error.

		Robustness of our primary results to this analysis can test the reasonableness of this implicit assumption.		
Require treatment initiation within 6 months of diagnosis metastatic disease	Patients will be required to initiate treatment within 6 months of the incident metastatic indicator used to establish 'time zero'	Patients with long time lapses between metastasis and treatment initiation may be receiving out-of-network care and therefore have incomplete capture of key variables for the analysis. These patients may also indicate measurement error in the assessment of initial metastasis. Analysis of patients who initiate treatment soon after metastasis may better reflect the target population of interest and patients who have more complete healthcare information in the database.	This sensitivity analysis indirectly tests the assumption that the initial occurrence of metastatic disease and the target population of interest are being ascertained accurately.	This sensitivity analysis may reduce sample size and precision of effect estimates. It is also indirectly testing assumptions that can't be verified empirically.
		Exploratory Analyses		
Differential treatment intensity	The rate of engagement with the health-system (e.g., frequency of office visits, radiology reports/scans, or encounters) will be assessed within each treatment group over the follow-up period.	Differential levels of care between the study groups may influence outcomes and serve as a proxy for residual or unmeasured confounding.	This analysis may provide a sense of confidence (or lack thereof) in the comparability of treatment groups.	Differential treatment intensity is not guaranteed to adequately capture all (or any) elements of residual or unmeasured confounding.
Censoring event distributions	The distribution of censoring event times over the study	A differential pattern of censoring over the follow-up	This analysis may provide a sense of confidence (or lack	A differential pattern of censoring events does not

	period will be assessed within each treatment group over the follow-up period	period between treatment groups could indicate the presence of informative censoring.	thereof) in the comparability of treatment groups.	directly indicate the reasons for the observed pattern and further explorations may be needed.

5.5. Data sources

Summary of data source: The data source used for this study is derived from electronic health records of 52 healthcare organizations from the US and the Asia-Pacific region. In particular, the data are drawn from a combination of structured fields in the health record (demographics, date-indexed encounters, diagnoses, procedures, and medications) and natural language processing of free-text provider notes. The data are enriched using tumor/cancer registries for oncology information and obituaries and Social Security records for mortality data. The healthcare organizations from which the data are drawn are varied and include acute care hospitals, networks of outpatient clinics, academic medical centers, pediatric hospitals, and cardiac care and surgical centers. Although the data are sourced from many institutions, those institutions are not identifiable in the data for privacy purposes.

Context and rationale for data sources

Reason for selection: The process for selecting a data source is described in a previously published manuscript.[12] Briefly, data sources meeting operational definitions for eligibility criteria, exposure, outcome, and confounders determined by the investigators were considered for use. Among these initially identified data sources, additional factors were considered to select a final fit-for-purpose dataset, including data quality, quantity (sample size), and provenance. Because data is sourced from entire health care organizations, which house several in-network hospitals, clinics, and other institutions, patients' experiences are captured more completely.

Strengths of data source(s): Major strengths of the data source used in this study include its linkage with a tumor/cancer registry, availability of all (including non-oncology) diagnosis codes contained in the health records, and presence of medication history through a variety of sources (medication reconciliation, drug ordering system, administrations, dispensations, etc.). Information is captured from natural language processing of unstructured clinical notes, as well as structured data fields within the health record. Importantly, results of tumor biomarkers and genomic testing are available, facilitating appropriate selection of our cohort of interest (i.e., patients without EGFR or ALK mutations in tumor cells). Collectively, these qualities enable ascertainment of patients' longitudinal clinical histories at a relatively high level of detail as well as identification of confounders that are not oncology-related (e.g., comorbidities).

Limitations of data source(s): Limitations include the inability to distinguish between medication record types (i.e., orders, administrations, and dispensations), unavailable information on out-of-network encounters (i.e., EHR discontinuity[13]), and limited drug dosing information. Natural language processing may also result in an unknown degree of measurement error, as the models used to derive variables have not been validated. Many of these

factors are limitations inherent to many EHR-based data sources. Mortality has not been validated in the database and is subject to measurement error as well.

