Emulation of the KEYNOTE-042 (NCT02220894) trial using specialty oncology electronic health records databases

NCT07485179

8th February, 2026

1. Title Page

Title	Emulation of the KEYNOTE-042 (NCT02220894) trial using specialty oncology electronic health records databases
Research question & Objectives	Emulation of the KEYNOTE-042 (NCT02220894) trial, which compared Pembrolizumab to Chemotherapy as the first-line treatment on the risk of death in patients with non-small cell lung cancer.
Protocol version	V1.0
Last update date	February 8, 2026
Contributors	**Primary investigators contact information:** Xiangzhong Xue, Shirley V. Wang **Contributor names:** Yuxin Wang (programmer)
Study registration	**Site:** clinicaltrials.gov **Identifier:** NCT07485179
Sponsor	**Organization:** Food and Drug Administration **Contact:** n/a
Conflict of interest	SVW has been an ad hoc consultant to ICA Group, Cytel Inc, and MITRE a federally funded research and development center for the Centers for Medicare and Medicaid
Protocol repository	Clinicaltrials.gov
Analytic code repository	https://gitlab.partners.org/drugepi/encore/keynote-042 (access only through Mass General Brigham network for authorized personnel)
Quarto study report (including annotated code and output)	https://gitlab.partners.org/drugepi/encore/keynote-042/-/tree/main/public?ref_type=headscc (access only through Mass General Brigham network for authorized personnel)
***encore.io*[1] version**	0.2.0 (see attached documentation *encore.io_0.2.0.pdf*)
encore.analytics[1] version	0.2.0 (https://janickweberpals.github.io/encore.analytics/)
[1] Internally-developed R package to streamline analytics across all available databases and to enhance consistency, transparency and reproducibility in variable definitions and analytic workflows across trial emulations.	

Table of contents

2. Abstract

This trial emulation study aims to emulate the KEYNOTE-042 trial (NCT02220894) using real-world specialty oncology electronic health record data and to investigate the concordance between the trial's original and emulated treatment effect estimates on overall survival (OS). KEYNOTE-042 was Phase III, open-label, randomised study assessing the efficacy and safety of pembrolizumab monotherapy (200 mg intravenously every 3 weeks for up to 35 cycles) versus investigator's choice of platinum-based chemotherapy (carboplatin plus paclitaxel or pemetrexed for 4–6 cycles, with optional pemetrexed maintenance for non-squamous histology) in patients with previously untreated advanced or metastatic non-small-cell lung cancer (NSCLC) without sensitising EGFR mutations or ALK translocations, and whose tumours expressed programmed death-ligand 1 (PD-L1) with a tumour proportion score (TPS) of 1% or greater.

3. Amendments and updates

Version date	Version number	Section of protocol	Amendment or update	Reason
February 8, 2026	V1.0	NA	Initial version	NA

4. Rationale and background

Randomized controlled trials (RCTs) are generally regarded as the gold-standard of evidence for establishing efficacy of medical products. However, real-world data (RWD) are increasingly used to complement evidence from RCTs. Yet, to have confidence in the accuracy of non-interventional studies medical products and their outcomes in oncology, investigators need to know what questions can be validly answered, with which non-interventional study designs, and which analysis methods are appropriate, given the data that is available. Building on a process from the RCT DUPLICATE initiative,[1–4] **EmulatioN of Comparative Oncology trials with Real-world Evidence** (**ENCORE)** is the trial emulation discussed in this protocol, which is part of the expansion project specific to oncology and aims to emulate 12 randomized oncology RCTs using multiple EHR data sources.

The purpose of this protocol is to describe the emulation of the KEYNOTE-042.[5] KEYNOTE-042 was Phase III, open-label, randomised study assessing the efficacy and safety of pembrolizumab monotherapy (200 mg intravenously every 3 weeks for up to 35 cycles) versus investigator's choice of platinum-based chemotherapy (carboplatin plus paclitaxel or pemetrexed for 4–6 cycles, with optional pemetrexed maintenance for non-squamous histology) in patients with previously untreated advanced or metastatic non-small-cell lung cancer (NSCLC) without sensitising EGFR mutations or ALK translocations, and whose tumours expressed programmed death-ligand 1 (PD-L1) with a tumour proportion score (TPS) of 1% or greater. The trial had 3 co-primary analyses that focused on patients with PD-L1 TPS ≥50%, PD-L1 TPS ≥20%, and PD-L1 TPS ≥1%, respectively.

The hazard ratio (HR) for overall survival was 0.69 (95% CI, 0.56–0.85) among patients with PD-L1 TPS ≥50%, 0.77 (95% CI, 0.64–0.92) among those with PD-L1 TPS ≥20%, and 0.81 (95% CI, 0.71–0.93) among those with PD-L1 TPS ≥1%. We will focus on the PD-L1 ≥50% subgroup as the primary analysis for this emulation because the vast majority of patients treated with pembrolizumab in clinical practice have PD-L1 ≥50%, The median overall survival was 16.7 months (95% CI 13.9–19.7) in the pembrolizumab group compared to 12.1 months (95% CI 11.3–13.3) in the chemotherapy arm. Pembrolizumab monotherapy received its first-line NSCLC

approval for the PD-L1 high-expression (TPS 50%) population based on KEYNOTE-024 on October 24, 2016, which had established a significant survival benefit for pembrolizumab monotherapy in previously untreated patients with a PD-L1 TPS expression of 50% or greater. Following the results of KEYNOTE-042, the FDA expanded the indication to include the first-line treatment of patients with locally advanced or metastatic NSCLC whose tumours express PD-L1 with a TPS of 1% or greater on April 11, 2019.

The 5-year pooled OS endpoint was published in the Journal of Clinical Oncology on October 28, 2022.[5]

5. Research question and objectives

The primary and secondary research question were summarized in Table 1.

A. Primary research question and objective

Table 1. Primary and secondary research questions and objective.

Objective:	To compare the overall survival [OS] among patients with advanced or metastatic NSCLC and PD-L1 TPS ≥50% who initiated pembrolizumab versus chemotherapy.
Hypothesis:	Initiation of pembrolizumab improves overall survival time as compared to initiation of chemotherapy
Population ***(mention key inclusion-exclusion criteria):***	• Age at least 18 years • Subjects with histologically or cytologically confirmed advanced or metastatic NSCLC that is not amenable to curative-intent treatment • ECOG 0 or 1 • No EGFR sensitizing (activating) mutation or an ALK translocation • PD-L1 ≥50% (Although KEYNOTE-042 reported three primary overall survival outcomes among patients with PD-L1 ≥50%, ≥20%, and ≥1%, we will focus on the PD-L1 ≥50% subgroup as the primary analysis for this emulation because the vast majority of patients treated with pembrolizumab in clinical practice have PD-L1 ≥50% and the trial results suggested heterogeneity by PD-L1 level.)

Exposure:	Initiation of pembrolizumab
Comparator:	Initiation of chemotherapy (carboplatin plus paclitaxel or pemetrexed for 4–6 cycles)
Outcome:	Time to all-cause mortality (OS)
Time *(when follow up begins and ends):*	From the day of the end of the treatment assessment window[a] until death or last observed clinical activity/last sign of the patient being alive or data cut-off, whichever occurred earlier
Setting:	1L advanced NSCLC
Main measure of effect:	Primary: Hazard ratio (95% CI) for overall survival Secondary: median overall survival time (difference) in % (95% CI)

aThe treatment assessment window is defined as the period from first-line treatment initiation to a vendor-specified time point in order to fully capture the first-line regimen for advanced NSCLC. The emulation of the main protocol elements of the KEYNOTE-042 is illustrated side by side in **Table 2**.

Table 2. Trial emulation table summarizing the main protocol elements of the KEYNOTE-042 trial and the planned emulation.

Protocol component	KEYNOTE-042 RCT	Emulation
Eligibility Criteria	Age ≥18 years at randomization	Age ≥18 years at treatment initiation
	No prior systemic chemotherapy for advanced/metastatic NSCLC	Patients with documentation of prior chemotherapy administration for advanced/metastatic NSCLC were excluded
	ECOG performance status 0–1	ECOG 0–1 (if available/ascertainable in the database)
	Histologically or cytologically confirmed advanced/metastatic NSCLC that is not amenable to curative-intent treatment	Line of therapy setting classified as “advanced” (EDB1) or “metastatic” (EDB2), or evidence of metastatic disease at treatment initiation (EDB4)
	No EGFR sensitizing mutation or ALK translocation	Pembrolizumab group: patients with documented EGFR/ALK positivity were excluded[a]; Chemotherapy group: patients with documented EGFR/ALK positivity or missing/unknown EGFR/ALK status were excluded
	PD-L1 positive	PD-L1 ≥50% required[b]; exclude if PD-L1 is missing/unknown or <50%.
	No investigational agent within 4 weeks prior to the first dose of trial treatment	Patients with any documentation of an investigational agent within 4 weeks prior to initiation of first-line pembrolizumab/chemotherapy were excluded
	No prior carboplatin + paclitaxel for squamous patients in the adjuvant setting	Squamous patients with any documentation of prior carboplatin plus paclitaxel were excluded
	No chemotherapy or biologic therapies within 3 weeks prior to the first dose of trial treatment	Patients with any documentation of chemotherapy[c] or biologic therapy[d] within 3 weeks prior to initiation of first-line pembrolizumab/chemotherapy were excluded
	No immunotherapy prior to the first dose of trial treatment	Patients with documentation of prior immunotherapy[e] administration were excluded
	No active malignancy or prior malignancy with the exception of basal cell carcinoma of the skin, superficial bladder cancer, squamous cell carcinoma of the skin, or in situ cervical cancer, or no evidence of that disease recurrence ≥ 5 years since initiation of that therapy.	Patients with any prior non-lung malignancy diagnosis (exceptions: basal cell carcinoma of the skin, superficial bladder cancer, squamous cell carcinoma of the skin, and in situ cervical cancer) were excluded

	No active/known/suspected autoimmune disease[f]	Patients with documented autoimmune diseases[f] within 2 years prior to treatment initiation were excluded
	No interstitial lung disease	Patients with documented interstitial lung disease were excluded
	No CNS metastases	Patients with documented CNS metastases were excluded
Treatment strategies	pembrolizumab versus chemotherapy	pembrolizumab versus chemotherapy
Assignment procedures	Random assignment to receive either pembrolizumab or chemotherapy in a 1:1 ratio	Randomization emulated via 1:1 matching of patients who initiate pembrolizumab as opposed to chemotherapy based on a propensity score that includes major risk factors for overall survival.
Follow-up period	Time from randomization until death from any cause or censoring at the last contact that the patient was known to be alive.	Time from the end of the treatment assessment window[g] defined by the vendors business rules to identify line of therapy until death from any cause or censoring of patients who did not die on the basis of the last recorded date that the patient was known to be alive
Outcome	Overall survival	Overall survival
Causal contrast	Intent-to-treat effect	As started (observational analogue of intent-to-treat)

[a] Patients with missing or unknown EGFR/ALK status were not excluded in pembrolizumab armunder the assumption that patients treated with pembrolizumab met the indication of having no EGFR mutation or ALK translocation.

[b]KEYNOTE-042 reported three primary overall survival outcomes among patients with PD-L1 ≥50%, ≥20%, and ≥1%. We will focus on the PD-L1 ≥50% subgroup as the primary analysis for this emulation because the vast majority of patients treated with pembrolizumab in clinical practice have PD-L1 ≥50% and the trial results suggested heterogeneity by PD-L1 level.

[c]Includes the following: Cisplatin, Carboplatin, Pemetrexed, Paclitaxel, Nab-paclitaxel, Docetaxel, Gemcitabine, Vinorelbine, Etoposide

[d]Includes the following: Bevacizumab, Ramucirumab, Necitumumab, Cetuximab

[e]Includes the following: Ipilimumab, Pembrolizumab, Pembrolizumab, Atezolizumab, Durvalumab, Avelumab, Cemiplimab, Tremelimumab, Dostarlimab

[f]Includes the following: Inflammatory bowel disease, Systemic lupus erythematosus, Dermatomyositis, Scleroderma, Vasculitis, Polyarteritis nodosa, Sarcoidosis, Immune thrombocytopenic purpura, Hemolytic anemia, Multiple sclerosis

[g]Proprietary business rules to define initiation of the line of therapy (including exposure assessment window) cannot be shared.

EDB1 =ENCORE DataBase 1; EDB2 =ENCORE DataBase 2; EDB4 =ENCORE DataBase 4; ECOG = Eastern Cooperative Oncology Group; CNS = central nervous system

6. Research methods

6.1. Data sources

6.1.1. Context and rationale for data sources

The overall ENCORE project uses data from a total four different oncology-specific electronic health records (EHR)-derived data sources: ConcertAI, COTA, Flatiron Health, McKesson/Ontada. For ENCORE, not all databases are available for each cancer indication and the names of the databases will henceforth be blinded and referred to as ENCORE DataBase (EDB) 1, 2, 3 and 4 (the numbering does not coincide with the above order of mention of the databases).

For this trial emulation, NSCLC-specific data are available for EDB1, EDB2 and EDB4. The fitness-for-purpose of the data for the given trial emulation were assessed and considered for the final selection of the databases.

Reason for selection: All considered databases draw from a comprehensive national sample of patients with cancer in the US with detailed EHR-derived information on the information necessary to study medication effectiveness in oncology.

Strengths of data source(s): Size and detailed clinical information on oncology-specific variables and outcomes (validated composite all-cause mortality sourced from different data sources[6,7]).

Limitations of data source(s): General limitations across all data sources include missing data, potential lack of data continuity, heterogeneous data provenance, quality/heterogeneous ascertainment of mortality endpoint data and the variability in how line of treatment is captured and curated (a more comprehensive discussion of the data sources and approaches is provided in section 7. After a comprehensive assessment of all data sources regarding their fitness for the purpose of emulating the KEYNOTE-042 trial, EDB2 was determined insufficient to be included in the main and sensitivity analyses and EDB4 was determined insufficient to be included in the main analyses. In addition, the primary analysis will focus on patients with PD-L1 ≥50% in EDB1, for the reasons described below.

- Rationale for excluding EDB2 from the main and sensitivity analyses: After applying the inclusion/exclusion criteria, EDB2 results in a cohort with a very small sample size after propensity score matching (18 patients in each group) an measured risk factors remained unbalanced after matching (Figure 6). Another consideration was the lack of availability of important prognostic covariates, which may contribute to biased effect estimates due to unmeasured confounding.

- Rationale for excluding EDB4 from the main analyses: As shown in Table 3, the patient identification period in EDB4 starts on 10/01/2018 which is after the official approval of pembrolizumab for 1L metastatic NSCLC. This means that a design which includes patients who initiate treatment before this date will suffer from immortal time bias (see Figure 11). Although we could restrict to patients who enter the cohort after 2018, there was an additional concern that in EDB4, the capture of key PD-L1 details needed for covariate adjustment was less comprehensive than in other databases. PD-L1 is recorded only as a binary variable (positive/negative), without quantitative PD-L1 levels, and assay/IHC type could not be verified. For these reasons, EDB4 will be used only in a sensitivity analysis designed to evaluate the potential impact of selection/immortal time bias.
- Rationale for focusing on PD-L1 ≥50% patients in EDB1 in main analyses: PD-L1 expression is an important treatment effect modifier of pembrolizumab in advanced NSCLC. After applying eligibility criteria, the distribution of PD-L1 expression differed significantly between KEYNOTE-042 and EDB1, with patients with PD-L1 ⩾50% comprising 47% of the trial population, compared with 70% of the eligible EDB1 population. Therefore, including the full PD-L1 ≥1% population in the emulation would bias the estimated treatment effect in favor of pembrolizumab.[5] Moreover, in real-world practice, patients with PD-L1 1%–49% are typically treated with chemoimmunotherapy unless they are considered too frail for chemotherapy.[8] Those treated with pembrolizumab monotherapy in this subgroup are likely highly selected, further increasing the potential for residual confounding

Data source provenance/curation: In brief, all databases provide EHR-derived oncology-specific patient-level information which are either derived directly (e.g., through structured data fields and dropdown menu selections) from EHR and/or undergo semi-automated abstraction processes from unstructured reports. The detailed data provenance, abstraction processes and implemented business rules to curate and prioritize certain variables may vary by database and can be found in legacy publications by the data partners.

Table 3. Metadata about data sources and software.

	EDB1	EDB2	EDB4
Data Source(s):	EHR-derived	EHR-derived	EHR-derived
Study Period:	Patient identification period: 01/01/2011-04/30/2024 with follow-up information through data cut-off date on 04/30/2024	Patient identification period: as of 2/20/2022 with follow-up information through data cut-off date on 02/24/2023	Patient identification period: 10/01/2018-09/30/2023 with follow-up information through data cut-off date on 09/30/2023.
Eligible Cohort Entry Period:	After the date when pembrolizumab was approved by the FDA (October 24, 2016)	After the date when pembrolizumab was approved by the FDA (October 24, 2016)	After the date when pembrolizumab was approved by the FDA (October 24, 2016)
Data Version (or date of last update):	Delivery: Jul 11, 2024	Delivery: Sep 8, 2023	Delivery: Oct 24, 2023

		Updated (LoT addition): Mar 11, 2024	Updated (demographics): Feb 29, 2024
Data sampling/extraction criteria:	Patients are sampled if they have a confirmed diagnosis of advanced NSCLC via abstraction on or after 1 Jan 2011, and at least 2 EHR visits on or after 1 Jan 2011. Both ICD-9 (162.x) and ICD-10 (C34x or C39.9) codes are used for the initial selection, and advanced diagnosis are then confirmed via abstraction (since ICD codes do not specify advanced diseases).	Patients are sampled if they were diagnosed with NSCLC and do not meet any of the following exclusion criteria: patient is <18 years of age at the time of diagnosis, does not have the malignancy of interest, is not evaluated at the accessible provider site for the malignancy of interest, has concurrent primary malignancies, has no date of diagnosis in EHR, patient chart has no clinician note available in the EHR, is diagnosed with a malignancy after the diagnosis of the malignancy of interest prior to evaluation for the malignancy of interest at the accessible provider site, is only seen once at the provider site for the malignancy of interest, is initially misdiagnosed and treated for another malignancy, but was later confirmed to have the malignancy of interest, is on therapy, active surveillance, or observation for a malignancy diagnosed prior to the diagnosis of the malignancy of interest at the time of diagnosis of the malignancy of interest, is metastatic (includes leukemias and multiple myeloma) for a malignancy diagnosed prior to the malignancy of interest, with	NSCLC Patients with an office visit in the reporting period will be included in the report with full patient history. Patients are sampled if they were diagnosed with NSCLC and with a documented visit date, within the defined reporting period, to one of the facilities and were at least 20 years of age at the time of first diagnosis. Patients who were on a clinical trial at any point in their treatment history are excluded.

		the presence of low grade (inclusive of grades 1 and 2) neuroendocrine histology	
Type(s) of data:	EHR-derived	EHR-derived	EHR-derived
Data linkage[1]:	Mortality/date of death is a composite endpoint of structured and unstructured data from the EHR, obituary data, and the social security death index	Mortality/date of death is a composite endpoint of structured and unstructured data from EHRs and commercially available obituary data including the Social Security Administration death master file	Mortality/date of death is a composite endpoint of structured and unstructured EHR data, supplemented with commercially available claims data, obituary data, and the Social Security Administration death master file
Conversion to CDM[2]:	No	No	No
Software for data management:	R 4.3.2	R 4.3.2	R 4.3.2

[1] Mortality/date of death is a composite endpoint that is often derived from various linked sources including social security death index/ Social Security Administration death master file, obituary data and EHR records

[2] CDM = Common Data Model

6.2. Data management

Data is stored on secure Mass General Brigham corporate provisioned and backed up servers physically located in our Mass General Brigham corporate data centers. Mass General Brigham corporate data centers are designed to insure availability of the affiliated hospitals' and research applications and IT systems in the event of a disaster. The Division follows Mass General Brigham workstation requirements which include: encryption at rest, up-to-date malware protection including antivirus, spyware detection and removal tools, Crowdstrike End Point protection installed, devices enrolled in enterprise Mobile Device Management (MDM) solution as appropriate, any laptop/computer used for business purposes must not be shared with family, friends, or other unauthorized individuals, and compliance with enterprise Password Requirements. Only authorized personnel have read-only access to raw data files.

