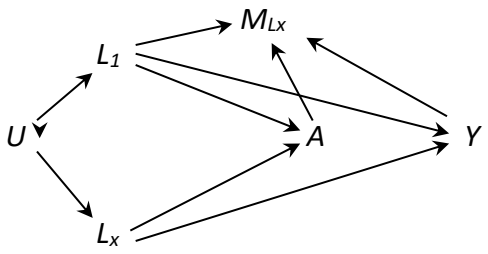
Supplemental Online Content

Merola D, Campbell U, Lenis D, et al. Calibrating observational health record data against a randomized trial. *JAMA Netw Open*. 2024;7(9):e2436535. doi:10.1001/jamanetworkopen.2024.36535

This supplemental material has been provided by the authors to give readers additional information about their work.

eMethods 1. Missing Data: Mechanisms, Assumptions, and Methodological Approach



Legend

A : A binary indicator of treatment group
Y : A binary indicator of the outcome event
L_1 : A vector of completely observed confounders
L_x : A vector of partially observed confounders
M_{Lx} : A vector of binary indicators for missingness among variables in the vector L_x

Causal Directed Acyclic Graph (DAG) depicting the hypothesized data-generating process of study variables and missing data for the intention-to-treat analysis.

Theoretical mechanisms of missing data, justification of methodological approach, and assumptions:
The data source is comprised of data drawn from over 50 healthcare networks, some of which may contribute to certain variables to varying degrees. That is, network-specific practices might lead to specific patterns of missing data. This is one plausible reason for missing values and, while identifying information on contributing healthcare networks is not available in the database, it can be proxied by patient regional location. This hypothesis is supported by an observed dependence of missing indicators for region-specific variables. Therefore, the data are believed to be Missing at Random (MAR). That is, the missing data mechanism depends on observed variables.

The hypothesized relationships between missingness, exposure, outcome, and confounders are conveyed in the Directed Acyclic Graph (DAG) above. In the DAG, missingness for each patient characteristic is represented as its own random variable, contained in the vector M_{Lx}. Missingness is proposed to be a function of the observed data, particularly regional location of each patient (a component of the L_1 vector). Our assumption that missingness is a function of the observed variables is conveyed by arrows from L_1, A, Y to the M_{Lx} vector. Note that the variables that contain missing values (L_x) or unobserved variables (U) are not assumed to cause missingness (M_{Lx}).

Multiple imputation with chained equations (fully conditional specification) will be employed to address missing values that are MAR. This approach can lead to valid estimates for data that are MAR under the following sets of assumptions,[1] which are encoded in the DAG above:

I. Traditional identifiability assumptions necessary for causal inference:

1. Exchangeability: The counterfactual outcome $Y^{A=a,M_{Lx}=0}$ in a world when we intervene on 'A' and 'M_{Lx}' is independent of A, conditional on (L_1, L_x)
2. Consistency: 'Interventions' of no missing data and treatment, represented by $M_{Lx} = 0$ and $A = a$, respectively, and the outcome, Y, are sufficiently well-defined such that the counterfactual outcome $Y^{A=a,M_{Lx}=0}$ corresponds to the outcome that would have occurred had these interventions been (factually) imposed on the study population
3. Positivity: The probability of being in a given treatment group is greater than 0 within every combination of levels of the confounding variables
4. No mis-specification of the imputation and propensity score models

II. Additional assumptions needed in the context of missing data:

1. There are no unmeasured common causes of the missingness indicators M_{Lx} and other variables in the causal DAG above, including other missingness indicators represented in the vector M_{Lx}
2. The variables represented in the vector L_x do not cause missingness in themselves or other variables
3. The missing indicators M_{Lx} do not cause any other variables in the causal DAG shown above

4. M_{Lx} is independent of L_x, conditional on (L_1, A, Y)

Using the same analytic approach described below, the effect of interest may also be validly estimated under different sets of relationships than those encoded in the DAG. For example, if missingness (M_{Lx}) was not caused by the exposure (A) and/or the outcome (Y), the effect of treatment could still be ascertained because all of the above assumptions would still hold.[1] However, results from this analysis may not be valid if the aforementioned assumptions are violated, particularly if missingness of a variable is caused by that variable itself or if the missing data was caused by unobserved factors. Imputation models that are incorrectly specified or not compatible with the substantive (outcome) model will also lead to biased estimates.[2]

Description of missing data:
There are 7 variables with missing values as shown in eTable 3. Notably, missing values in 'Performance Status' and 'PD-L1 Percent' appear to occur differentially with respect to treatment group, while missingness in all other variables appears to occur non-differentially.