Table 11. Data source provenance and completeness

Provenance	**Information type**	**Availability**	**Notes**
Cancer registry data	Cancer type/diagnosis	X	Cancer type is derived from standardized (ICD-O) codes drawn from the cancer registry
	Tumor histology	X	Tumor histology is derived from standardized (ICD-O) codes drawn from the cancer registry
	Disease state (e.g., TNM, AJCC staging)	X	Staging information is derived from standardized (ICD-O) codes drawn from the cancer registry
	Time of onset/original diagnosis	X	
EHR (structured fields)	Cancer type/diagnosis	X	Cancer type is derived from standardized (ICD) codes drawn from the structured fields of the health record.
	Tumor histology		
	Disease state (e.g., TNM, AJCC staging)		
	Time of onset/original diagnosis		
	Comorbidities	X	Provided in the form of standardized codes (e.g., ICD)
	Procedures	X	Provided in the form of standardized codes (e.g., HCPCS/CPT, ICD)
	Lab results	X	Provided in the form of standardized codes (e.g., LOINC)
	Medication orders	X	Medication orders, administrations, and dispensations are indistinguishable
	Medication administration	X	Medication orders, administrations, and dispensations are indistinguishable

	Medication dispensation	X	Medication orders, administrations, and dispensations are indistinguishable
	Demographics	X	
	Vitals	X	Provided in the form of standardized codes (e.g., LOINC)
EHR (unstructured data (e.g., notes))	Cancer type/diagnosis		
	Tumor histology		
	Disease state (e.g., TNM, AJCC staging)		
	Time of onset/original diagnosis		
	Comorbidities	X	Some comorbidity data derived from natural language processing of clinical notes.
	Procedures	X	Some procedure data is derived from natural language processing of clinical notes.
	Lab results	X	Some laboratory data is derived from natural language processing (e.g., eGFR estimated from serum creatinine and other components of the Cockcroft-Gault equation).
	Medication orders	X	Some medication use data is derived from natural language processing of clinical notes.
	Medication administration	X	Some medication use data is derived from natural language processing of clinical notes.
	Medication dispensation	X	Some medication use data is derived from natural language processing of clinical notes.
	Demographics		
	Vitals	X	Some vitals data is derived from natural language processing of clinical notes.
External laboratory data	Molecular diagnostics	X	Provided in the form of standardized codes (e.g., HGVS)

Insurance claims	Outpatient diagnoses		
	Outpatient procedures		
	Outpatient medications		
	Inpatient diagnoses		
	Inpatient procedures		
	Inpatient medications		

Table 12. Metadata about data sources and software

Data Source(s):	[Redacted]
Study Period:	3/6/1946 – 1/17/2023
Eligible Cohort Entry Period:	3/6/1946
Data Version (or date of last update):	1/18/2023
Data sampling/extraction criteria:	* Must Have: C34 Malignant neoplasm of bronchus and lung * Group 1A Tumor Registry: The terms in this group occurred at any time * Must Have: C34 Malignant neoplasm of bronchus and lung [ACIDOPHIL CARCINOMA OR ACINAR CELL CARCINOMA OR ADENOCA IN FAMIL POLYP COLI OR ADENOCA. IN ADENOMA. POLYP OR ADENOCA. WITH METAPLASIA OR ADENOCARCINOFIBROMA OR ADENOCARCINOMA, NOS OR ADENOID CYSTIC & CRIBRIFORM CA. OR ADENOSQUAMOUS CARCINOMA OR ADRENAL CORTICAL CARCINOMA OR ALVEOLAR RHABDOMYOSARCOMA OR AMELANOTIC MELANOMA OR AMELOBLASTIC FIBROSARCOMA OR AMELOBLASTIC ODONTOSARCOMA OR AMELOBLASTOMA, MALIGNANT OR ANGIOLIPOMA OR ASTROBLASTOMA OR ASTROCYTOMA, NOS OR BASAL CELL CARCINOMA, NOS OR BASOPHIL CARCINOMA OR BENIGN LIPOMA OR BLOOD VESSEL TUMORS OR BLUE NEVUS, MALIGNANT OR BRENNER TUMOR, MALIGNANT OR BRONCHIOLO-ALVEOLAR ADENOCA. OR CARCINOID TUMOR, MALIGNANT OR CARCINOMA, NOS OR CARCINOMA, UNDIFF., NOS OR CARCINOSARCOMA, NOS OR CELLULAR BLUE NEVUS OR CEREBELLAR SARCOMA, NOS OR CERUMINOUS ADENOCARCINOMA OR CHOLANGIOCARCINOMA OR