Cleaned and analysis-ready datasets, i.e., +/- imputed one-row-per-patient tables with all required exposure, outcome and covariate variables, are stored in separate sub-directories dedicated for the specific emulated trial.

6.3. *Quality control*

Upon delivery, data quality procedures included checks on delivered tables and variables, per table checks, descriptives on most important measures such as demographic and stage distributions by sex at time of initial diagnosis, regimen/exposure frequency counts and time-trends and overall survival benchmarks against literature and general cancer registry statistics. The R code to reproduce the quality assessments is deposited on the Mass General Brigham-provisioned GitLab server https://gitlab.partners.org/drugepi/encore/quality (repository is only accessible within the Mass General Brigham network and additionally only to authorized study personnel).

6.4. *Study design*

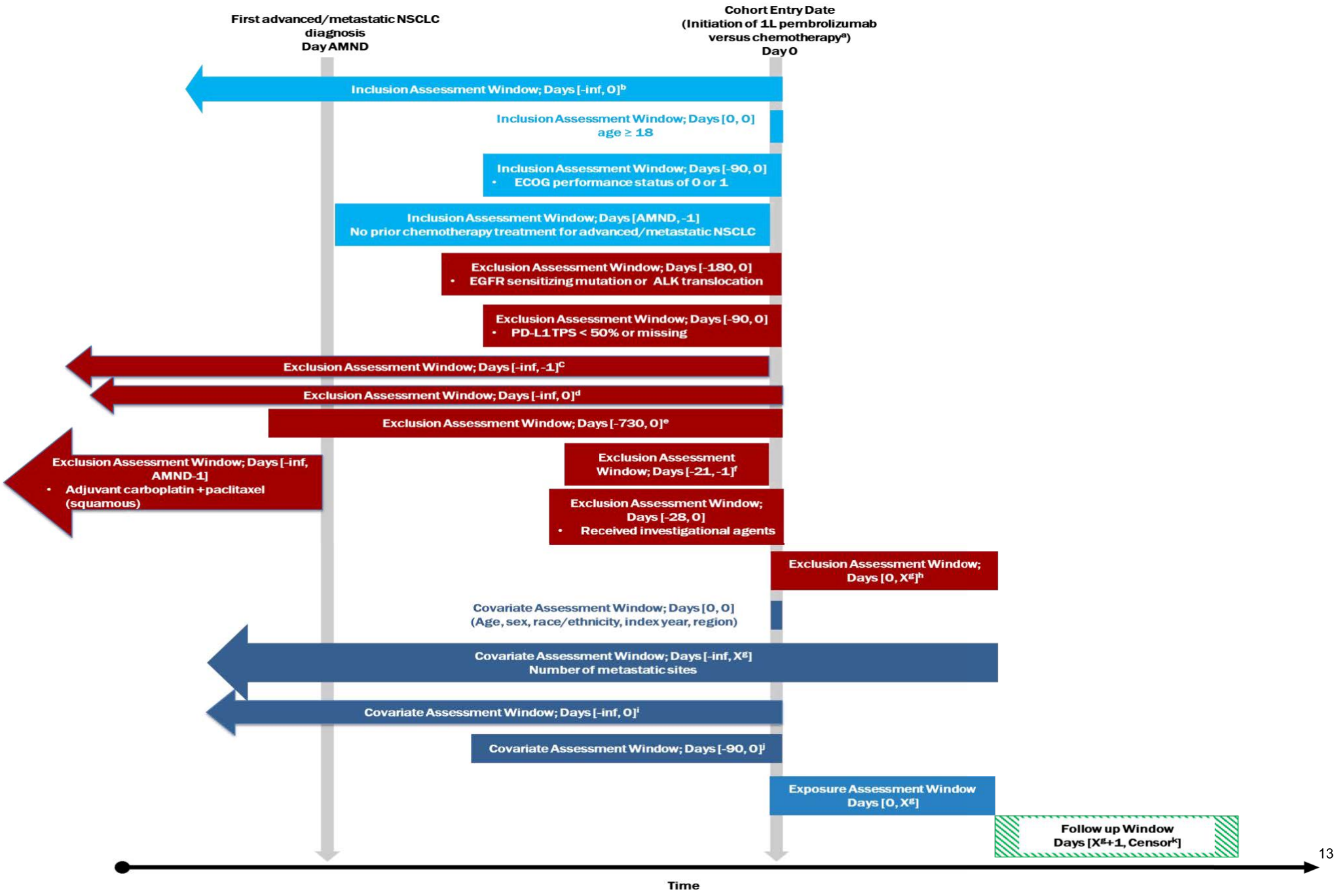
Research design (e.g. cohort, case-control, etc.): Cohort study

Rationale for study design choice: Resembles the principles of the (target) trial emulation framework.[9]

6.5. *Study design diagram*

Figure 1 depicts study design and variable measurement considerations for the emulation of the KEYNOTE-042 trial. The selection of key confounders/prognostic factors is driven by expert knowledge and additionally based on covariates included in the real-world prognostic score (ROPRO), which is a published and validated pan-tumor and cancer-specific prognostic score framework for overall survival.[10–14]

Figure 1. Study design illustration for KEYNOTE-042 trial emulation.


First advanced/metastatic NSCLC diagnosis
Day AMND
Cohort Entry Date
(Initiation of 1L pembrolizumab versus chemotherapy[a])
Day 0
Inclusion Assessment Window; Days [-inf, 0][b]
Inclusion Assessment Window; Days [0, 0]
age ≥ 18
Inclusion Assessment Window; Days [-90, 0]
• ECOG performance status of 0 or 1
Inclusion Assessment Window; Days [AMND, -1]
No prior chemotherapy treatment for advanced/metastatic NSCLC
Exclusion Assessment Window; Days [-180, 0]
• EGFR sensitizing mutation or ALK translocation
Exclusion Assessment Window; Days [-90, 0]
• PD-L1 TPS < 50% or missing
Exclusion Assessment Window; Days [-inf, -1][c]
Exclusion Assessment Window; Days [-inf, 0][d]
Exclusion Assessment Window; Days [-730, 0][e]
Exclusion Assessment Window; Days [-inf, AMND-1]
• Adjuvant carboplatin + paclitaxel (squamous)
Exclusion Assessment Window; Days [-21, -1][f]
Exclusion Assessment Window; Days [-28, 0]
• Received investigational agents
Exclusion Assessment Window; Days [0, X[g]][h]
Covariate Assessment Window; Days [0, 0]
(Age, sex, race/ethnicity, index year, region)
Covariate Assessment Window; Days [-inf, X[g]]
Number of metastatic sites
Covariate Assessment Window; Days [-inf, 0][i]
Covariate Assessment Window; Days [-90, 0][j]
Exposure Assessment Window
Days [0, X[g]]
Follow up Window
Days [X[g]+1, Censor[k]]
Time

a. Carboplatin + paclitaxel or pemetrexed

b. Advanced or metastatic NSCLC that is not amenable to curative-intent treatment; no history of non–lung malignancies (except BCC/SCC of skin, superficial bladder cancer, or cervical CIS)

c. Prior immune checkpoint inhibitor therapy, including ipilimumab, nivolumab, pembrolizumab, atezolizumab, durvalumab, avelumab, cemiplimab, tremelimumab, or dostarlimab

d. History of interstitial lung disease, central nervous system metastases, or carcinomatous meningitis

e. History of autoimmune disease, including inflammatory bowel disease, systemic lupus erythematosus, dermatomyositis, scleroderma, vasculitis, polyarteritis nodosa, sarcoidosis, immune thrombocytopenic purpura, hemolytic anemia, or multiple sclerosis

f. Prior cytotoxic chemotherapy or biologic therapy, including cisplatin, carboplatin, pemetrexed, paclitaxel, nab-paclitaxel, docetaxel, gemcitabine, vinorelbine, etoposide, bevacizumab, ramucirumab, necitumumab, or cetuximab

g. Proprietary business rules to define initiation of the line of therapy (including exposure assessment window) cannot be shared.

h. Systemic anticancer therapy other than pembrolizumab and chemotherapy

i. De novo metastatic status, time from initial diagnosis to T0, time from first evidence of metastatic disease to T0, smoking, family history, race/ethnicity, etc.

j. Labs (albumin, hemoglobin, etc.) and vitals (BMI, etc.) that are part of the ROPRO prognostic score; see Becker, Weberpals, et al. Ann Oncol 2020 (doi: 10.1016/j.annonc.2020.07.013)

Intention-to-treat: death due to any reason or last observed clinical activity/sign of patient being alive or data cut-off date (whichever occurred earlier)

AMND = advanced/metastatic non-small cell lung cancer diagnosis; 1L = First-line antineoplastic systemic therapy; BCC = basal cell carcinoma; SCC = squamous cell carcinoma; CIS = carcinoma in situ; TPS = tumor proportion score; EGFR = epidermal growth factor receptor; ALK = anaplastic lymphoma kinase; ILD = interstitial lung disease

6.6. *Setting*

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

Time 0 in this database study is defined as the date a patient initiated pembrolizumab (exposure) or chemotherapy (comparator) as 1L systemic antineoplastic treatment for advanced or metastatic NSCLC. This aims to emulate the date of randomization and cohort entry in the RCT (the time from randomization to first dose is not reported in the clinicaltrials.gov study reports (KEYNOTE-042) or the trial articles).

6.6.2. Context and rationale for study inclusion criteria

Study inclusion criteria were defined to emulate all key inclusion criteria for the trial that were deemed both clinically relevant and measurable in EHR data. See Excel appendix table 1 (Table1_I_E) for a one-by-one evaluation. A summary of the operational definitions of the inclusion criteria that were applied for each database can be found in the Excel appendix table 2. A flowchart of the study cohort assembly is provided in Appendix 3.

6.6.3. Context and rationale for study exclusion criteria

Study exclusion criteria were defined to emulate all key exclusion criteria for the trial that were deemed both clinically relevant and measurable in EHR data. See Excel appendix table 1 (Table1_I_E) for a one-by-one evaluation. A summary of the operational definitions of the exclusion criteria that were applied for each database can be found in the Excel appendix table 2. A flowchart of the study cohort assembly is provided in Appendix 3.

6.6.4. Context and rationale for exposure(s) of interest

The exposure and comparator were defined to emulate the agents compared in the trial—pembrolizumab versus chemotherapy in a 1L advanced setting.

- **EDB1**: Exposure is derived using a manually curated line of therapy (LOT) table provided by the data partner that programmatically categorizes treatment regimens into a coherent line of treatment. That is, each patient is represented with one row per curated line of therapy with corresponding information on line number, regimens, as well as start and end dates. Based on this table, patients who received pembrolizumab or chemotherapy treatment regimens are identified by their generic names (string match) in 1L. The LOT implicitly only considers regimens that were given in an advanced disease setting. More details and annotated code to identify initiators can be found in the 'Derive cohort EDB1' Quarto report (access within MGB network only).
- **EDB2**: Exposure is derived using a manually curated line of therapy (LOT) table provided by the data partner that programmatically categorizes treatment regimens into a coherent line of treatment. That is, each patient is represented with one row per curated line of therapy with corresponding information on line number, regimens, as well as start and end dates. Based on this table, patients who received pembrolizumab or chemotherapy treatment regimen by their generic names (string match) in 1L are identified. The LOT implicitly only considers regimens that were given as part of a metastatic disease setting. More details and annotated code to identify initiators can be found in the 'Derive cohort EDB2' Quarto report (access within MGB network only).

- **EDB4:** For the EDB4 database, the following logic is applied.
    - Identify patients with evidence of a metastasis from the diagnosis table in which the earliest date associated with evidence of metastasis is captured as a structured field (metastasis date).
    - Identify all potential antineoplastic drugs typically used in advanced/metastatic NSCLC (see list below*). Only these are considered.
    - Identify patients who initiated the KEYNOTE-042 regimens as their first-line antineoplastic treatment after the date of metastasis. Treatment line identification is established according to the business rules set by the data vendor.

***Antineoplastic drugs considered:** adagrasib, ado-trastuzumab emtansine, afatinib, alectinib, amivantamab-vmjw, atezolizumab, atezolizumab and hyaluronidase-tqjs, bevacizumab, bevacizumab-adcd, bevacizumab-awwb, bevacizumab-bvzr, bevacizumab-maly, bevacizumab-tnjn, binimetinib, brigatinib, cabozantinib, capmatinib, carboplatin, cemiplimab-rwlc, ceritinib, cetuximab, cisplatin, crizotinib, dabrafenib, dacomitinib, datopotomab, chemotherapy, durvalumab, encorafenib, ensartinib, entrectinib, erdafitinib, erlotinib, etoposide, fam-trastuzumab deruxtecan-nxki, gefitinib, gemcitabine, ipilimumab, larotrectinib, lazertinib, lorlatinib, mobocertinib, pembrolizumab, pembrolizumab-hyaluronidase, osimertinib, paclitaxel, paclitaxel protein bound, pembrolizumab, pembrolizumab and berahyaluronidase alfa-pmph, pemetrexed, pralsetinib, ramucirumab, repotrectinib, selpercatinib, sotorasib, sunvozertinib, taletrectinib, telisotuzumab vedotin-tllv, tepotinib, trametinib, tremelimumab-actl, vandetanib, vemurafenib, vinorelbine, zenocutuzumab-zbco, zongertinib.

6.6.5. Context and rationale for outcome(s) of interest

The primary outcome for the database study was defined to emulate the primary outcome for the trial, time from index to death due to any reason (overall survival). Operational definitions:

- **EDB1**: Time in {days, months and years} from index date to death due to any reason. The date of death is de-identified to month-level granularity or (rarely) to year-level granularity and the date of death is therefore imputed to the 15th of a month or mid-year/July 2 of the year of death, respectively. If there is no indication that a patient died during the study period, the patient is censored. The censoring date is defined as the last visit or treatment encounter or data cut-off date, whichever occurred earlier. The overall survival endpoint is operationalized using a parameterized R function *edb1_get_os()* and more details can be found in the attached pdf documentation.
- **EDB2**: Time in {days, months and years} from index date to death due to any reason. If there is no indication that a patient died during the study period, the patient is censored. The censoring date is defined as the last observed activity date or data cut-off date, whichever occurred earlier. Activity dates are defined as documented in Table 4. Dates used to derive time to all-cause mortality may have some associated imprecision such that the date of death is either known completely, the year and month is known or only the year is known. The overall survival endpoint is operationalized using a parameterized R function *edb2_get_os()* and more details can be found in the attached pdf documentation.

Table 4. Relevant clinical activities considered to derive last activity date for censoring.

Table / clinical activity considered	Dates considered
Adverse events	Event date
Therapy (cellular, systemic, radiation, surgery)	Start and end dates or declined intervention date, surgery date, assessed resection dates
Palliative care referral	Referral date
Visits	Contact/visit date
Vitals	Assessed date
Labs	Lab date
Biomarkers	Specimen collection date
Patient observation period	Start and end dates
Demographics	Date of most recent contact with provider, date patient was diagnosed with a second primary malignancy
Performance assessments	Documented date, reported date
Secondary diagnoses	Diagnosis date
Progression, histology, lymphovascular invasion, metastatic sites, pancoast tumor, perineural invasion	Assessed date
Stage/TNM	Assessed date
Smoking	Assessed date

- EDB4: Time in {days, months and years} from index date to death due to any reason. The date of death is de-identified to month-level granularity and the day of death is therefore imputed to the 15th of a month. If there is no indication that a patient died during the study period, the patient is censored. The censoring date is defined as the last date of vital signs recorded as proof that the patient was alive at that time (de-identified to week-level granularity) or

data cut-off date, whichever occurred earlier. The overall survival endpoint is operationalized using a parameterized R function *edb4_get_os()* and more details can be found in the attached pdf documentation.

6.6.6. Context and rationale for follow up

Only intention-to-treat (ITT) analyses will be conducted. Although cross-over from the comparator to the exposure can be expected, which biases the exposure treatment effect more towards the null, this also applies to the RCT. Although the KEYNOTE-042 protocol prohibited formal crossover from the chemotherapy group to pembrolizumab, 126 patients (20%) in the chemotherapy arm received subsequent immunotherapy after the trial's treatment discontinuation.[15]

An as-treated analysis is not considered since in the context of oncology, reasons for discontinuation usually are due to toxicity, death or progression/non-response to the current treatment, all of which are highly correlated with the outcome under study which would hence lead to bias due to informative censoring.