Methodological Approach:
For both, the intention-to-treat and per-protocol analyses, a series of conditional models were be specified to estimate the parameters of a joint distribution from which the imputed values will be drawn for each variable with missingness shown in eTable 3. Predictive mean matching, ordered logistic regression, and multinomial logistic models were used to impute continuous, ordinal, and unordered categorical variables that contain missing values, respectively. Imputation models were specified in a manner that is compatible with the substantive model for the outcome analysis.[2] For example, if the substantive (outcome) model specifies a linear relation between two variables, this will be maintained in the imputation models. All observed variables used in the outcome analysis, including the event indicator for the outcome and follow-up time, were included in the imputation models.[1–4] For the per-protocol analysis, additional interaction terms between time and other analysis variables were included where possible, such that the imputation was conducted within 'blocks' of proximal time intervals.[4]

The 'mice' package (v3.14.0) in R was used to conduct the imputation and generate 10 imputed datasets. The substantive (outcome) analysis conveyed in the protocol was then conducted within each of these datasets and final estimates were then pooled together using Rubin's rules to incorporate uncertainty due to missingness into the variance estimates.[5]

eMethods 2. Contemporaneous Cohort Sensitivity Analysis

Description: The main analysis utilized historical controls, resulting in treatment groups that differ with respect to calendar time. In this sensitivity analysis, the study cohort was restricted to patients that had a cohort entry date of January 1, 2018 or later. Then, the analyses were repeated among this cohort (e.g., weighting, imputation, outcome estimation, etc.).

Results & Interpretation: The main intention-to-treat analysis was robust to the assumptions made regarding the use of 'historical controls' (eTable 5). There does not appear to be a strong indication that time was a confounder based on this sensitivity analysis.

eMethods 3. Exposure Definition Sensitivity Analysis

Description: The main analysis utilized RxNorm, NDC, and HCPCS/CPT procedure codes to ascertain exposure events. Because this approach did not clearly distinguish medication orders, administrations, and dispensings, this could have led to exposure misclassification. In this sensitivity analysis, treatments were only defined using HCPCS/CPT procedure codes, which are standardized medical billing codes for medication administration. Then, the main analyses were repeated among the resultant cohort.

Results & Interpretation: The results of the main intention-to-treat analysis were robust to the exposure definition (eTable 6). There is no evidence of misclassification bias due to the approach of treatment ascertainment used in the main analysis.

eMethods 4. Treatment Start Time Restriction Sensitivity Analysis

Description: In the main analysis, patients may have initiated treatment long after having an initial indication of metastatic disease in the data source. This may have resulted in a population that is substantially different than that observed in the KEYNOTE-189 trial, where patients initiated treatment soon after randomization. In this sensitivity analysis, the time from initial occurrence of metastatic disease to treatment initiation was restricted to 6 months or less. Then, the main analyses were repeated among the resultant cohort.

Results & Interpretation: The results of the main intention-to-treat analysis were robust to the time allowed between first indication of metastatic disease and treatment initiation (eTable 7). There was no evidence that this time period could explain the differences observed between the real-world evidence study and KEYNOTE-189 trial.

eMethods 5. Differential Treatment Intensity Exploratory Analysis

Description: Unmeasured confounding may have been present in the main analysis. In this exploratory analysis, the mean rate of radiotherapy and vitals encounters over follow-up was estimated by treatment group. Assessment of radiotherapy and vitals events was intended to capture the intensity in which patients are being followed by providers, which may indirectly indicate the presence of unmeasured confounding if non-differential with respect to treatment group.