CHONDROBLASTOMA, MALIGNANT OR CHONDROSARCOMA, NOS OR CHORDOMA OR CHORIOCARCINOMA OR CHROMOPHOBE CARCINOMA OR CHRONIC MYELOPROLIFERATIVE DIS. OR CLEAR CELL ADENOCARCINOMA, NOS OR CLEAR CELL SARC/NEPHROBLASTOMA OR COMB HEPATOCEL CA. & CHOLANG OR CRANIOPHARYNGIOMA OR CYSTADENOCARCINOMA, NOS OR DUCT CARCINOMA OR EMBRYONAL CARCINOMA, NOS OR EMBRYONAL RHABDOMYOSARCOMA OR ENDOCRINOMAS OR ENDOMETRIOID ADENOCARCINOMA OR EPENDYMOMA, NOS OR EPITHELIAL NEOPLASM OR EPITHELIOID CELL MELANOMA OR ERYTHROID LEUKEMIA OR EWING SARCOMA OR EXTRA-ADRENAL PARAGANG., MAL OR FIBRILLARY ASTROCYTOMA OR FIBROMA OR FIBROMATOUS NEOPLASMS OR FIBROUS HISTIOCYTOMA, MAL. OR FOLLIC. & MARGINAL LYMPH, NOS OR FOLLICULAR ADENOCARCINOMA, NOS OR GANGLIONEUROBLASTOMA OR GERM CELL TUMORS OR GIANT & SPINDLE CELL CARCINOMA OR GLIOBLASTOMA, NOS OR GLIOMA OR GLOMANGIOSARCOMA OR GONADAL NEOPLASMS OR GRANULAR CELL CARCINOMA OR GRANULAR CELL TUMOR OR GRANULOSA CELL TUMOR, MAL. OR GT. CELL TUMOR OF BONE, MAL. OR HEMANGIOBLASTOMA OR HEMANGIOENDOTHELIOMA OR HEMANGIOPERICYTOMA OR HEPATOCELLULAR CARCINOMA, NOS OR HILUS CELL TUMOR OR HODGKIN LYMPHOMA OR HODGKIN LYMPHOMA, NOD. SCLER. OR IMMUNOPROLIFERATIVE DISEASES OR INFLAMMATORY CARCINOMA OR INTRADERMAL NEVUS OR JUVENILE FIBROADENOMA OR JUXTACORTICAL OSTEOSARCOMA OR KAPOSI SARCOMA OR LEUKEMIA, NOS OR LEYDIG CELL TUMOR, MALIGNANT OR LIPID CELL TUMOR, MAL. OR LIPOSARCOMA NEOPLASMS OR LOBULAR AND OTHER DUCTAL CA. OR LUTEOMA OR LYMPHANGIOSARCOMA OR LYMPHOEPITHELIAL CARCINOMA OR LYMPHOID LEUKEMIA, NOS OR MAL. MEL. IN JUNCT. NEVUS OR MAL. MELAN. IN GIANT PIGMT. NEVUS OR MALIGNANT LYMPHOMA, NOS OR MAST CELL TUMORS OR MEDULLARY CARCINOMA, NOS OR MEDULLOBLASTOMA, NOS OR MEGAKARYOBLASTIC LEUKEMIA OR MENINGIOMA OR MESENCHYMOMA, MALIGNANT OR MESONEPHROMA, MALIGNANT OR MESOTHELIOMA, MALIGNANT OR MIXED TUMOR, MALIGNANT, NOS OR ML, LARGE B-CELL, DIFFUSE OR ML, SMALL B-CELL LYMPHOCYTIC OR MONOCYTIC/OTHER LEUKEMIA, NOS OR MUCINOUS ADENOCARCINOMA OR MUCINOUS CYSTADENOCARC., NOS OR MUCOEPIDERMOID CARCINOMA OR MULLERIAN MIXED TUMOR OR MYELODYSPLASTIC SYNDROMES OR MYELOID LEUKEMIA, NOS OR MYELOID SARCOMA OR MYELOLIPOMA OR MYELOPLASTIC/MYELOPROLIFERATIVE NEOPLASMS OR MYOMATOUS NEOPLASMS OR MYXOSARCOMA OR NEOPLASM OR NEOPLASMS OF HISTIOCYTES AND ACCESSORY LYMPHOID CELLS OR NEURILEMMOMA OR NEUROBLASTOMA, NOS OR NEUROFIBROSARCOMA OR NEVI & MELANOMAS OR NON-SMALL CELL CARCINOMA OR NONENCAPSUL. SCLEROSING CA. OR ODONTOGENIC CARCINOSARCOMA OR ODONTOGENIC NEOPLASM OR ODONTOGENIC TUMOR OR ODONTOGENIC TUMOR, MAL. OR ODONTOMA OR OLFACTORY NEUROGENIC TUMOR OR OLIGODENDROBLASTOMA OR OLIGODENDROGLIOMA, NOS OR OSSEOUS & CHONDROMATOUS NEOPLASMS OR OSTEOBLASTOMA OR OSTEOCHONDROMA OR OSTEOSARCOMA, NOS OR OTHER LEUKEMIAS OR OTHER MYELOID LEUKEMIAS OR OTHER SPEC. NON-HODGKIN LYMPHOMA OR OVARIAN STROMAL TUMOR, MAL. OR OXYPHILIC ADENOCARCINOMA OR PAGET DISEASE, MAMMARY OR PANCREATOBLASTOMA OR PAPILLARY & FOLLICULAR ADENOCA. OR PAPILLARY ADENOCARCINOMA, NOS OR PAPILLARY CARCINOMA, NOS OR