6.6.7. Context and rationale for covariates

We identified a series of covariates that are strong prognostic factors for the outcome and auxiliary covariates which may be useful to impute missing data. Such covariates comprise demographics, covariates indicating disease-severity, cancer-specific covariates as well as pathological and genetic factors. In addition, selected labs and vitals are considered since they were shown to carry a high amount of prognostic information as described in Becker, Weberpals, et al.[10] For these variables, additional plausibility checks and transformations are carried out. In detail, labs and vitals are individually checked if they cross a certain biologically implausible threshold (e.g., a heart rate of 0) in which cases the values are set missing and imputed in a next step. These thresholds were compiled by experienced practicing physicians and medical oncologists and are listed in appendix **Table 10** and **Table 11**.

Note that not all covariates are available across all databases used for this trial emulation. In the analytical stage, the most comprehensive model will be fit for each database individually.

Table 5. Operational definitions of key covariates used for trial emulation.

Characteristic	Harmonized analysis variable name	R function to derive covariate (see pdf in appendix)	Details[1]	Variable encoding	Assessment window
Age at initial diagnosis	dem_age_initial_diagnosis	edbx_get_demographics()	Age measured at initial diagnosis of eligible primary tumor	Nominal (<60, 60-69, 70-79, 80+)	[-inf;0] at initial diagnosis of primary cancer

Characteristic	Harmonized analysis variable name	R function to derive covariate (see pdf in appendix)	Details[1]	Variable encoding	Assessment window
Age at index date	dem_age_index	edbx_get_demographics()	Age measured at index date	Binary (<60, 65+); modelled continuously in ROPRO[10]	[0;0]
Sex	dem_sex	edbx_get_demographics()	Sex	Binary (Male, Female)	[0;0]
Year of index date	c_year_index	De novo derived from dt_index	Calendar year in which patient initiated study treatment	Nominal (<2018, 2018+)	[0;0]
Family history	dem_family_history	edbx_get_demographics()	Family history of cancer	Logical (TRUE, FALSE)	[0;0] (no specific date is associated)
Race	dem_race	edbx_get_demographics()	Race categorized as in the original RCT	Binary (Asian vs non-Asian)	[0;0]
Ethnicity	dem_ethnicity	edbx_get_demographics()	Ethnicity	Hispanic, Non-Hispanic	[0;0]
Region	dem_region	edbx_get_demographics()	US region patient receives care in; if given on a state level, region is manually mapped (see Table 12)	Nominal (Northeast, South, West, Midwest)	[0;0]
Practice type	dem_practice	edbx_get_demographics()	Setting patient is receiving care at	Nominal (academic, community, academic & community)	[-inf;0]
Socio-economic status	dem_ses	edbx_get_demographics()	Socioeconomic status (SES) index based on residence area of patient	Nominal (from '1 - Lowest SES' through '5 - Highest SES)	[-inf;0]

Characteristic	Harmonized analysis variable name	R function to derive covariate (see pdf in appendix)	Details[1]	Variable encoding	Assessment window
Smoking	c_smoking_history	edbx_get_demographics()	History of current or former (= TRUE) or never (= FALSE) smoking on or anytime before index date; if there are multiple records per patient, any evidence of former/current smoking is prioritized	Binary logical (TRUE, FALSE)	[-inf;0]
PD-L1 expression status	c_pdl1_status	EDB1, EDB2: edbx_get_pdl1() EDB4: edb4_get_biomarker()	Evidence of PD-L1 expression level. If a patient has multiple PD-L1 measurements, the measurement closest to the index date is prioritized.	EDB1, EDB2: PD-L1 expression was reported as a percent staining value (TPS, %) or as a range (e.g., 1–10%). we harmonize PD-L1 into three trial-aligned groups: <1%, 1–49%, and ≥50% EDB4: Binary (negative, positive)	[-90;0]
PD-L1 assay type/scoring metric	c_pdl1_status	EDB1, EDB2: edbx_get_pdl1()	Evidence of PD-L1 expression scoring metric. If a patient has multiple PD-L1 measurements, the measurement closest to the index date is prioritized.	EDB1: IHC assay type (e.g., 22C3, 28-8, SP142, SP263); EDB2: PD-L1 scoring metric (e.g., TPS, IC, or TC)	

Characteristic	Harmonized analysis variable name	R function to derive covariate (see pdf in appendix)	Details[1]	Variable encoding	Assessment window
ECOG	c_ecog	edbx_get_ecog()	ECOG performance status measured closest to index date within assessment window. In case of ties, the lower ECOG value is selected	Nominal (0, 1, 2, 3, 4); modelled as ordinal numeric in ROPRO[10]; due to I/E criteria ECOG is modelled as a binary (0, 1) covariate	[-90;0]
Stage	c_stage_initial_dx	edbx_get_diagnosis_solid()	AJCC summary group stage at initial diagnosis	Ordinal numeric (from 0 to IV with sub-categories, e.g., IA1)[10]	[-inf;0] at initial diagnosis of primary cancer
De novo metastatic status	c_de_novo_mets_dx	edbx_get_diagnosis_solid()	Evidence of presence of one or multiple metastases at/before initial diagnosis	Binary logical (TRUE, FALSE)	[-inf;0] at initial diagnosis of primary cancer
Evidence of metastases	c_met_pre_index_aw	edbx_get_diagnosis_solid()	Evidence of any metastasis before the start of follow-up	Binary logical (TRUE, FALSE)	[-inf;X]; Proprietary business-rule-based covariate assessment window
Number of metastatic sites	c_number_met_sites_aw	edbx_get_diagnosis_solid()	Number of metastatic sites for a given patient before/on the start of follow-up	Integer	[-inf;X]; Proprietary business-rule-based covariate assessment window
Time between initial diagnosis to index date	c_time_dx_to_index	edbx_get_diagnosis_solid()	Time in days between initial diagnosis to index date	Continuous	[-initial dx;0]

Characteristic	Harmonized analysis variable name	R function to derive covariate (see pdf in appendix)	Details[1]	Variable encoding	Assessment window
Time between earliest evidence of a metastatic and index date	c_time_met_dx_to_index	edbx_get_diagnosis_solid()	Time in days between earliest evidence of a metastatic and index date	Continuous	[-inf;0]
Histology (adenocarcinoma)	c_histology_match	edbx_get_histology()	Evidence of adenocarcinoma histology (non-squamous cell for EDB1)	Binary logical (TRUE, FALSE)	[-inf;0]
Albumin	c_albumin_g_l_cont	edbx_get_labs()	Closest albumin measurement (in serum/plasma) relative to index date in g/L. In case of ties, the lower is selected	continuous	[-90;0]
Alkaline phosphatase (ALP) [2]	c_alp_u_l_cont	edbx_get_labs()	Closest alkaline phosphatase measurement (in serum/plasma) relative to index date in U/L. In case of ties, the lower is selected	continuous	[-90;0]
Alanine aminotransferase (ALT) [2]	c_alt_u_l_cont	edbx_get_labs()	Closest alanine transaminase measurement (in serum/plasma) relative to index date in U/L. In case of ties, the lower is selected	continuous	[-90;0]
Aspartate aminotransferase (AST)	c_ast_u_l_cont	edbx_get_labs()	Closest aspartate aminotransferase measurement (in serum/plasma) relative to	continuous	[-90;0]

Characteristic	Harmonized analysis variable name	R function to derive covariate (see pdf in appendix)	Details[1]	Variable encoding	Assessment window
			index date in U/L. In case of ties, the lower is selected. Used to compute AST-ALT ratio		
AST/ALT ratio	c_ast_alt_ratio_cont	edbx_get_labs()	AST/ALT ratio calculated from c_ast_u_l_cont/c_alt_u_l_cont	continuous	[-90;0]
Bilirubin[2]	c_bilirubin_mg_dl_cont	edbx_get_labs()	Closest bilirubin measurement (in serum/plasma) relative to index date in mg/dL. In case of ties, the lower is selected	continuous	[-90;0]
Calcium[2]	c_calcium_mg_dl_cont	edbx_get_labs()	Closest calcium measurement (in serum/plasma) relative to index date in mg/dL. In case of ties, the lower is selected	continuous	[-90;0]
Chloride	c_chloride_mmol_l_cont	edbx_get_labs()	Closest chloride measurement (in serum/plasma) relative to index date in mmol/L. In case of ties, the lower is selected	continuous	[-90;0]
Eosinophils/100 leukocytes[2]	c_eosinophils_leukocytes_ratio_cont	edbx_get_labs()	Eosinophils/100 leukocytes in blood. In case of ties, the lower	continuous	[-90;0]
Glucose[2]	c_glucose_mg_dl_cont	edbx_get_labs()	Closest glucose measurement (in serum/plasma) relative to index date in mmol/L. In case of ties, the lower is selected	continuous	[-90;0]

Characteristic	Harmonized analysis variable name	R function to derive covariate (see pdf in appendix)	Details[1]	Variable encoding	Assessment window
Hemoglobin	c_hemoglobin_g_dl_cont	edbx_get_labs()	Closest hemoglobin measurement (in blood) relative to index date in g/L. In case of ties, the lower is selected	continuous	[-90;0]
Lactate dehydrogenase (LDH)[3]	c_ldh_u_l_cont	edbx_get_labs()	Closest LDH measurement (in serum or plasma) relative to index date in U/L. In case of ties, the lower is selected	continuous	[-90;0]
Lymphocytes	c_lymphocyte_10_9_l_cont	edbx_get_labs()	Closest lymphocytes measurement (in blood) relative to index date in 10^9/L. In case of ties, the lower is selected. Used to compute neutrophil/lymphocyte ratio	continuous	[-90;0]
Lymphocyte/leukocyte ratio[2]	c_lymphocyte_leukocyte_ratio_cont	edbx_get_labs()	Closest lymphocyte/leukocyte ratio measurement (in blood) relative to index date. In case of ties, the lower is selected. Used to compute neutrophil/lymphocyte ratio	continuous	[-90;0]
Monocytes[2]	c_monocytes_10_9_l_cont	edbx_get_labs()	Closest monocytes measurement (in blood) relative to index date in 10^9/L. In case of ties, the lower is selected.	continuous	[-90;0]

Characteristic	Harmonized analysis variable name	R function to derive covariate (see pdf in appendix)	Details[1]	Variable encoding	Assessment window
Neutrophils	c_neutrophil_10_9_l_cont	edbx_get_labs()	Closest neutrophils measurement (in blood) relative to index date in 10^9/L. In case of ties, the lower is selected. Used to compute neutrophil/lymphocyte (NLR) ratio	continuous	[-90;0]
Neutrophil/lymphocyte ratio[2]	c_neutrophil_lymphocyte_ratio_cont	edbx_get_labs()	Neutrophil/lymphocyte (NLR) ratio calculated from c_neutrophil_10_9_l_cont/ c_lymphocyte_10_9_l_cont	continuous	[-90;0]
Platelets	c_platelets_10_9_l_cont	edbx_get_labs()	Closest platelets measurement (in blood) relative to index date in 10^9/L. In case of ties, the lower is selected	continuous	[-90;0]
Protein	c_protein_g_l_cont	edbx_get_labs()	Closest protein measurement (in serum/plasma) relative to index date in g/L. In case of ties, the lower is selected	continuous	[-90;0]
Urea nitrogen[2]	c_urea_nitrogen_mg_dl_cont	edbx_get_labs()	Closest urea nitrogen measurement (in serum/plasma) relative to index date in mg/L. In case of ties, the lower is selected	continuous	[-90;0]

Characteristic	Harmonized analysis variable name	R function to derive covariate (see pdf in appendix)	Details[1]	Variable encoding	Assessment window
Systolic blood pressure[2]	c_sbp_cont	edbx_get_vitals()	Closest systolic blood pressure (in mmHg) measurement. In case of ties, the lower is selected	continuous	[-90;0]
Body mass index (BMI)[2]	c_bmi_cont	edbx_get_vitals()	Closest BMI measurement (in kg/m^2) relative to index date. In case of ties, the lower is selected. For ROPRO and in EDB2, BMI is (additionally) computed from individual height and weight measurements	continuous	[-90;0]
Heart rate[2]	c_hr_cont	edbx_get_vitals()	Closest heart rate measurement (in bpm) relative to index date. In case of ties, the lower is selected	continuous	[-90;0]
Oxygen saturation	c_oxygen_cont	edbx_get_vitals()	Closest heart rate measurement (in bpm) relative to index date. In case of ties, the lower is selected	continuous	[-90;0]

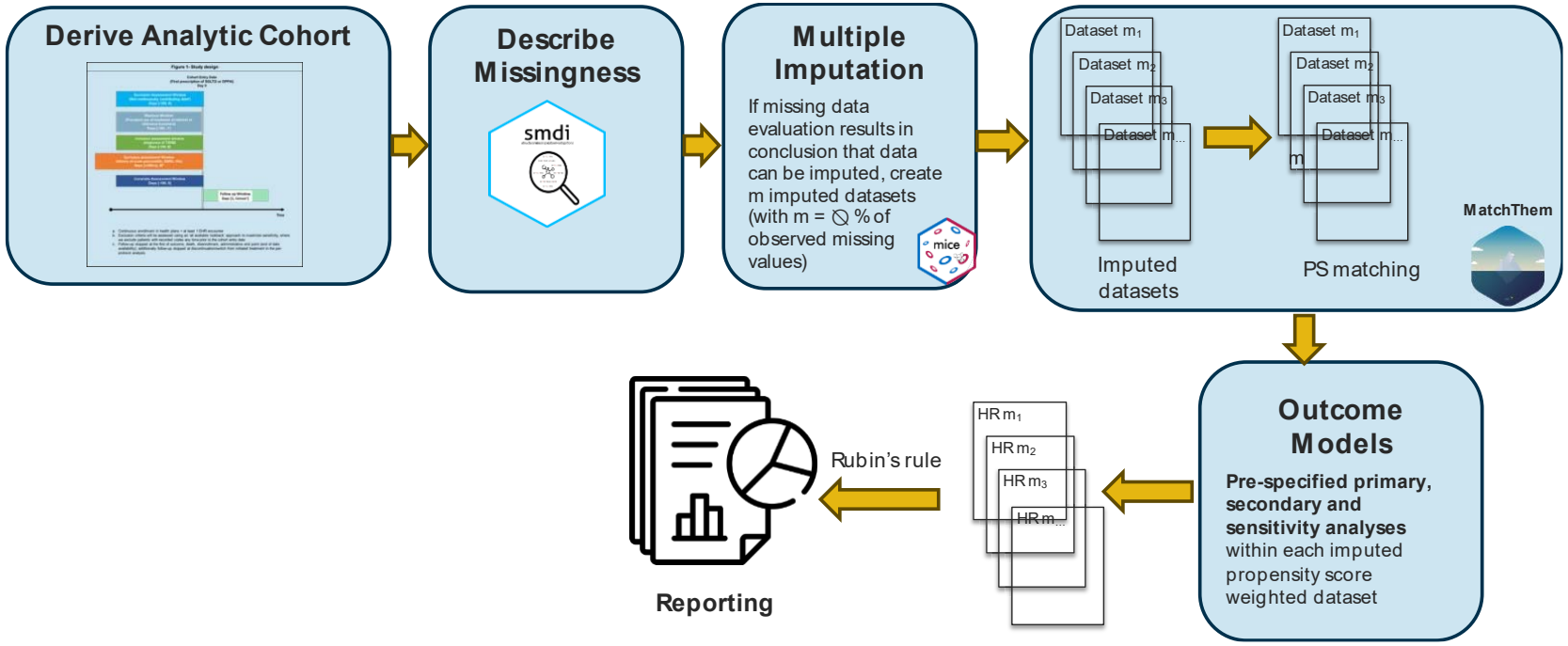
[1] x stands for the pseudonymized number of the respective database, i.e., EDB1, EDB2 or EDB4
[2] For calculation of ROPRO prognostic score[10], this variable is log transformed.
[3] For calculation of ROPRO prognostic score[10], this variable is log-log transformed.

6.7. *Data analysis*

6.7.1. Context and rationale for analysis plan

To emulate the KEYNOTE-042 trial, the following analytical workflow will be used (Figure). First, an analytical cohort with covariates on key eligibility criteria and prognostic factors will be derived across all databases. To ensure reproducibility and consistency throughout the entire ENCORE project, the internally developed *encore.io* R package streamlines this process using the functions referenced in Table 5. Operational definitions of key covariates used for trial emulation. Table 5 (code documentation see appendix).[16] The analytical cohort will be derived by first identifying an advanced NSCLC inception cohort of initiators of pembrolizumab or chemotherapy in the first-line setting as described in section 6.6.4. Next, key eligibility criteria will be applied in which patients with missing values are considered eligible in the respective attrition steps to allow thorough missing data investigations.

Figure 2. Illustration of principled analytical workflow.



Once a full analytic cohort is built, principled missing data investigations will be employed to empirically assess assumptions on potentially underlying missingness mechanisms according to Rubin's classification of missing data (i.e., missing completely at random [MCAR], missing at random [MAR] and missing not at random [MNAR]).[17] To that end, we will adopt a principled process on missing data that was developed as part of a FDA Sentinel Innovation Center causal inference

workstream that empirically evaluates different aspects across partially observed covariates based on three group diagnostics (Table 6).[18,19] In brief, the first group diagnostics computes distributions and absolute standardized mean differences (ASMD) between patients with and without an observed value for a given partially observed covariates. If missingness can be explained by observed covariates such as in MAR mechanisms, patient characteristics will significantly differ which will (in analogy to propensity scores) be indicated by ASMDs > 0.1. In addition, Hoteling's[20] and Little's[21] tests additionally provide formal hypothesis tests for such comparisons in which high test statistics and a rejection of the null hypothesis would provide evidence for differences in the distribution of patient characteristics and suggest the underlying mechanism is not MCAR or MNAR. Group 2 diagnostics assess the ability to predict missingness based on observed covariates by fitting a classification model to predict the missingness indicator of the partially observed covariate. To that end, we will fit a random forest (RF) classification model using observed covariates with a 70/30 train-test split of the complete cohort. A sufficiently high area under the receiver operating characteristic curve (AUC) metric of the test dataset may demonstrate that missingness can be predicted well and could point towards MAR as a likely mechanism as opposed to an AUC~0.5 which would suggest MCAR or MNAR. Group 3 diagnostics evaluates the association between the missingness indicator of the partially observed covariates and the outcome (OS). If the missingness of a confounder cannot be explained or approximated by observed covariates and a difference in the outcome is observed depending on the missingness indicator (e.g., $HR_{\text{missingness indicator}} \neq 1$), this may be indicative of an underlying MNAR mechanism. These empirical diagnostics will be implemented through the smdi R package[22] and be further enhanced by clinical expert knowledge.