	PAPILLARY CYSTADENOCA., NOS OR PAPILLARY SEROUS CYSTADENOCA OR PAPILLARY TRANS. CELL CARCINOMA OR PAPILLOMATOSIS OR PARAGANGLIOMA OR PERINEURIOMA OR PERIPHERAL NEUROECTODERMAL TUMOR OR PHEOCHROMOCYTOMA OR PHYLLODES TUMOR,MAL. OR PILOMATRIX CARCINOMA OR PLASMA CELL TUMORS OR PLEXIFORM NEUROFIBROMA OR POLYCYTHEMIA VERA OR PRECURS. CELL LYMPHOBLASTIC LYMPH. OR PRECURSOR LYMPHOID NEOPLASMS OR PROLYMPH/PRECURS LEUKEMIA OR PROTOPLASMIC ASTROCYTOMA OR REFRACTORY ANEMIA OR RENINOMA OR RETINOBLASTOMA, NOS OR RHABDOMYOSARCOMA, NOS OR SARCOMA, NOS OR SEBACEOUS/ECCRINE ADENOCA. OR SERTOLI CELL CARCINOMA OR SIGNET RING CELL CARCINOMA OR SKIN APPENDAGE CARCINOMA OR SMALL CELL CARCINOMA OR SOLID CARCINOMA, NOS OR SQUAMOUS CELL CARCINOMA, NOS OR STROMAL SARCOMA OR STRUMA OVARII, MALIGNANT OR SWEAT GLAND ADENOCARCINOMA OR SYNOVIAL SARCOMA, NOS OR T-CELL LYMPHOMAS OR TERATOMA OR THECOMA, MALIGNANT OR THERAPY RELATED AC. MYEL. LEUK. OR THYMOMA, MALIGNANT OR TRABECULAR ADENOCARCINOMA OR TRANSITIONAL CELL CARCINOMA, NOS OR TRICHILEMMOCARCINOMA] * AND * Group 2A Non-Squamous: The terms in this group occurred at any time * Must Have: C34 Malignant neoplasm of bronchus and lung [ADENOCARCINOMA, NOS OR Solid carcinoma, NOS OR SOLID CARCINOMA, NOS OR null OR Bronchiolo-alveolar carcinoma, non-mucinous OR Adenocarcinoma with mixed subtypes OR PAPILLARY ADENOCARCINOMA, NOS OR CLEAR CELL ADENOCARCINOMA, NOS OR Fetal adenocarcinoma OR Mucinous cystadenocarcinoma, NOS OR Signet ring cell carcinoma OR SIGNET RING CELL CARCINOMA OR ACINAR CELL CARCINOMA OR Acinar cell carcinoma OR CARCINOMA, NOS OR Carcinoma, NOS OR NON-SMALL CELL CARCINOMA OR Non-small cell carcinoma OR null OR Giant cell and spindle cell carcinoma OR Giant cell carcinoma OR Spindle cell carcinoma, NOS OR Pseudosarcomatous carcinoma OR Basaloid carcinoma OR Mucoepidermoid carcinoma OR MUCOEPIDERMOID CARCINOMA OR Adenosquamous carcinoma OR ADENOSQUAMOUS CARCINOMA OR Pulmonary blastoma OR Carcinosarcoma, NOS OR CARCINOSARCOMA, NOS]
Type(s) of data:	Electronic Health Record
Software for data management:	RStudio Pro (v4.1.0) via Posit Workbench

5.6. Data management

Data management and processing will be done using RStudio Pro (v4.1.0) via Posit Workbench and will proceed through the following phases:

1. Data explorations: During this phase, the investigator will become acquainted with the data source and identify the specific variables that can be used to ascertain eligibility criteria, exposure, outcome, and confounders. These variables and the data tables that they may be found in will be listed in a spreadsheet. No analyses will be conducted outside of simple univariate summary distributions of the variables in the raw data. The purpose of this phase will be to explore missingness patterns and determine the utility of specific variables in key variable (e.g., exposure, outcome, confounder, selection criteria) ascertainment.
2. Cohort selection: Cohort selection will be carried out in a manner specified by the current version of the protocol. Cohort(s) generated in this phase will be used to conduct feasibility analyses that will further inform protocol development or, if no further protocol changes are needed, creation of the analytic file.
3. Analytic file: Person-level and person-time level analytic files will be created for the primary and secondary analyses, respectively. Each variable will be created based on operational definitions provided in this protocol for the exposure, outcome, and confounders of interest.

5.7. Quality control

Upon receiving the data, programmers will visually inspect all data tables to ensure correct formatting and labelling of variables. Summary distributions will be inspected during the "Data explorations" phase of the data management process described in the data management section of this protocol.

Software code written for the purposes of data management and statistical analysis will be reviewed using a formal code review process established at Aetion, which has been adapted from industry standards, best practices, and the experience of internal team members. A central GitHub repository will be used to share and review code. One to two reviewers will be responsible for reviewing each line of code and ensuring correct application of functions and methods (e.g., through spot-checking randomly selected samples of observations before and after application of code).