Table 6. Diagnostics to empirically differentiate and characterize missing data mechanisms.

	Group 1 Diagnostics		Group 2 Diagnostics	Group 3 Diagnostics
Diagnostic metric	**Absolute standardized mean difference (ASMD)**	**P-value Hoteling[20]/ Little[21]**	**Area under the receiver operating curve (AUC)**	**Log HR (missingness indicator)**
Purpose	Comparison of distributions between patients with vs. without observed value of the partially observed covariate.		Assessing the ability to predict missingness based on observed covariates.	Check whether missingness of a covariate is associated with the outcome (differential missingness).
Example value	ASMD = 0.1	p-value < 0.001	AUC = 0.5	log HR = 0.1 (0.05 to 0.2)
Interpretation	<0.1[a]: no imbalances in observed patient characteristics; missingness may be likely completely at random or not at random (~MCAR, ~MNAR).	High test statistics and low p-values indicate differences in baseline covariate distributions and null hypothesis would be rejected (~MAR).	AUC values ~ 0.5 indicate completely random or not at random prediction (~MCAR, ~MNAR). Values meaningfully above 0.5 indicate stronger relationships between	No association in either univariate or adjusted model and no meaningful difference in the log HR after full adjustment (~MCAR).

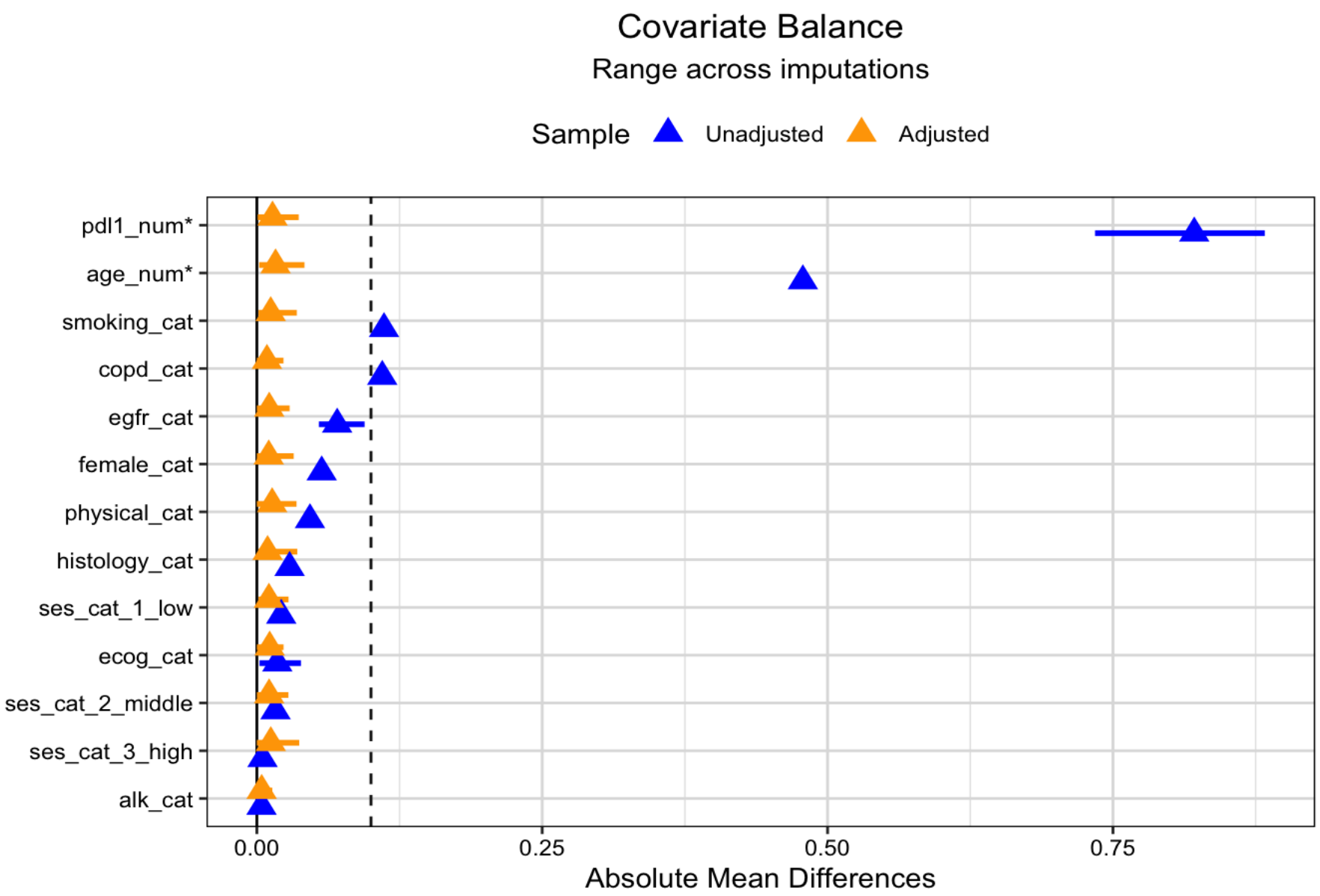
	>0.1[a]: imbalances in observed patient characteristics; missingness may be likely at random (~MAR).		covariates and missingness (~MAR).	Association in univariate but not fully adjusted model (~MAR). Meaningful difference in the log HR also after full adjustment (~MNAR).

Abbreviations: ASMD = Median absolute standardized mean difference across all covariates, AUC = Area under the curve, CI = Confidence interval, MAR = Missing at random mechanism in which the missingness probability depends on observed covariates, MCAR = Missing completely at random mechanism in which each patients has the same missingness probability, MNAR(unmeasured) = Missing not at random mechanism in which the missingness can only be explained by a covariate which is not observed in the underlying dataset, MNAR(value) = Missing not at random mechanism in which the missingness just depends on the actual value of the partially observed confounder of interest itself.

[a] Analogous to propensity score-based balance measures.[23][a] Analogous to propensity score-based balance measures.[23]

While the MAR assumption is a strong assumption to hold across all considered covariates, it was shown that especially in the context of partially observed covariate data (as opposed to missing exposure and outcome data), only mechanisms in which a covariate causes its own missingness leads to critical bias (MNAR).[24] In such situations, multiple imputation can have significant advantages over a complete case analysis since additional information (auxiliary covariates and missing indicator variables) can be included in imputation algorithms which can make the MAR assumption holding more plausible and increase efficiency in treatment effect estimates since all patients and critical covariates can be retained and variances can be realistically estimated, considering both the general sampling error and the error introduced by missing data.[25,26] Hence, multiple imputation with flexible, non-parametric random forest imputation algorithms[27] (mice R package[28]) will be used for this trial emulation. The number of imputed datasets (*m*) will be determined for each database separately based on the average proportion missingness observed in the analytic cohort and results from the above-referenced missing data investigations will inform the choice of appropriate sensitivity analyses.

Figure 3. Covariate balance across imputed datasets (simulated example).



To estimate the treatment effects for pembrolizumab using propensity score matching across imputed datasets we will apply the "within" approach using the "MatchThem" R package.[29,30] That is, propensity score matching and the estimation of the treatment effect are performed in each imputed dataset separately and resulting treatment effect estimates are combined using Rubin's rule. In this study, this will be implemented by matching eligible patients on their propensity to initiate pembrolizumab using a 1:1 nearest neighbor matching algorithm without replacement and a caliper of 1% of the standard deviation of the propensity score. The resulting covariates balance will be assessed by computing and visualizing ASMDs before and after matching across datasets. As compared to a single dataset matching approach, this can lead to a range of ASMDs per covariate due to random variation across imputed datasets for which an example (using simulated data) is illustrated in Figure . If sufficient balance can be established, a Cox proportional hazards regression model will be fit to estimate the marginal average treatment effect in the matched population. Since in most databases there are more pembrolizumab patients than patients in the control arm, the estimand will rather correspond to the ATC than the ATT. Confidence intervals will be estimated using cluster-robust standard errors.[31] As a secondary endpoint, we will additionally estimate the median OS survival time difference between the two treatment groups using the Kaplan-Meier method. It should be noted that due to administrative and de-identification purposes, the date of death is often only available at the month- or year-granularity level, in which case the date of death will be imputed to the 15th of a month or July 2nd of a year, respectively (depending on the database). In rare cases, this can lead to negative/implausible follow-up times if the date of death is very close to the index date. These patients will be excluded from the analysis.

The final hazard ratio and median OS survival time difference estimates for each database will then be combined using Rubin's rule.[28,32] A summary of the analytic approach is summarized in Table 7 and an example workflow with simulated data and annotated code can be found here.

A. Primary analysis

Table 7. Primary and subgroup analysis specification

Hypothesis:	Initiation of pembrolizumab decreases the hazard of all-cause mortality as compared to initiation of chemotherapy in patients with NSCLC and PD-L1 ≥50%.
Exposure contrast:	Initiation of pembrolizumab vs. initiation of chemotherapy
Outcome:	Time to all-cause mortality
Databases used:	EDB1
Analytic software:	R 4.3.2. Version control of code and R packages will be established through git and Posit package manager, respectively. All packages are frozen to their most recent version as of April 24, 2024.
Model(s): *(provide details or code)*	See example code here. The annotated code for the trial emulation will be hosted at https://gitlab.partners.org/drugepi/encore/keynote-042 (access only through MGB network for authorized personnel)
Confounding adjustment method	*Name method and provide relevant details, e.g. bivariate, multivariable, propensity score matching (specify matching algorithm ratio and caliper), propensity score weighting (specify weight formula, trimming, truncation), propensity score stratification (specify strata definition), other.*
	1:1 propensity score nearest neighbor matching without replacement and a caliper of 1% of propensity score standard deviation
Missing data methods	*Name method and provide relevant details, e.g. missing indicators, complete case, last value carried forward, multiple imputation (specify model/variables), other.*
	Multiple imputation by chained equations using a random forest imputation model across all covariate types. The number of imputed datasets will be determined by the average proportion of missing values across all partially observed covariates. Imputation models will include all variables of the substantiative model, i.e., exposure, outcome, confounders/prognostic factors and additional auxiliary covariates.

Subgroup Analyses	*List all subgroups*
	In subgroup analysis, propensity score matching and balance assessment will be conducted within each subgroup separately. The treatment effect will be estimated for each stratum separately (stratum-specific effects). 1. Age (<65 vs ≥65 years) 2. Sex (Male vs Female) 3. ECOG performance status (0 vs 1) 4. Histologic features (Squamous vs Nonsquamous) 5. Chemotherapy regimen (Pemetrexed and carboplatin vs Paclitaxel and carboplatin) 6. Disease status (Locally advanced vs Metastatic)

B. Secondary Analysis

Table 8. Secondary analysis specification.

Hypothesis:	Initiation of pembrolizumab increases overall survival as compared to initiation of chemotherapy in patients with NSCLC and PD-L1 ≥50%.
Exposure contrast:	Initiation of pembrolizumab vs. initiation of chemotherapy
Outcome:	Median overall survival time for patients, i.e., time until 50% of the patients in each stratum becomes deceased
Databases used:	EDB1
Analytic software:	R 4.3.2. Version control of code and R packages will be established through git and Posit package manager, respectively. All packages are frozen to their most recent version as of April 24, 2024.
Model(s): ***(provide details or code)***	See example code here. The annotated code for the trial emulation will be hosted at https://gitlab.partners.org/drugepi/encore/keynote-042/ (access only through MGB network for authorized personnel)
Confounding adjustment method	***Name method and provide relevant details, e.g. bivariate, multivariable, propensity score matching (specify matching algorithm ratio and caliper), propensity score weighting (specify weight formula, trimming, truncation), propensity score stratification (specify strata definition), other.***
	1:1 propensity score nearest neighbor matching without replacement and a caliper of 1% of propensity score standard deviation
Missing data methods	***Name method and provide relevant details, e.g. missing indicators, complete case, last value carried forward, multiple imputation (specify model/variables), other.***
	Multiple imputation by chained equations using a random forest imputation model across all covariate types. The number of imputed datasets will be determined by the average proportion of missing values across all partially observed covariates. Imputation models will include all variables of the substantiative model, i.e., exposure, outcome, confounders/prognostic factors and additional auxiliary covariates.

Subgroup Analyses	*List all subgroups*
	In subgroup analysis, propensity score matching, and balance assessment will be conducted within each subgroup separately. The treatment effect will be estimated for each stratum separately (stratum-specific effects). 1. Age (<65 vs ≥65 years) 2. Sex (Male vs Female) 3. ECOG performance status (0 vs 1) 4. Histologic (Squamous vs Nonsquamous) 5. Chemotherapy regimen (Pemetrexed and carboplatin vs Paclitaxel and carboplatin) 6. Disease status (Locally advanced vs Metastatic)

Table 9. Sensitivity 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 #1	Caliper matching on ROPRO prognostic score instead of propensity score	Matching patients on validated prognostic score may be more beneficial to control for (unmeasured) confounding	Matches patients on validated prognostic score that incorporates weights of key prognostic factors	Limited experience on how to optimally use prognostic scores and should be seen as an experimental sensitivity analysis
Sensitivity #2	ATO weighting instead of matching	Weights that resemble the average treatment effect in the overlap population (ATO) create a clinical equipoise population which is comparable to an RCT	ATO weighting usually results in excellent balance and clinical equipoise	Estimates the average treatment effect among the overlap patients which may not be comparable to target population anymore
Sensitivity #3	SMR/ATT weighting instead of matching. Here symmetric trimming (i.e., setting all weights lower/higher than that at a given quantile to the weight at the quantile) of extreme weights may be considered with the quantiles chosen based on weight distribution and resulting balancing performance.	SMR weighting retains all patients and resembles the same estimand as matching	ATT weighting retains all patients	Patients with extreme weights after trimming may bias the analysis
Sensitivity #4	Censoring date is changed to 3 months before data cut-off date	For all databases, information on mortality comes from different data sources which are updated	Approach implements a more conservative censoring rule	Approach addresses ghost-time bias by censoring

		asynchronously. To account for the potential lag of updated mortality information (*ghost-time bias*[34]), the censoring date for patients without mortality event in the whole patient identification period will be moved to last sign of patients being alive/visit or 3 months before data cut-off date, whichever occurred earlier.[35]		patients without a recorded death event earlier
Sensitivity #5	Delta imputation models for MNAR (tipping point analysis)	Primary multiple imputation analysis assumes MAR which may not hold for every covariate	Estimates impact of deviations from MAR assumption on final treatment effect estimates for key covariates	Delta parameters must be assumed and results are complex to interpret in multivariate missingness settings; just must important covariates or those with highest suspicion of being MNAR will be evaluated
Sensitivity #6	Re-weighting of strong risk factors and/or treatment effect modifiers distribution to match that of KEYNOTE-042	In the presence of effect modification, treatment effect estimates may be different if the distribution of strong risk factors/effect modifiers is different in the emulated cohort versus the trial cohort	Re-weighting adjusts for differences in distributions of key risk factors and/or treatment effect modifiers (see subgroup analysis in Table 7)	Re-weighting risk factors/potential effect modifiers to match the KEYNOTE-042 trial and simultaneously balancing them across treatment groups may be challenging due to differences in measurement

Sensitivity #7	Including patients who have had at least 1 visit 90 days prior to treatment initiation	EHR are often lacking data continuity, and this analysis uses the requirement of 1 visit as a proxy for continuous observation periods	Considers aspect of data continuity	There may be patients who are put on treatment immediately in which case they are falsely excluded
Sensitivity #8	Use EDB4 to estimate the treatment effects	Evaluate the potential impact of selection/immortal time bias on the treatment effects.	N/A	See section 6.1.1 for the limitations of EDB 4
Sensitivity #9	Inclusion of patients with chemotherapy exposure before October 24, 2016	Evaluate the impact of including patients who used chemotherapy in earlier calendar years (before the pembrolizumab marketing).	Increase the sample size in the comparison group	Treatment patterns and care quality may vary across calendar years.
Sensitivity #10	Missingness is handled by restricting to patients with complete observations on a subset of the most important confounders ("complete cases").	Instead of imputing data, this sensitivity analysis restricts the analysis cohort to patients with complete observations on key confounders	Data will not be imputed and missingness is assumed to be missing completely at random	The restriction to complete cases will significantly decrease sample size. To limit the attrition of patients with partially observed covariates, it won't be possible to use all covariates used in the main analysis propensity score model, but only consider key covariates with overall low proportions of missingness (age, sex, etc.)
Sensitivity #11	Expand the study population to PD-L1 ≥20%	This subgroup corresponded to one of the prespecified primary OS analysis	Expanding the analytic population to patients with PD-L1 ⩾20% corresponds	The trial provided evidence that there was treatment effect heterogeneity related

		populations in KEYNOTE-042.	to a planned co-primary analysis from the trial.	to differences in PD-L1 level. Because the distribution of the PD-L1 levels differs between KEYNOTE-042 and the effect estimates would be expected to diverge.
Sensitivity #12	Expand the study population to PD-L1 ≥1%	This group corresponded to one of the prespecified primary OS analysis populations in KEYNOTE-042.	Expanding the analytic population to patients with PD-L1 ⩾1% corresponds to a planned co-primary analysis from the trial.	The trial provided evidence that there was treatment effect heterogeneity related to differences in PD-L1 level. Because the distribution of the PD-L1 levels differs between KEYNOTE-042 and the effect estimates would be expected to diverge.
Sensitivity #13	Restrict to metastatic (stage IV) patients only	The majority of patients (89.2%) in KEYNOTE-042 are in the metastatic stage. In contrast, 58% of patients in the emulation chemotherapy arm had locally advanced disease. After propensity score matching, the emulation cohort still included a higher proportion of locally advanced patients than the trial population.	Potentially better aligned with the trial's treatment population.	Does not fully benchmark against the trial population and reduces statistical power due to a smaller analytic sample.

7. Limitations of the methods

- Missingness in prognostic factors is a major challenge which is addressed in this emulation by multiple imputation using a non-parametric imputation algorithm. Multiple imputation usually assumes that missingness can be explained by observed characteristics, which may be empirically evaluated using principled missingness diagnostics, but the true underlying missingness mechanisms are usually unknown. Nevertheless, multiple imputation makes use of additional information (auxiliary covariates) which can render the underlying missingness assumptions more plausible. In addition, assumptions for alternative missing data approaches like complete case analysis or the "missing indicator approach" come with even stronger assumptions and additionally have the limitation of significantly reduced sample sizes, especially when comprehensively adjusting for known confounders and prognostic factors.
- Data continuity is a major challenge in EHR databases since "guaranteed" observable periods (such as continuous enrolment periods in administrative claims data) do not exist which may lead to measurement error in key covariates and exposure misclassification. Sensitivity analysis #8 tries to address this requiring patients to have had at least one visit before the index date which increases the likelihood that a patients was not only diagnosed at the respective center but is also regularly seen.
- Balancing patients on calendar year is not possible since calendar year shows instrumental variable-like behaviours (see Figure 9), i.e., it perfectly predicts treatment assignment and does not have any association with the outcome other than through the exposure. This assumption is not directly testable using observational data, and calendar time is likely influenced not only by the introduction of pembrolizumab but also by subsequent approvals and uptake of other immunotherapies and new indications over time. The improvements of radiation of brain metastases may be the only exception, but it is expected that this may be of negligible significance for the scope of this emulation.
- Given the documented clinical benefit and the corresponding rapid uptake of pembrolizumab after October 2016, it could be assumed that patients who initiated the "old" standard-of-care after October 2016 did so for specific reasons. This means these patients may differ in their baseline characteristics and prognosis for reasons that may not be measurable.

8. Protection of human subjects

This study has been approved by the Brigham and Women's Hospital Institutional Review Board.

9. References

1. Jm F, A P, D M, et al. Nonrandomized Real-World Evidence to Support Regulatory Decision Making: Process for a Randomized Trial Replication Project. *Clin Pharmacol Ther*. 2020;107(4). doi:10.1002/cpt.1633

2. Franklin JM, Patorno E, Desai RJ, et al. Emulating Randomized Clinical Trials With Nonrandomized Real-World Evidence Studies: First Results From the RCT DUPLICATE Initiative. *Circulation*. 2021;143(10):1002-1013. doi:10.1161/CIRCULATIONAHA.120.051718

3. Franklin JM, Glynn RJ, Suissa S, Schneeweiss S. Emulation Differences vs. Biases When Calibrating Real-World Evidence Findings Against Randomized Controlled Trials. *Clin Pharmacol Ther*. 2020;107(4):735-737. doi:10.1002/cpt.1793

4. Wang SV, Schneeweiss S, RCT-DUPLICATE Initiative. Emulation of Randomized Clinical Trials With Nonrandomized Database Analyses: Results of 32 Clinical Trials. *JAMA*. 2023;329(16):1376-1385. doi:10.1001/jama.2023.4221

5. De Castro G, Kudaba I, Wu YL, et al. Five-Year Outcomes With Pembrolizumab Versus Chemotherapy as First-Line Therapy in Patients With Non–Small-Cell Lung Cancer and Programmed Death Ligand-1 Tumor Proportion Score ≥ 1% in the KEYNOTE-042 Study. *J Clin Oncol*. 2023;41(11):1986-1991. doi:10.1200/JCO.21.02885

6. Curtis MD, Griffith SD, Tucker M, et al. Development and Validation of a High-Quality Composite Real-World Mortality Endpoint. *Health Serv Res*. 2018;53(6):4460-4476. doi:10.1111/1475-6773.12872

7. Dong S, Kansagra AJ, Kaur G, et al. Validation of a Composite Real-World Mortality Variable Among Patients with Hematologic Malignancies Treated in the United States. *Blood*. 2023;142:5145.

8. Govindan R, Aggarwal C, Antonia SJ, et al. Society for Immunotherapy of Cancer (SITC) clinical practice guideline on immunotherapy for the treatment of lung cancer and mesothelioma. *J Immunother Cancer*. 2022;10(5):e003956. doi:10.1136/jitc-2021-003956

9. Hernán MA, Wang W, Leaf DE. Target Trial Emulation: A Framework for Causal Inference From Observational Data. *JAMA*. Published online December 12, 2022. doi:10.1001/jama.2022.21383

10. Becker T, Weberpals J, Jegg AM, et al. An enhanced prognostic score for overall survival of patients with cancer derived from a large real-world cohort. *Ann Oncol*. 2020;31(11):1561-1568. doi:10.1016/j.annonc.2020.07.013

11. Becker T, Mailman M, Tan S, Lo E, Bauer-Mehren A. Comparison of overall survival prognostic power of contemporary prognostic scores in prevailing tumor indications. *Med Res Arch*. 2023;11(4). doi:10.18103/mra.v11i4.3638

12. Arkenau HT, Barriuso J, Olmos D, et al. Prospective validation of a prognostic score to improve patient selection for oncology phase I trials. *J Clin Oncol Off J Am Soc Clin Oncol*. 2009;27(16):2692-2696. doi:10.1200/JCO.2008.19.5081

13. Loureiro H, Becker T, Bauer-Mehren A, Ahmidi N, Weberpals J. Artificial Intelligence for Prognostic Scores in Oncology: a Benchmarking Study. *Front Artif Intell*. 2021;4:625573. doi:10.3389/frai.2021.625573

14. Loureiro H, Roller A, Schneider M, Talavera-López C, Becker T, Bauer-Mehren A. Matching by OS Prognostic Score to Construct External Controls in Lung Cancer Clinical Trials. *Clin Pharmacol Ther*. n/a(n/a). doi:10.1002/cpt.3109

15. Mok TSK, Wu YL, Kudaba I, et al. Pembrolizumab versus chemotherapy for previously untreated, PD-L1-expressing, locally advanced or metastatic non-small-cell lung cancer (KEYNOTE-042): a randomised, open-label, controlled, phase 3 trial. *The Lancet*. 2019;393(10183):1819-1830. doi:10.1016/S0140-6736(18)32409-7

16. Weberpals J, Wang SV. The FAIRification of research in real-world evidence: A practical introduction to reproducible analytic workflows using Git and R. *Pharmacoepidemiol Drug Saf*. 2024;33(1):e5740. doi:10.1002/pds.5740

17. RUBIN DB. Inference and missing data. *Biometrika*. 1976;63(3):581-592. doi:10.1093/biomet/63.3.581

18. Weberpals J, Raman SR, Shaw PA, et al. A Principled Approach to Characterize and Analyze Partially Observed Confounder Data from Electronic Health Records. *Clin Epidemiol*. 2024;16:329-343. doi:10.2147/CLEP.S436131

19. Sondhi A, Weberpals J, Yerram P, et al. A systematic approach towards missing lab data in electronic health records: A case study in non-small cell lung cancer and multiple myeloma. *CPT Pharmacomet Syst Pharmacol*. 2023;12(9):1201-1212. doi:10.1002/psp4.12998

20. Hotelling H. The Generalization of Student's Ratio. *Ann Math Stat*. 1931;2(3):360-378. doi:10.1214/aoms/1177732979

21. Little RJA. A Test of Missing Completely at Random for Multivariate Data with Missing Values. *J Am Stat Assoc*. 1988;83(404):1198-1202. doi:10.1080/01621459.1988.10478722

22. Weberpals J, Raman SR, Shaw PA, et al. smdi: an R package to perform structural missing data investigations on partially observed confounders in real-world evidence studies. *JAMIA Open*. 2024;7(1):ooae008. doi:10.1093/jamiaopen/ooae008

23. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivar Behav Res*. 2011;46(3):399-424. doi:10.1080/00273171.2011.568786

24. Moreno-Betancur M, Lee KJ, Leacy FP, White IR, Simpson JA, Carlin JB. Canonical Causal Diagrams to Guide the Treatment of Missing Data in Epidemiologic Studies. *Am J Epidemiol*. 2018;187(12):2705-2715. doi:10.1093/aje/kwy173

25. Bartlett JW, Seaman SR, White IR, Carpenter JR. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Stat Methods Med Res*. 2015;24(4):462-487. doi:10.1177/0962280214521348

26. Weberpals J, Shaw PA, Lin KJ, et al. High-dimensional multiple imputation (HDMI) for partially observed confounders including natural language processing-derived auxiliary covariates. *arXiv*. Preprint posted online May 17, 2024:arXiv:2405.10925. doi:10.48550/arXiv.2405.10925

27. Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *Am J Epidemiol*. 2014;179(6):764-774. doi:10.1093/aje/kwt312

28. Buuren S van, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *J Stat Softw*. 2011;45:1-67. doi:10.18637/jss.v045.i03

29. Leyrat C, Seaman SR, White IR, et al. Propensity score analysis with partially observed covariates: How should multiple imputation be used? *Stat Methods Med Res*. 2019;28(1):3-19. doi:10.1177/0962280217713032

30. Pishgar F, Greifer N, Leyrat C, Stuart E. MatchThem: Matching and Weighting after Multiple Imputation. *R J*. 2021;13(2):292-305. doi:10.32614/RJ-2021-073

31. Austin PC. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Stat Med*. 2014;33(7):1242-1258. doi:10.1002/sim.5984

32. Rubin DB. Multiple imputation. In: *Flexible Imputation of Missing Data, Second Edition*. Chapman and Hall/CRC; 2018:29-62.

33. Schwarzer G, Carpenter JR, Rücker G. *Meta-Analysis with R*. Vol 4784. Springer; 2015.

34. Jacobs EJ, Newton CC, Wang Y, Campbell PT, Flanders WD, Gapstur SM. Ghost-time bias from imperfect mortality ascertainment in aging cohorts. *Ann Epidemiol*. 2018;28(10):691-696.e3. doi:10.1016/j.annepidem.2018.06.002

35. Chen L, Fajardo O, Huntley M, Meyer AM, Taylor M. Use of last clinical activity date in overall survival analysis with real world data. In: *PHARMACOEPIDEMIOLOGY AND DRUG SAFETY*. Vol 30. WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA; 2021:116-116.

36. Schoenfeld DA. Sample-size formula for the proportional-hazards regression model. *Biometrics*. 1983;39(2):499-503.

10. Appendices

10.1. CONSORT diagrams

The following CONSORT attrition diagrams depict the process to select eligible KEYNOTE-042-like populations in EDB1, EDB2 and EDB4 for the main analysis, respectively.

Figure 2. CONSORT attrition to select eligible KEYNOTE-042-like populations in EDB1.

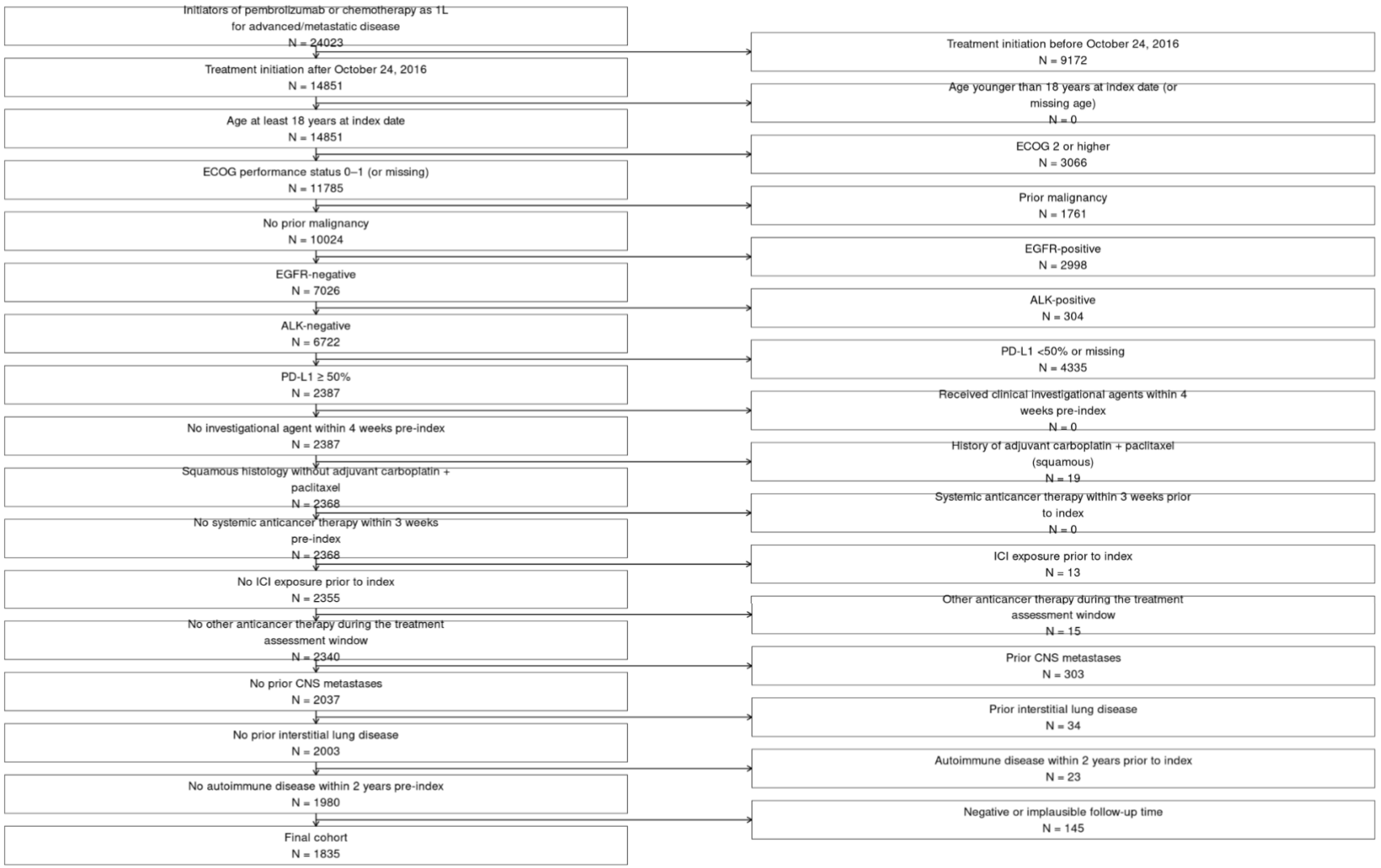
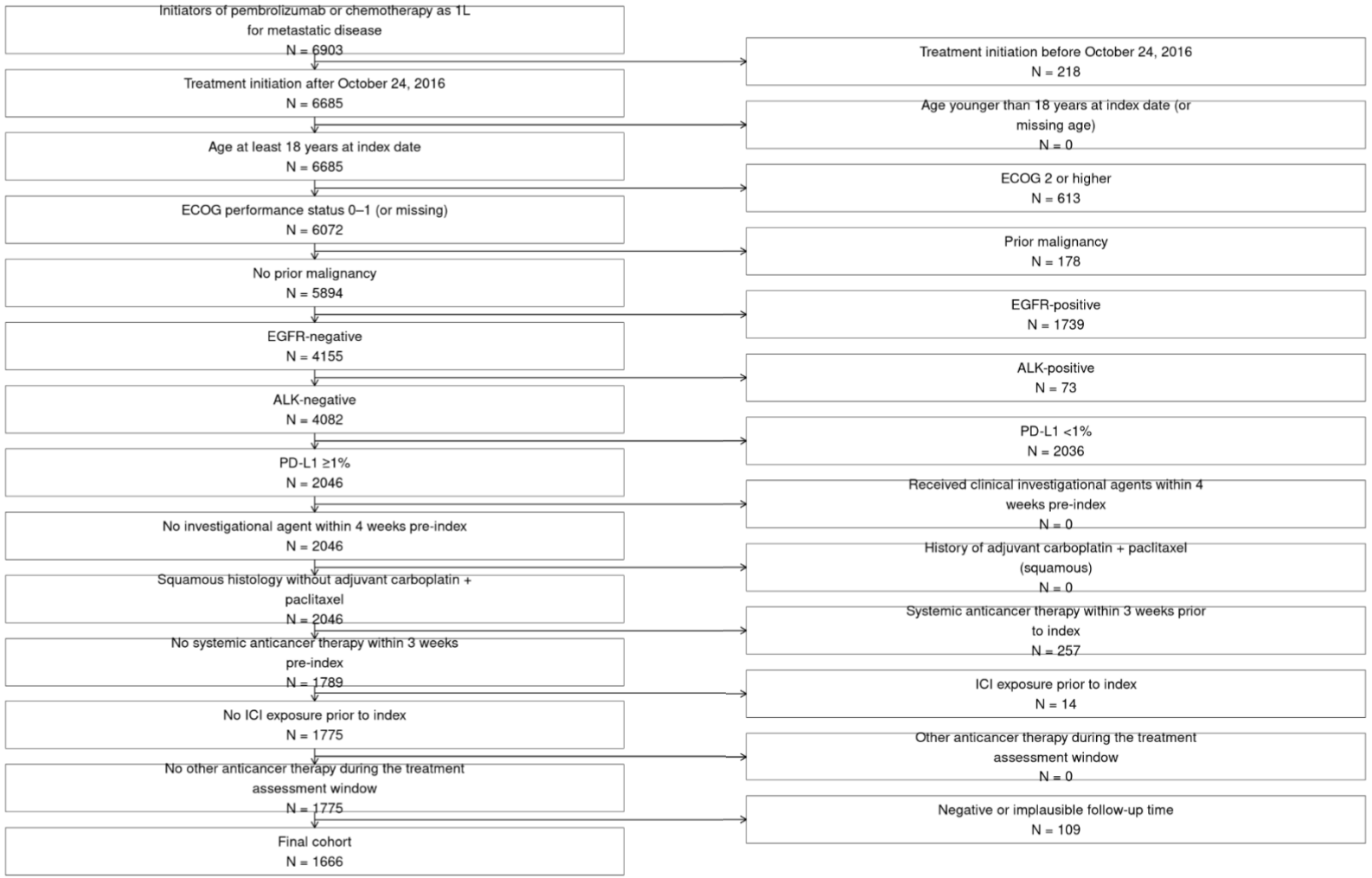


Figure 3. CONSOR attrition to select eligible KEYNOTE-042-like populations in EDB2.

Figure 4. CONSORT attrition to select eligible KEYNOTE-042-like populations in EDB4.



10.2. Covariate balance figures

The following figures illustrate the balance of key covariates included in propensity score models among eligible KEYNOTE-042-like populations in EDB1, EDB2 and EDB4, respectively.

Figure 5. EDB1 covariate balance of covariates included in propensity score model before and after matching.