5.8. Study size and feasibility

5.8.1. CONSORT-style diagram for the primary intention-to-treat cohort

	Patients remaining in cohort (N)	**Patients lost from cohort (N)**
Total Patients in Dataset		
Inclusion Criteria		

Stage 4 or metastatic lung cancer		
Non-small cell histology		
No epidermal growth factor receptor (EGFR) or anaplastic lymphoma kinase (ALK) mutations		
No prior systemic treatment for metastatic disease		
Exclusion Criteria		
Squamous cell morphology		
Biologic cancer therapy for metastatic disease		
Prior non-lung malignancy		
CNS metastases		
Treatment with disease modifying agents, corticosteroids, or immunosuppressive drugs		
Chronic corticosteroids		
Prior treatment with PD-L1 or PD-L2 agent or antibody targeting other immuno-regulatory receptors or mechanisms		
Total Patients that Received Exposure Drug Regimen		
Total Patients that Received Comparator Drug Regimen		

5.8.2. Initial analysis of study cohort, unstratified by exposure (primary intention-to-treat cohort)

An analysis unstratified by the exposure group will be conducted to examine the sample size, distribution of outcome events, follow-up time, and censoring events in the entire study population prior to matching to explore alignment with the KEYNOTE-189 trial and feasibility.

Table 13. Initial analysis of study cohort, unstratified by exposure

Description	**Intention-to-treat cohort (N=)**
Exposure group – number of patients (n, %)	
Comparator group – number of patients (n, %)	
Outcome events (n)	

Risk of outcome event (%)	
Rate of outcome event (n per 1,000 person-years)	
Censoring events (n, %)	
Loss to follow-up (n, %)	
Administrative end of data (n, %)	
End of study period (n, %)	
Follow-up time in days (median [IQR])	
Exposure group (median [IQR])	
Comparator group (median [IQR])	

A power assessment[14] will be conducted in a 1:1 propensity score matched population to determine whether the RWE study is powered to detect a point estimate similar to that observed in the KEYNOTE-189 trial. Only age and sex will be used in this first iteration as previously described.[12] Conservative estimates of statistical power are desired; therefore, power calculations will be conducted on a 1:1 matched population, where some patients will likely be discarded due to the matching process or missing data. The outcome of interest will not be assessed within exposure strata at this stage to eliminate the possibility of a priori knowledge of the outcome informing the study design.

The mortality hazard ratio in the intention-to-treat analysis of the KEYNOTE-189 trial was 0.49, favouring the pembrolizumab-exposed treatment group. There were 410 patients randomized to the pembrolizumab group and 206 patients randomized the placebo-containing group in the trial.[1] This observational study will seek to achieve a power of at least 80% to detect a hazard ratio of the same magnitude as the KEYNOTE-189 trial at a two-sided alpha level of 0.05.

Table 14. Power and sample size in 1:1 propensity score matched population (matched using basic covariates only)

Number of matched patients	
Exposed	
Comparator	
Risk of outcome event	
Desired HR from KEYNOTE-189	
Alpha (2-sided)	

Number of expected outcome events	
Power	

Provide a plot of the propensity score distribution stratified by treatment group before and after matching.
Provide a table of baseline characteristics before and after matching and stratified by treatment group, including absolute standardized differences.

5.8.3. Final power and covariate balance assessment, unstratified by exposure (primary intention-to-treat cohort)

A final power assessment[14] will be conducted in a 1:1 propensity score matched population to determine whether the RWE study is powered to detect a point estimate similar to that observed in the KEYNOTE-189 trial. Numerous covariates will be used in this second iteration of the feasibility analysis as previously described.[12] Not all covariates may be used due to the extent of missing data and matching will be conducted among complete cases only. The outcome of interest will also not be assessed within exposure strata at this stage.

Table 15. Power and sample size in 1:1 propensity score matched population (matched using many covariates)

Number of matched patients	
Exposed	
Comparator	
Risk of outcome event	
Desired HR from KEYNOTE-189	
Alpha (2-sided)	
Number of expected outcome events	
Power	

Provide a plot of the propensity score distribution stratified by treatment group before and after matching.
Provide a table of baseline characteristics before and after matching, including absolute standardized differences and quantification of missing values for each potential confounder variable, stratified by treatment group.

6. Limitation of the methods

This study has several methodologic limitations, as well as limitations that pertain to the study's comparability to the KEYNOTE-189 trial.

As with all epidemiologic studies, the possibility of residual or unmeasured confounding cannot be ruled out. Residual confounding may result from misspecification of statistical models (e.g., when estimating inverse probability treatment and censoring weights). To mitigate the risk of this bias, a flexible functional form of covariates in each model will be used where possible (e.g., including polynomial and interaction terms) to make fewer assumptions about the relation between the independent and dependent variables. An advantage of the data source is that patients' medical histories are available through standardized diagnosis and procedure codes, as well as medication use. Collectively, these histories can be leveraged to ascertain many known prognosticators of mortality to further reduce the risk of unmeasured confounding accordingly. In the oncology population, the possibility of time-varying confounding is an important consideration, as patients' health can rapidly decompensate over study follow-up. Furthermore, cancer treatment regimens tend to vary over time, involving several drugs. To account for treatments and prognosticators that vary over follow-up, a secondary analysis will be conducted using a marginal structural model to estimate a per-protocol effect. This approach addresses time-varying confounding by generating a pseudo-population, in which treatment at each discrete time interval during follow-up is independent of confounder history (under the assumption that all confounders were measured (i.e., "conditional exchangeability"), consistency, positivity, and no model misspecification).[7]

In addition to confounding, selection bias may occur due to the presence of missing data or informative censoring. Missing data may induce selection bias if the mechanism giving rise to missing values cannot be adequately described or proxied by all or some of the observed variables in the analysis (i.e., data are "missing not at random"). Provided that the observed variables do fully relate to missingness (i.e., data are either "missing completely at random" or "missing at random"), several methods are available to correct for potential bias. Once missing data are fully described, including the extent of missingness for key variables in each treatment group, the mechanism underlying missingness will be hypothesized and addressed accordingly. Informative censoring arises when patients who remain within each risk set at a given discrete time during follow-up have a differential risk of the outcome than patients who were censored. Our secondary (per-protocol) analysis plan accounts for this by using inverse probability of censoring weights, under the assumption that all common causes of censoring and the outcome at each discrete time during follow-up are measured and not misspecified in the weight models.

Information bias is a possibility in our analysis. Inaccurate or miscoding of standardized codes in the data could lead to measurement error. Furthermore, in the data source, medication orders, dispensations, and administrations are not always distinguishable. In the present analysis, all such medication records are treated equal, and may not represent the patient's true experience. For this reason, a sensitivity analysis in which medication administrations using HCPCS/CPT codes will be used to define treatment exposures for patients. Notably, the mortality outcome has not been validated and is subject to measurement error, particularly as only month and year or mortality are provided in the data source (we assume date of death is first of the month reported). The pattern of mortality is unlikely to vary across treatment group (i.e., non-differential), which would pose no risk of bias to relative effect measures, such as hazard ratios. However, systematic under-recording may affect absolute effect measures, such as the 12-month survival probabilities.

Although the primary intention-to-treat analysis is the same estimand from the KEYNOTE-189 trial, it is agnostic to adherence. Since adherence may differ substantially between the trial participants and routine clinical practice, a per-protocol effect that accounts for treatment adherence will be conducted in this study. An additional limitation of the primary (intention-to-treat) analysis is the misalignment of treatment initiation with respect to calendar time. We address this with a sensitivity analysis restricting the cohort entry dates (time zero) to be contemporaneous with respect to exposure group.

Lastly, this study is being performed to assess the ability of real-world data in this specific treatment setting to reach similar conclusions to randomized clinical trials using the KEYNOTE-189 trial as a benchmark for causal inference. This study may fail to align with the trial's results for many reasons that are not rooted in bias, including but not limited to differences in study samples (particularly with respect to effect modifiers), random sampling variability, differences in adherence and follow-up, and differences in the measurement of study eligibility criteria, exposures, and outcomes. Therefore, it will be critical to evaluate the results of this study in light of all reasons that could explain any observed differences (or similarities).