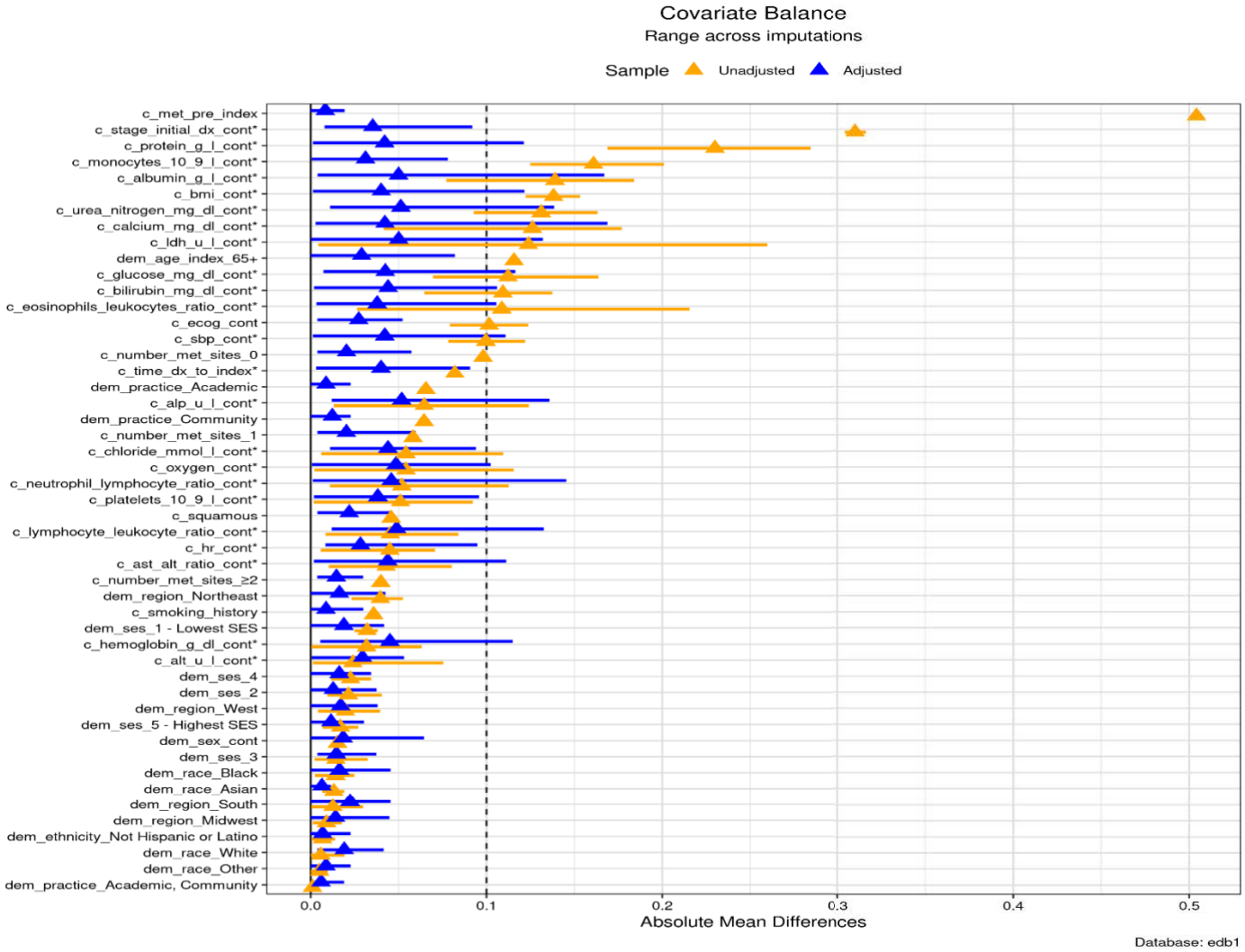


Figure 6. EDB2 covariate balance of covariates included in propensity score model before and after matching.

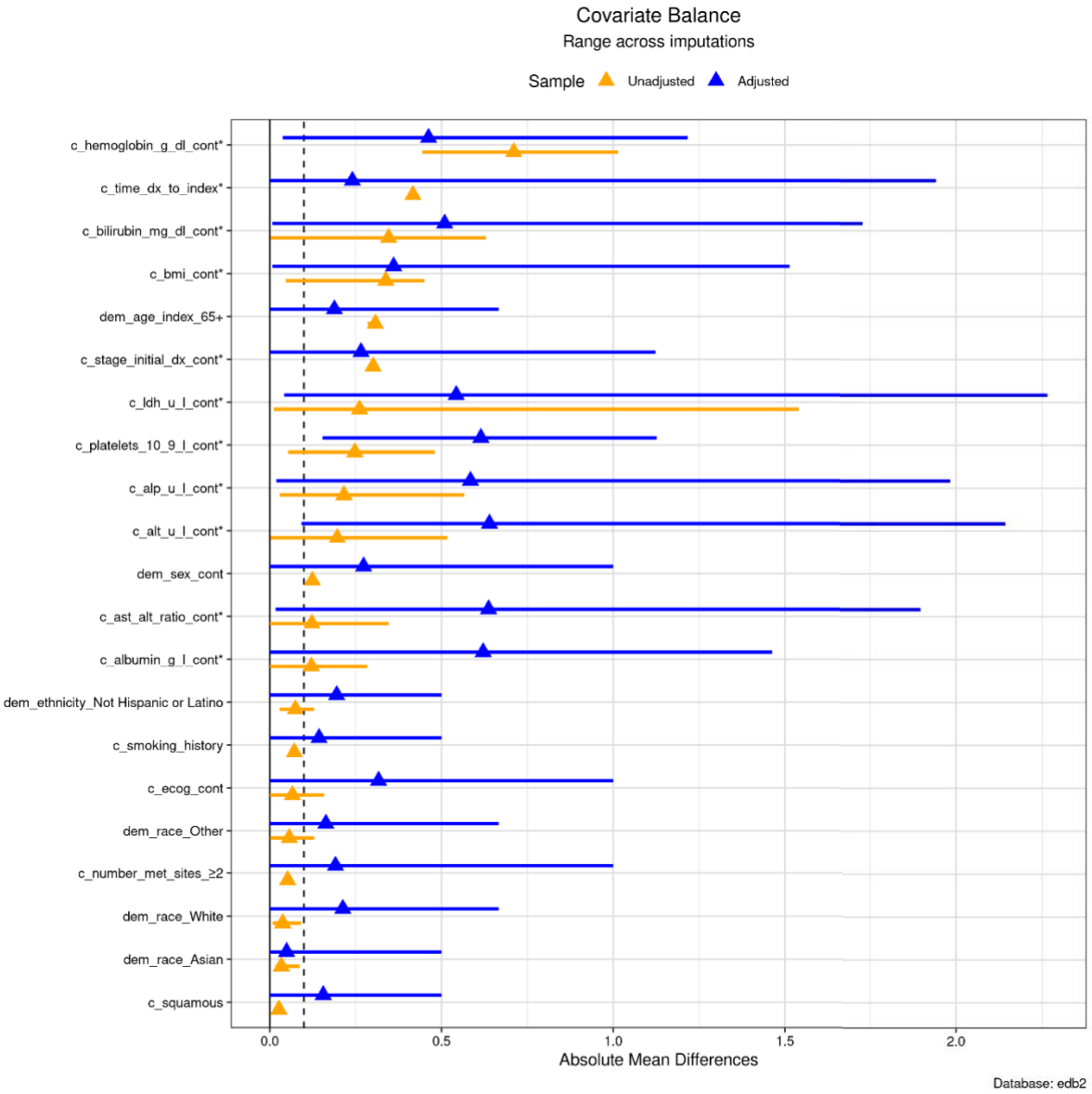
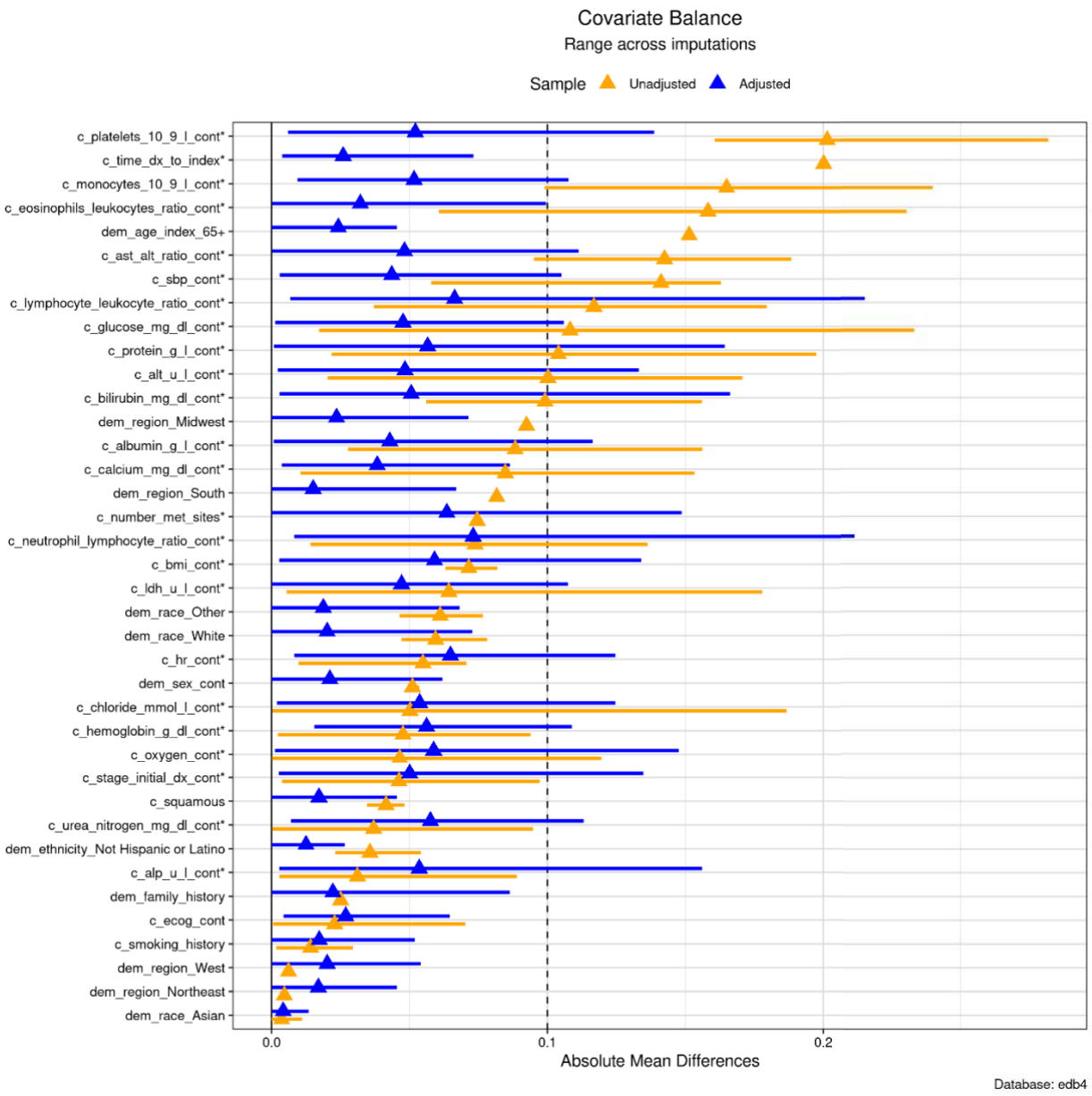


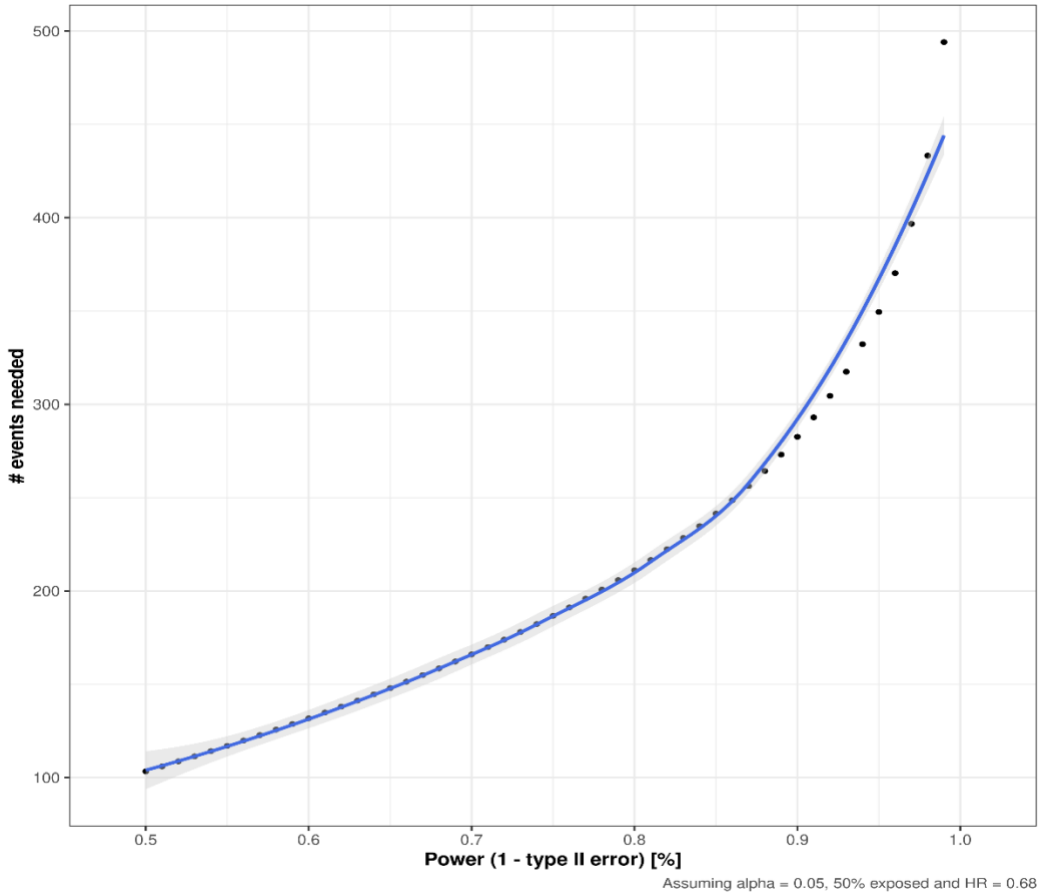
Figure 7. EDB4 covariate balance of covariates included in propensity score model before and after matching.



10.3. Sample size/power calculations

Power estimations are computed based on the average number of observed events (N=335, unstratified by treatment group) across imputed and matched datasets according to the methodology described by Schoenfeld.[36] Considering EDB1 only for primary analysis and assuming an HR = 0.69 and an alpha = 0.05 (2-sided), this results in an estimated statistical power (1- estimated type II error) of 92.4%.

Figure 8. Number of events needed to achieve x% power.



10.4. Additional Figures and Tables

Table 10. Lab measurement plausibility thresholds.

Lab and standardized unit	Lower plausibility threshold	Upper plausibility threshold
c_albumin_g_l	10	200
c_alp_u_l	1	2000
c_alt_u_l	1	90000
c_ast_u_l	1	90000
c_bilirubin_mg_dl	0.1	80
c_calcium_mg_dl	0.1	20
c_chloride_mmol_l	0.1	200
c_eosinophils_leukocytes_ratio	0	100
c_glucose_mg_dl	0.1	2000
c_granulocytes_leukocytes_ratio	0	100
c_hemoglobin_g_dl	0.1	20
c_ldh_u_l	0.1	Inf
c_lymphocyte_10_9_l	0	1e+06
c_lymphocyte_leukocyte_ratio	0	100
c_monocytes_10_9_l	0	1e+06
c_neutrophil_10_9_l	0	1e+06
c_platelets_10_9_l	0	5000
c_protein_g_l	1	300
c_urea_nitrogen_mg_dl	0.1	250

Table 11. Vital sign measurement plausibility thresholds.

Vital sign	Lower plausibility threshold	Upper plausibility threshold
c_sbp	50	250
c_dbp	30	150
c_bmi	10	80
c_bsa	0.5	3.5
c_height	0.5	3
c_oxygen	50	100
c_pain	0	10
c_hr	20	250
c_resp	5	50
c_temp	86	113
c_weight	20	300

Table 12. Mapping from State to Region.

State	Region
CT	Northeast
ME	Northeast
MA	Northeast
NH	Northeast
RI	Northeast
VT	Northeast
DE	Northeast
NJ	Northeast
NY	Northeast
PA	Northeast
IL	Midwest
IN	Midwest
MI	Midwest
OH	Midwest
WI	Midwest
IA	Midwest
KS	Midwest
MN	Midwest
MO	Midwest
NE	Midwest

ND	Midwest
SD	Midwest
FL	South
GA	South
MD	South
NC	South
SC	South
VA	South
DC	South
WV	South
AL	South
KY	South
MS	South
TN	South
AR	South
LA	South
OK	South
TX	South
AZ	West
CO	West
ID	West
MT	West

NV	West
NM	West
UT	West
WY	West
AK	West
CA	West
HI	West
OR	West
WA	West

Figure 9. Treatment initiation trends by calendar year and treatment in EDB1.

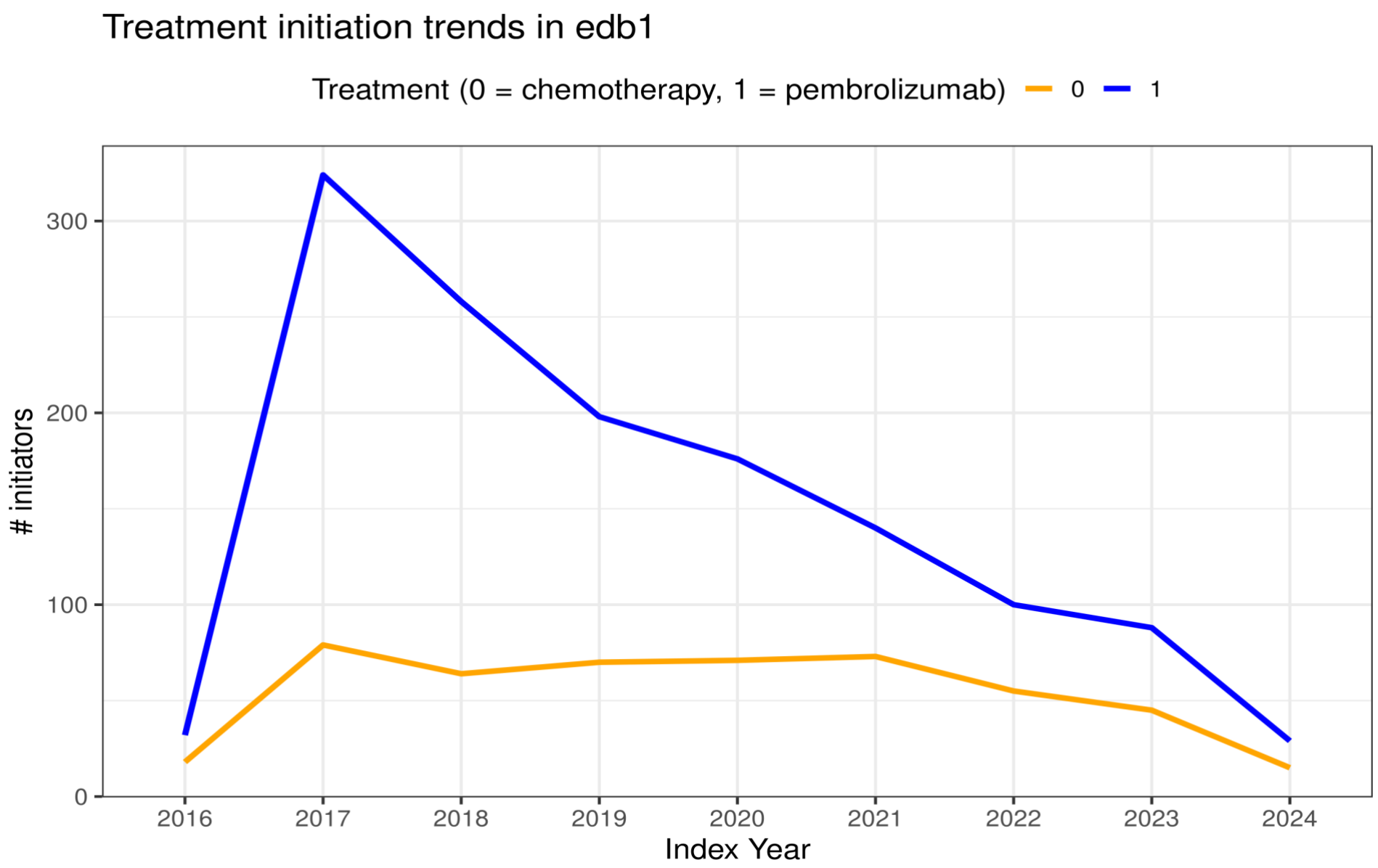


Figure 10. Treatment initiation trends by calendar year and treatment in EDB2.

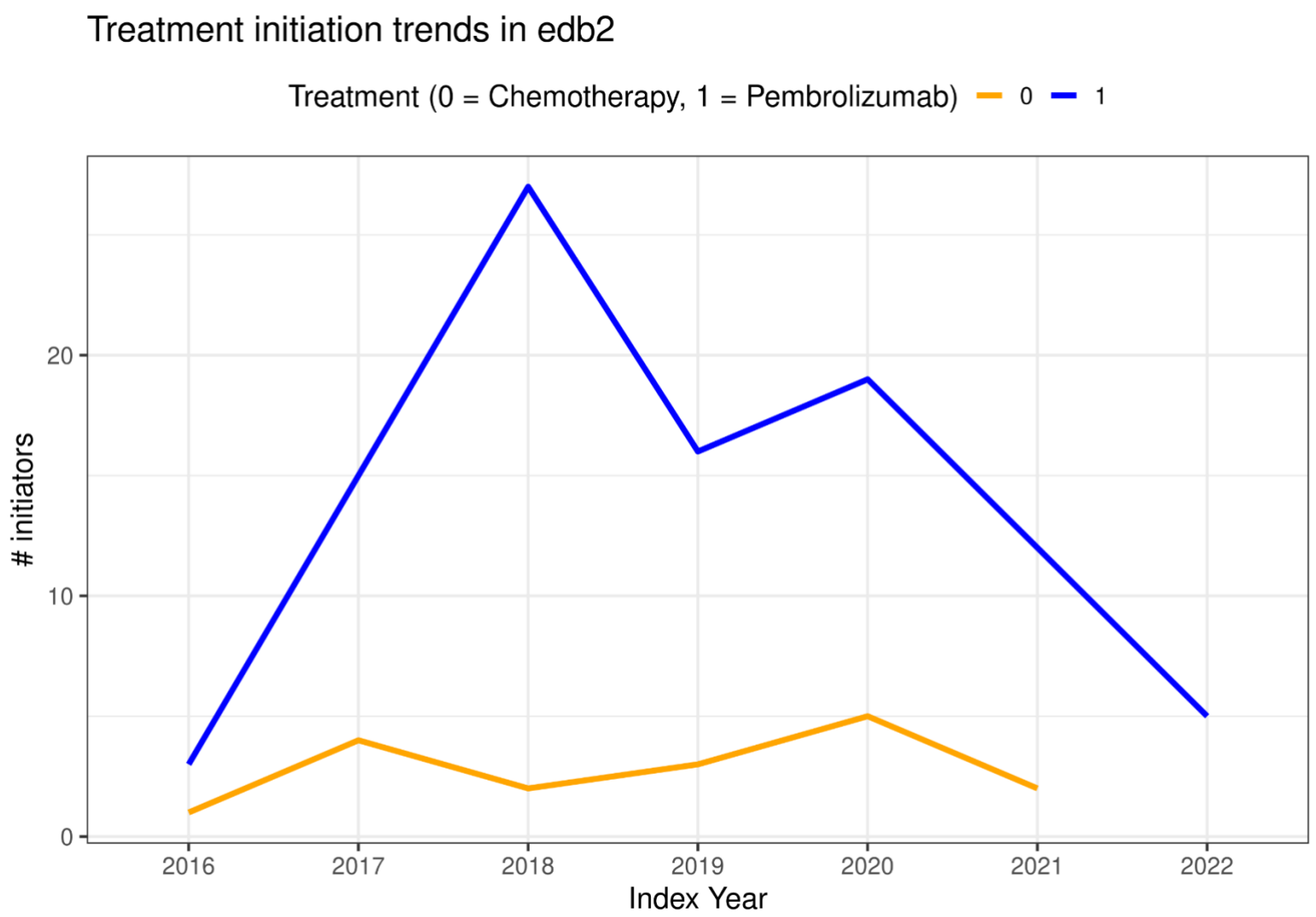
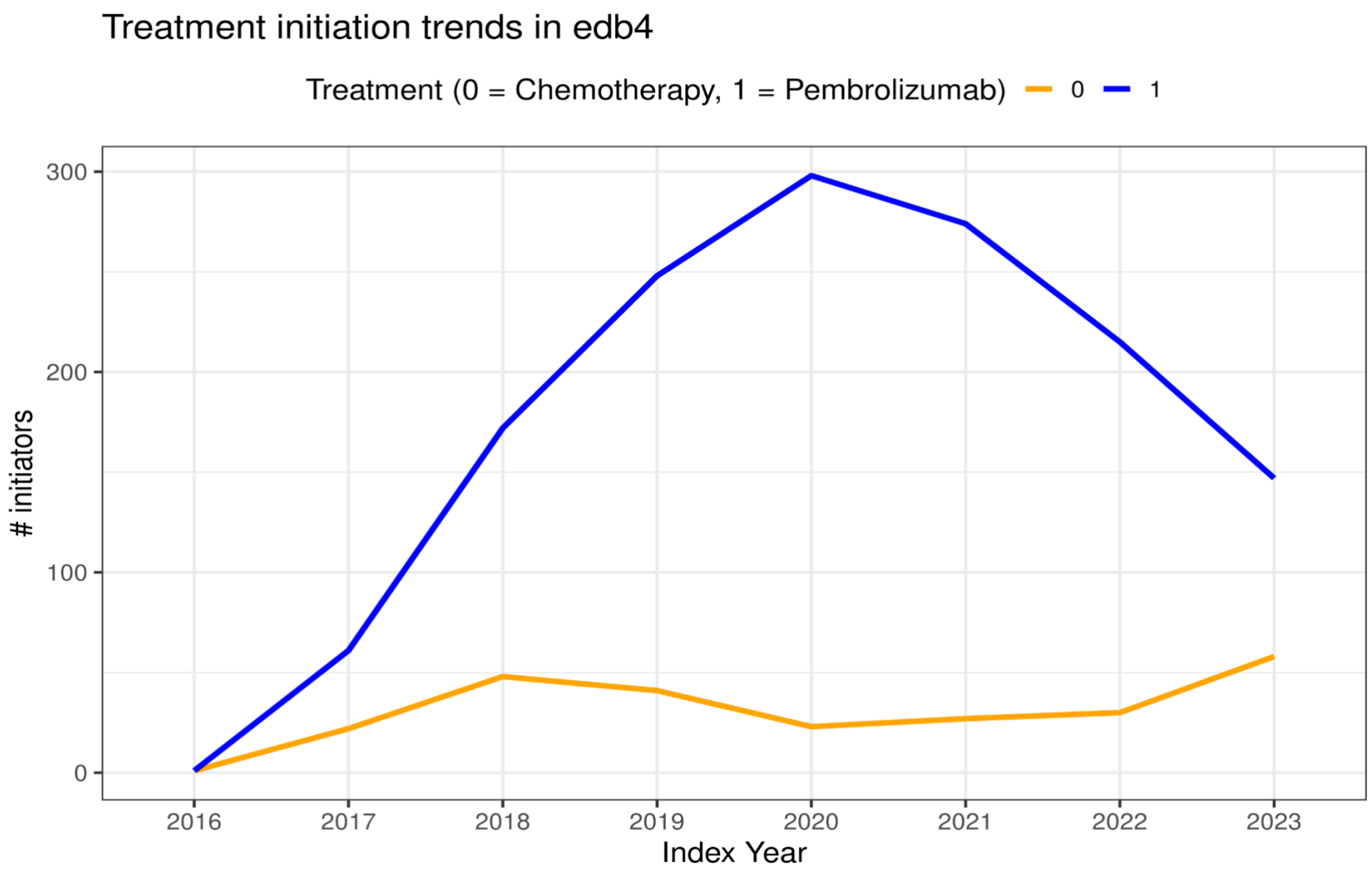


Figure 11. Treatment initiation trends by calendar year and treatment in EDB4.



NCTID	NCT02220894
Acronym	**Keynote-042**
Protocol	https://cdn.clinicaltrials.gov/large-docs/94/NCT02220894/Prot_SAP_001.pdf
SAP	https://cdn.clinicaltrials.gov/large-docs/94/NCT02220894/Prot_SAP_001.pdf
PMID	https://ascopubs.org/doi/10.1200/JCO.21.02885
Indication	Non–Small-Cell Lung Cancer and Programmed Death Ligand-1 Tumor Proportion Score ≥ 1%
Line of Therapy	1
Exposures	pembrolizumab 200 mg once every 3 weeks for 35 cycles
Comparisons	chemotherapy (carboplatin + paclitaxel or pemetrexed) for 4-6 cycles with optional maintenance pemetrexed.
Emulated outcome	Overall survival (primary end point)

Study of Pembrolizumab (MK-3475) Versus Platinum-Based Chemotherapy for Participants With Programmed Cell Death-Ligand 1 (PD-L1)-Positive Advanced or Metastatic Non-Small Cell Lung Cancer (MK-3475-042/KEYNOTE-042)

Measurement eligibility criteria						
Criteria	**Criteria rule as defined in original protocol**	**Clinical relevance**	**Emulation [EDB1]**	**Emulation [EDB2]**	**Emulation [EDB4]**	**Comment**
Inclusion 1	Have measurable disease based on RECIST 1.1 as determined by the site.	Limited Relevance	Not implementable	Not implementable	Not implementable	We assumed that patients initiating first-line therapy had RECIST 1.1 1.1-measurable disease.
Inclusion 2	Be ≥18 years of age on the day of signing informed consent.	Relevant	Possible	Possible	Possible	To protect privacy, most databases only provide month- or year-level granularity of dates
Inclusion 3	Have a life expectancy of at least 3 months.	Relevant	Not implementable	Not implementable	Not implementable	
Inclusion 4	Have not received prior systemic chemotherapy treatment for their advanced/metastatic NSCLC.	Relevant	Possible	Possible	Possible	The first systemic anticancer regimen for advanced/metastatic NSCLC
Inclusion 5	Have a performance status of 0 or 1 on the Eastern Cooperative Oncology Group (ECOG) Performance Status.	Relevant	Possible	Possible	Possible	ECOG implementation possible; high % missingness likely
Inclusion 6	Hematological: Absolute neutrophil count (ANC) ≥1,500/mcL, Platelets ≥100,000/mcL, Hemoglobin ≥9 g/dL or ≥5.6 mmol/L;	Limited Relevance	Not implementable	Not implementable	Not implementable	Not well captured
Inclusion 7	Have no history of prior malignancy, with the exception of basal cell carcinoma of the skin, superficial bladder cancer, squamous cell carcinoma of the skin, or in situ cervical cancer, or have undergone potentially curative therapy with no evidence of that disease recurrence for 5 years since initiation of that therapy.	Relevant	Possible	Possible	Possible	Patients with prior malignancy were excluded
Inclusion 8	Have provided formalin-fixed tumor tissue sample from a biopsy of a tumor lesion either at the time of or after the diagnosis of advanced or metastatic disease has been made AND from a site not previously irradiated to assess for PD-L1 status.	Limited Relevance	Not implementable	Not implementable	Not implementable	
Inclusion 9	Have a histologically or cytologically confirmed diagnosis of advanced or metastatic NSCLC and not have an EGFR sensitizing (activating) mutation or an ALK translocation.	Relevant	Limited	Limited	Limited	Pembrolizumab group: Exclude patients with documented EGFR-positive or ALK-positive tumors EGFR/ALK status is allowed to be missing/unknown. According to the FDA-approved label and the National Comprehensive Cancer Network (NCCN) Clinical Practice Guidelines, pembrolizumab monotherapy is indicated for aNSCLC without EGFR or ALK genomic tumor aberrations. Chemotherapy group: EGFR/ALK negativity required; exclude if EGFR/ALK is missing/unknown, or positive.
Inclusion 10	Have a PD-L1 positive (TPS≥1%) tumor as determined by IHC at a central laboratory. Note: Only PD-L1 positive (TPS≥1%) subjects will be randomized. If the tumor specimen is not evaluable for PD-L1 expression by the central laboratory, the subject is not eligible to participate in the study.	Relevant	Limited	Limited	Limited	PD-L1 ≥50% required; exclude if PD-L1 is missing/unknown or <50%. KEYNOTE-042 reported three primary overall survival outcomes among patients with PD-L1 ≥50%, ≥20%, and ≥1%, we will focus on the PD-L1 ≥50% subgroup as the primary analysis for this emulation because the vast majority of patients treated with pembrolizumab in clinical practice have PD-L1 ≥50%.
Inclusion 11	Female subjects must have a negative urine or serum pregnancy test at screening (within 72 hours of first dose of study medication) if of childbearing potential or be of non-child bearing potential. If the urine test is strong or cannot be confirmed as negative, a serum pregnancy test will be required. The serum pregnancy test must be negative for the subject to be eligible. Non-childbearing potential is defined as (by other than medical reasons): a. ≥45 years of age and has not had menses for greater than 1 year, b. Amenorrheic for <2 years without a hysterectomy and oophorectomy and an FSH value in the postmenopausal range upon pretrial (screening) evaluation, c. Whose status is post hysterectomy, oophorectomy, or tubal ligation. Documented hysterectomy or oophorectomy must be confirmed with medical records of the actual procedure or confirmed by an ultrasound. Tubal ligation must be confirmed with medical records of the actual procedure otherwise the subject must be willing to use two adequate barrier methods throughout the study, starting with the screening visit through 120 days after the last dose of study therapy. Please see Section 5.7.2 for a list of acceptable birth control methods. Information must be captured appropriately within the site's source documents.	Limited Relevance	Not implementable	Not implementable	Not implementable	
Inclusion 12	If of childbearing potential, female subjects must be willing to use two adequate barrier methods or a barrier method plus a hormonal method throughout the study, starting with the screening visit (Visit 1) through 120 days after the last dose of pembrolizumab is received and through 180 days after last dose of chemotherapeutic agents as specified in the protocol. Such methods of contraception, or true abstinence from heterosexual activity, when this is in line with the preferred and usual lifestyle of the subject, are required (periodic abstinence, e.g., calendar, ovulation, symptothermal, post-ovulation methods and withdrawal are not acceptable methods of contraception). Please see Section 5.7.2 for a list of acceptable birth control methods.	Relevant	Not implementable	Not implementable	Not implementable	This criterion was not directly captured in the data and was assumed to be met under routine clinical care for patients receiving systemic anticancer therapy (e.g., contraception counseling, pregnancy testing when applicable, and avoidance of breastfeeding).