7. References

1. Gandhi L, Rodríguez-Abreu D, Gadgeel S, et al. Pembrolizumab plus Chemotherapy in Metastatic Non–Small-Cell Lung Cancer. *N Engl J Med*. 2018;378(22):2078-2092. doi:10.1056/NEJMoa1801005

2. Food and Drug Administration. FDA grants regular approval for pembrolizumab in combination with chemotherapy for first-line treatment of metastatic nonsquamous NSCLC. FDA. Published November 3, 2018. Accessed April 4, 2023. https://www.fda.gov/drugs/resources-information-approved-drugs/fda-grants-regular-approval-pembrolizumab-combination-chemotherapy-first-line-treatment-metastatic

3. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available: Table 1. *Am J Epidemiol*. 2016;183(8):758-764. doi:10.1093/aje/kwv254

4. Chemotherapy in non-small cell lung cancer: a meta-analysis using updated data on individual patients from 52 randomised clinical trials. Non-small Cell Lung Cancer Collaborative Group. *BMJ*. 1995;311(7010):899-909.

5. Scagliotti GV, Parikh P, von Pawel J, et al. Phase III Study Comparing Cisplatin Plus Gemcitabine With Cisplatin Plus Pemetrexed in Chemotherapy-Naive Patients With Advanced-Stage Non–Small-Cell Lung Cancer. *J Clin Oncol*. 2008;26(21):3543-3551. doi:10.1200/JCO.2007.15.0375

6. Schiller JH, Harrington D, Belani CP, et al. Comparison of Four Chemotherapy Regimens for Advanced Non–Small-Cell Lung Cancer. *N Engl J Med*. 2002;346(2):92-98. doi:10.1056/NEJMoa011954

7. Hernán MA, Robins JM. *Causal Inference*: *What If*. Chapman & Hall/CRC; 2020. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

8. Zeng X, Zhu S, Xu C, et al. Effect of Comorbidity on Outcomes of Patients with Advanced Non-Small Cell Lung Cancer Undergoing Anti-PD1 Immunotherapy. *Med Sci Monit Int Med J Exp Clin Res*. 2020;26:e922576. doi:10.12659/MSM.922576

9. Zhao L, Leung LH, Wang J, et al. Association between Charlson comorbidity index score and outcome in patients with stage IIIB-IV non-small cell lung cancer. *BMC Pulm Med*. 2017;17(1):112. doi:10.1186/s12890-017-0452-0

10. Buuren S van. *Flexible Imputation of Missing Data*. CRC Press; 2012.

11. Westreich D, Cole SR, Tien PC, et al. Time scale and adjusted survival curves for marginal structural cox models. *Am J Epidemiol*. 2010;171(6):691-700. doi:10.1093/aje/kwp418

12. Merola D, Campbell U, Gautam N, et al. The Aetion Coalition to Advance Real-World Evidence through Randomized Controlled Trial Emulation Initiative: Oncology. *Clin Pharmacol Ther*. Published online December 7, 2022:cpt.2800. doi:10.1002/cpt.2800

13. Joshua Lin K, Jin Y, Gagne J, et al. Longitudinal Data Discontinuity in Electronic Health Records and Consequences for Medication Effectiveness Studies. *Clin Pharmacol Ther*. 2022;111(1):243-251. doi:10.1002/cpt.2400

14. Chow SC, Shao J, Wang H, eds. *Sample Size Calculations in Clinical Research*. 2nd ed. Chapman & Hall/CRC; 2008.

8. Appendices

Appendix I. Key variable descriptions

Appendix II. Comparison of protocol elements between the real-world evidence study and the KEYNOTE-189 trial

Appendix III. Standardized billing codes used to define key variables

Appendix IV. Missing data: Mechanisms, assumptions, and methodological approach