Inclusion 13	Male subjects with a female partner(s) of child-bearing potential must agree to use two adequate barrier methods or a barrier method plus a hormonal method throughout the trial starting with the screening visit (Visit 1) through 120 days after the last dose of pembrolizumab is received and through 180 days after the last dose of chemotherapeutic agents as specified in the protocol. Such methods of contraception, or true abstinence from heterosexual activity, when this is in line with the preferred and usual lifestyle of the subject, are required (periodic abstinence, e.g., calendar, ovulation, symptothermal, post-ovulation methods and withdrawal are not acceptable methods of contraception). Males with pregnant partners must agree to use a condom; no additional method of contraception is required for the pregnant partner. Please see Section 5.7.2 for a list of acceptable birth control methods.	Relevant	Not implementable	Not implementable	Not implementable	
Inclusion 14	Have voluntarily agreed to participate by giving written informed consent/assent for the trial. The subject may also provide consent/assent for Future Biomedical Research. However, the subject may participate in the main trial without participating in Future Biomedical Research.	Limited Relevance	Not implementable	Not implementable	Not implementable	
Exclusion 1	Has an EGFR sensitizing mutation and/or ALK translocation. Note: For patients enrolled who are known to have a tumor of predominantly squamous histology, molecular testing for EGFR mutation and ALK translocation will not be required as this is not standard of care and is not part of current diagnostic guidelines.	Relevant	Limited	Limited	Limited	Pembrolizumab group: Exclude patients with documented EGFR- or ALK-positive tumors; EGFR/ALK status is allowed to be missing/unknown. According to the FDA-approved label and the National Comprehensive Cancer Network (NCCN) Clinical Practice Guidelines, pembrolizumab monotherapy is indicated for aNSCLC without EGFR or ALK genomic tumor aberrations. Chemotherapy group: EGFR/ALK negativity required; exclude if EGFR/ALK is missing/unknown, or positive.
Exclusion 2	Is currently participating and receiving study therapy or has participated in a study of an investigational agent and received study therapy or used an investigational device within 4 weeks of the first dose of trial treatment.	Limited Relevance	Limited	Limited	Limited	We excluded patients who received an investigational agent.
Exclusion 3	Tumor specimen is not evaluable for PD-L1 expression by the central laboratory. If an additional tumor specimen is submitted AND evaluable for PD-L1 expression, the subject will be eligible to participate if PD-L1 expression is assessed as positive (TPS ≥1%) by the central laboratory.	Relevant	Limited	Limited	Limited	PD-L1 ≥50% required; exclude if PD-L1 is missing/unknown or <50% and the trial results suggested heterogeneity of effects by PD-L1 level. KEYNOTE-042 reported three primary overall survival outcomes among patients with PD-L1 ≥50%, ≥20%, and ≥1%. We will focus on the PD-L1 ≥50% subgroup as the primary analysis for this emulation because the vast majority of patients treated with pembrolizumab in clinical practice have PD-L1 ≥50%.
Exclusion 4	Subjects with squamous histology who received carboplatin in combination with paclitaxel in the adjuvant setting.	Limited Relevance	Limited	Limited	Limited	Patients with squamous histology who previously received carboplatin in combination with paclitaxel were excluded.
Exclusion 5	Is receiving systemic steroid therapy ≤3 days prior to the first dose of trial treatment or receiving any other form of immunosuppressive medication. Notes: a. Corticosteroid use on study after Cycle 1 for management of AEs, SAEs and ECLs, as a pre-medication for the control chemotherapies, as a pre-medication for IV contrast allergies/reactions or if considered necessary for a subject's welfare is allowed. b. Subjects who receive daily steroid replacement therapy serve as an exception to this rule. Daily prednisone at doses of 5-7.5 mg is an example of replacement therapy. c. Equivalent hydrocortisone doses are also permitted if administered as a replacement therapy.	Limited Relevance	Not implementable	Not implementable	Not implementable	Not well captured
Exclusion 6	The subject's NSCLC can be treated with curative intent with either surgical resection and/or chemoradiation.	Relevant	Possible	Possible	Possible	Operationalized by excluding patients with stage potentially amenable to curative-intent therapy (stage IIIA) where staging data were available.
Exclusion 7	Is expected to require any other form of systemic or localized antineoplastic therapy while on trial (including maintenance therapy with another agent for NSCLC, radiation therapy, and/or surgical resection).	Relevant	Not implementable	Not implementable	Not implementable	
Exclusion 8	Has received any prior systemic cytotoxic chemotherapy, biological therapy OR had major surgery within 3 weeks of the first dose of trial treatment; received lung radiation therapy of >30 Gy within 6 months of the first dose of trial treatment.	Limited Relevance	Limited	Limited	Limited	Patients who had received any prior systemic cytotoxic chemotherapy or biologic therapy within 3 weeks before index date were excluded.
Exclusion 9	Has received prior therapy with an anti-PD-1, anti-PD-L1, anti-PD-L2, anti-CD137, or anti-cytotoxic T-lymphocyte-associated antigen-4 (CTLA-4) antibody (including ipilimumab or any other antibody or drug specifically targeting T-cell co-stimulation or checkpoint pathways).	Limited Relevance	Possible	Possible	Possible	Patients with any prior exposure to immune checkpoint inhibitors were excluded, including ipilimumab, nivolumab, pembrolizumab, atezolizumab, durvalumab, avelumab, cemiplimab, tremelimumab, and dostarlimab.
Exclusion 10	Has known central nervous system metastases and/or carcinomatous meningitis. Note: Subjects with previously treated brain metastases may participate provided they are clinically stable (neurologically asymptomatic) and have no evidence of new or enlarging brain metastasis by imaging at least 4 weeks after treatment of the brain metastases (e.g., surgery, RT) and are off steroids for at least 3 days prior to the first dose of study medication.	Limited Relevance	Possible	Possible	Possible	We excluded patients with CNS involvement, defined by diagnoses of brain metastases, cerebral meningeal metastases, or metastases to other parts of the nervous system.
Exclusion 11	Has active autoimmune disease that has required systemic treatment in past 2 years (i.e., with use of disease modifying agents, corticosteroids or immunosuppressive drugs). Note: Replacement therapy (i.e., thyroxine, insulin, or physiologic corticosteroid replacement therapy for adrenal or pituitary insufficiency, etc.) is not considered a form of systemic treatment. Subjects that require inhaled corticosteroids would not be excluded from the study. Subjects with vitiligo or resolved childhood asthma/atopy would not be excluded from the study. Subjects that require local steroid injections would not be excluded from the study.	Relevant	Limited	Limited	Limited	Patients with a history of clinically significant autoimmune disease within two years prior to treatment initiation were excluded.Autoimmune diseases are likely underdocumented in the data.
Exclusion 12	Has had an allogeneic tissue/solid organ transplant.	Relevant	Not implementable	Not implementable	Not implementable	
Exclusion 13	Has interstitial lung disease OR has had a history of pneumonitis that has required oral or IV steroids.	Relevant	Limited	Limited	Limited	Exclude patients with prior diagnosis of interstitial lung disease.
Exclusion 14	Has received or will receive a live vaccine within 30 days prior to the first administration of study medication. Seasonal flu vaccines that do not contain live vaccine are permitted.	Limited Relevance	Not implementable	Not implementable	Not implementable	Not well captured
Exclusion 15	Has an active infection requiring intravenous systemic therapy.	Limited Relevance	Not implementable	Not implementable	Not implementable	
Exclusion 16	Has known history of Human Immunodeficiency Virus (HIV) (HIV 1/2 antibodies). Note: HIV testing is required at screening as indicated in Section 6.0 – Trial Flow Chart.	Limited Relevance	Not implementable	Not implementable	Not implementable	
Exclusion 17	Has known active Hepatitis B or C. Subjects with a positive HBsAg result would be excluded. Active Hepatitis C is defined by a known positive Hep C Ab result and known quantitative HCV RNA results greater than the lower limits of detection of the assay. Note: Hepatitis B and Hepatitis C testing is required at screening as indicated in Section 6.0 – Trial Flow Chart.	Limited Relevance	Not implementable	Not implementable	Not implementable	
Exclusion 18	Has a history or current evidence of any condition, therapy, or laboratory abnormality that might confound the results of the trial, interfere with the subject's participation for the full duration of the trial, or is not in the best interest of the subject to participate, in the opinion of the treating Investigator.	Relevant	Not implementable	Not implementable	Not implementable	
Exclusion 19	Has known psychiatric or substance abuse disorders that would interfere with cooperation with the requirements of the trial.	Relevant	Not implementable	Not implementable	Not implementable	
Exclusion 20	Is, at the time of signing informed consent, a regular user (including "recreational use") of any illicit drugs or had a recent history (within the last year) of substance abuse (including alcohol).	Relevant	Not implementable	Not implementable	Not implementable	
Exclusion 21	Is pregnant or breastfeeding, or expecting to conceive or father children within the projected duration of the trial, starting with the screening visit (Visit 1) through 120 days after the last dose of pembrolizumab or 180 days after the last dose of SOC chemotherapy.	Relevant	Not implementable	Not implementable	Not implementable	

Autoimmune diseasi	inflammatory bowel disease; systemic lupus erythematosus; dermatomyositis; scleroderma; vasculitis; polyarteritis nodosa; sarcoidosis; immune thrombocytopenic purpura; hemolytic anemia; multiple sclerosis

NCTID NCT02220894
Acronym **Keynote-042** Study of Pembrolizumab (MK-3475) Versus Platinum-Based Chemotherapy for Participants With Programmed Cell Death-Ligand 1 (PD-L1)-Positive Advanced or Metastatic Non-Small Cell Lung Cancer (MK-3475-042/KEYNOTE-042)
Protocol https://cdn.clinicaltrials.gov/large-docs/94/NCT02220894/Prot_SAP_001.pdf
SAP https://cdn.clinicaltrials.gov/large-docs/94/NCT02220894/Prot_SAP_001.pdf
PMID https://ascopubs.org/doi/10.1200/JCO.21.02885
Indication Non–Small-Cell Lung Cancer and Programmed Death Ligand-1 Tumor Proportion Score ≥ 1%
Line of Therapy 1
Exposures pembrolizumab 200 mg once every 3 weeks for 35 cycles
Comparisons chemotherapy (carboplatin + paclitaxel or pemetrexed) for 4-6 cycles with optional maintenance pemetrexed.
Emulated outcome Overall survival (primary end point)

Measurement eligibility criteria							**Comment**	**encore.io function**
Criteria	**Criteria in trial protocol**	**Criteria rule as defined in original protocol**	**Time point/period of emulated measurement [days]**	**Emulation [EDB1]**	**Emulation [EDB2]**	**Emulation [EDB4]**		
Inclusion	Inclusion 2	Men and women ≥ 18 years of age	[0;0]	Age at index date (year granularity level)	Age at index date (year granularity level)	Age at index date (year granularity level)		edbx_get_demographics()
Inclusion	Inclusion 4	Have not received prior systemic chemotherapy treatment for their advanced/metastatic NSCLC.	[AMND;-1]	No systemic chemotherapy between the initial diagnosis of advanced/metastatic disease and the index date.	No systemic chemotherapy between the initial diagnosis of metastatic disease and the index date.	No systemic chemotherapy between the initial diagnosis of metastatic disease and the index date.		
Inclusion	Inclusion 5	Eastern Cooperative Oncology Group (ECOG) performance status of ≤ 1	[-90;0]	ECOG = 0 | 1	ECOG = 0 | 1	ECOG = 0 | 1		edbx_get_ecog()
Inclusion	Inclusion 7	Have no history of prior malignancy, with the exception of basal cell carcinoma of the skin, superficial bladder cancer, squamous cell carcinoma of the skin, or in situ cervical cancer, or have undergone potentially curative therapy with no evidence of that disease recurrence for 5 years since initiation of that therapy.	[-inf;0]	Record of non-lung cancer diagnosis other than basal cell carcinoma of the skin, superficial bladder cancer, squamous cell carcinoma of the skin, or in situ cervical cancer prior to or on the index date.	Record of non-lung cancer diagnosis other than basal cell carcinoma of the skin, superficial bladder cancer, squamous cell carcinoma of the skin, or in situ cervical cancer prior to or on the index date.	Record of non-index cancer treatment prior to or on the index date.	5-year recurrence-free could not be fully operationalized. No diagnosis table for EDB4 available	
Inclusion	Inclusion 9; exclusion 6	Have histologically or cytologically confirmed advanced or metastatic NSCLC that is not amenable to curative-intent treatment (surgical resection and/or definitive chemoradiation).	[-inf;0]	Line of therapy needs to be for "Advanced" setting (LoT table)	Line of therapy needs to be for "Metastatic" setting (LoT table)	Any evidence of at least one distant metastasis at any time before the index date (inclusive). This captures both de novo metastatic patients and those who progressed/developed metastases before/on the index date.	EDB1 included patients diagnosis with stage IIIB through IV or recurrence/progression to metastastic after earlier stage diagnosis. Since advanced NSCLC not amenable to curative surgery or radiotherapy is difficult to ascertain in EHR, EDB2 and EDB4 are restricted to metastatic patients only which is easier to assess.	EDB1 and EDB2: Part of exposure definition EDB4: Derived form edbx_get_diagnosis()
Exclusion	Inclusion 9; exclusion 1	Have EGFR sensitizing (activating) mutation or an ALK translocation.	[-180;0]	Pembrolizumab group: Exclude patients with documented EGFR-positive or ALK-positive tumors; EGFR/ALK status is allowed to be missing/unknown. Chemotherapy group: EGFR/ALK negativity required; exclude if EGFR/ALK is missing/unknown, or positive.	Pembrolizumab group: Exclude patients with documented EGFR-positive or ALK-positive tumors; EGFR/ALK status is allowed to be missing/unknown. Chemotherapy group: EGFR/ALK negativity required; exclude if EGFR/ALK is missing/unknown, or positive.	Pembrolizumab group: Exclude patients with documented EGFR-positive or ALK-positive tumors; EGFR/ALK status is allowed to be missing/unknown. Chemotherapy group: EGFR/ALK negativity required; exclude if EGFR/ALK is missing/unknown, or positive.	High % missingness likely. EGFR sensitizing mutations and ALK rearrangements were captured in EDB1, EDB2, and EDB4 within 180 days before or on the index date. According to the FDA-approved label and the National Comprehensive Cancer Network (NCCN) Clinical Practice Guidelines, pembrolizumab monotherapy is indicated for aNSCLC without EGFR or ALK genomic tumor aberrations. Therefore, patients with no documented EGFR/ALK aberrations were considered eligible for pembrolizumab monotherapy.	
Exclusion	Inclusion 10; exclusion 3	The tumor specimen is not evaluable for PD-L1 expression or PD-L1-negative (TPS < 1%) as determined by IHC at a central laboratory.	[-90;0]	PD-L1 ≥50% required; exclude if PD-L1 is missing/unknown or <50%. Acceptable PD-L1 assays: Pembrolizumab group: includes 22C3, 28-8, E1L3N, other assays, or unknown/not documented; SP142 and SP263 are not permitted. Chemotherapy group: PD-L1 testing must be performed using 22C3, 28-8, or E1L3N.	PD-L1 ≥50% required; exclude if PD-L1 is missing/unknown or <50%. Pembrolizumab group: Accept PD-L1 results reported as TPS, TC, or missing/unknown. Chemotherapy group: Accept PD-L1 results reported as TPS or TC.	Require PD-L1 positivity; Exclude patients with missing/unknown or negative PD-L1 results.	High % missingness likely. KEYNOTE-042 reported three primary overall survival outcomes among patients with PD-L1 ≥50%, ≥20%, and ≥1%. We will focus on the PD-L1 ≥50% subgroup as the primary analysis for this emulation because the vast majority of patients treated with pembrolizumab in clinical practice have PD-L1 ≥50% and the trial results suggested heterogeneity in effect by PD-L1 status.	

Exclusion	Exclusion 2	Have received study therapy or have participated in a study of an investigational agent and received study therapy or used an investigational device within 4 weeks of the first dose of trial treatment.	[-28;0]	Exclude patients who received investigational agents within 4 weeks prior to the index date.	Exclude patients who received investigational agents within 4 weeks prior to the index date.	Exclude patients who received investigational agents within 4 weeks prior to the index date.		
Exclusion	Exclusion 3	Subjects with squamous histology who received carboplatin in combination with paclitaxel in the adjuvant setting.	[-inf;AMND-1]	Identify adjuvant carboplatin + paclitaxel prior to advanced/metastatic diagnosis among patients with squamous histology	Identify adjuvant carboplatin + paclitaxel prior to metastatic diagnosis among patients with squamous histology	Identify adjuvant carboplatin + paclitaxel prior to metastatic diagnosis among patients with squamous histology		
Exclusion	Exclusion 4	Has received any prior systemic cytotoxic chemotherapy, biological therapy OR had major surgery within 3 weeks of the first dose of trial treatment	[-21;-1]	Prior systemic cytotoxic chemotherapy or biologic therapy within 21 days before index date	Prior systemic cytotoxic chemotherapy or biologic therapy within 21 days before index date	Prior systemic cytotoxic chemotherapy or biologic therapy within 21 days before index date		
Exclusion	Exclusion 9	Has received prior therapy with an anti–PD-1, anti–PD-L1, anti–PD-L2, anti-CD137, or anti–CTLA-4 antibody.	[-inf;-1]	Exclude patients who received Immunotherapy before the index date	Exclude patients who received Immunotherapy before the index date	Exclude patients who received Immunotherapy before the index date		
Exclusion	Exclusion 10	Has known central nervous system metastases and/or carcinomatous meningitis.	[-inf;0]	Exclude patients who had a diagnosis of brain metastases, cerebral meningeal metastases, or metastases to other parts of nervous system before the index date	Exclude patients who had a diagnosis of brain metastases, cerebral meningeal metastases, or metastases to other parts of nervous system before the index date	Exclude patients who had a diagnosis of brain metastases, cerebral meningeal metastases, or metastases to other parts of nervous system before the index date		
Exclusion	Exclusion 11	Has active autoimmune disease that has required systemic treatment in past 2 years.	[-730;0]	Patients with a history of autoimmune disease within 2 years before treatment initiation were excluded.	Patients with a history of autoimmune disease within 2 years before treatment initiation were excluded.	Patients with a history of autoimmune disease within 2 years before treatment initiation were excluded.	Autoimmune disease history is likely underdocumented in the data.	
Exclusion	Exclusion 13	Has interstitial lung disease OR has had a history of pneumonitis that has required oral or IV steroids.	[-inf;0]	Exclude patients who were diagnosed with interstitial lung disease before the index date.	Exclude patients who were diagnosed with interstitial lung disease before the index date.	Exclude patients who were diagnosed with interstitial lung disease before the index date.	Interstitial lung disease is likely underdocumented in the data.	

AMND = advanced/metastatic non-small cell lung cancer diagnosis

Immunotherapy	Ipilimumab, Nivolumab, Pembrolizumab, Atezolizumab, Durvalumab, Avelumab, Cemiplimab, Tremelimumab, Dostarlimab
Autoimmune diseases	Inflammatory bowel disease, Systemic lupus erythematosus, Dermatomyositis, Scleroderma, Vasculitis, Polyarteritis nodosa, Sarcoidosis, Immune thrombocytopenic purpura, Hemolytic anemia, Multiple sclerosis
systemic cytotoxic chemotherapy	Cisplatin, Carboplatin, Pemetrexed, Paclitaxel, Nab-paclitaxel, Docetaxel, Gemcitabine, Vinorelbine, Etoposide
biological therapy	Bevacizumab, Ramucirumab, Necitumumab, Cetuximab