Results & Interpretation: The rate of radiotherapy and vitals encounters over follow-up was similar between the treatment groups (eTable 8). There is no strong indication of unmeasured confounding based on this analysis.

eMethods 6. Censoring Event Distribution Exploratory Analysis

Description: The censoring time distribution of patients without outcome (mortality) events and who were not censored due to administrative end of study period in the intention-to-treat cohort were examined visually for differences between treatment group. Differences in the distribution of these censoring times could suggest the presence of informative censoring. Frequency and density plots before and after inverse probability of treatment weighting were reported. Only plots from the first imputed dataset were shown for brevity, as the distribution was similar across each imputed dataset.

Results & Interpretation: There was a similar bi-modal distribution of the censoring times between the exposed and comparator groups (eFigure 3) and no strong indication of informative censoring based on this analysis.

eMethods 7. Exposure Regimen Ascertainment Window Expansion Post-Hoc Analysis

Description: The 14-day exposure regimen ascertainment window used to ascertain exposure in the main intention-to-treat analysis may have resulted in misclassification, particularly among comparator patients that were awaiting insurance coverage of the exposure drug. In this post-hoc analysis, which was not pre-specified in the protocol, the exposure ascertainment window was expanded from 14 to 45 days.

Results & Interpretation: The results of the main intention-to-treat analysis were robust to this adjustment of the exposure ascertainment window (eTable 9).

eMethods 8. Comparator Percent Crossover Post-Hoc Analysis

Description: Crossover of the comparator patients into the exposure group during follow-up may result in a bias towards the null. For this reason, a post-hoc analysis that was not pre-specified in the study protocol was conducted to evaluate the percent of comparator patients that crossed over during follow-up.

Results & Interpretation: A moderate degree of cross-over from the comparator group to the exposure group was observed, which could (at least in part) explain the null finding in the real-world evidence study relative to the KEYNOTE-189 trial (eTable 10).

eMethods 9. De-Novo Metastatic Disease Post-Hoc Analysis

Description: De-novo metastatic disease is a negative prognosticator in non-small cell lung cancer that was not specified as a confounder in the protocol (i.e., a potential unmeasured confounder). The frequency of patients with de-novo metastatic disease was quantified in each treatment group was reported.

Results & Interpretation: A larger proportion of patients in the comparator group were identified as having de-novo metastatic disease (eTable 11). Because de-novo metastatic disease is a negative prognosticator, we would expect to see a bias away from the null, favoring the exposure group. This was contrary to what was observed in the main (intention-to-treat) analyses.

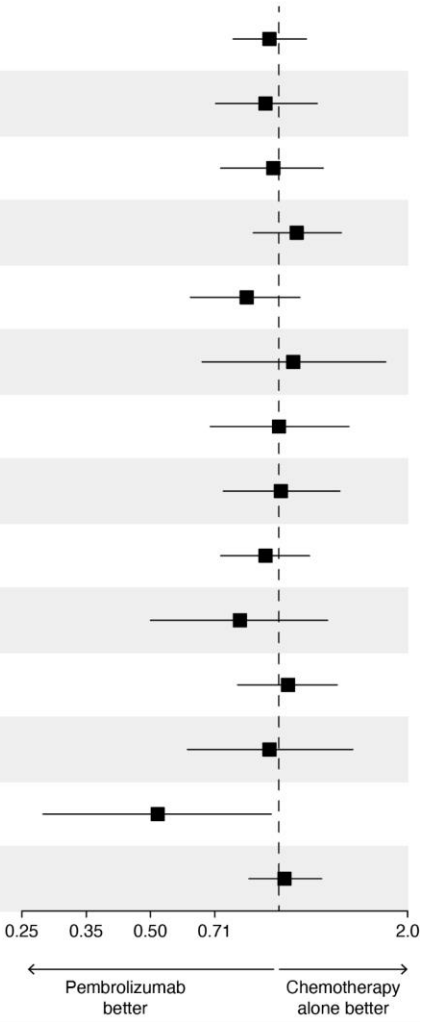
eMethods 10. Follow-up Time Distribution Post-Hoc Analysis

Description: An insufficient follow-up time may preclude observation of the treatment effect. The density of follow-up times in the intention-to-treat and per-protocol cohorts were plotted.

Results & Interpretation: Follow-up in the per-protocol cohort (eFigure 5) was not long enough for an effect to materialize, but appeared sufficient in the intention-to-treat cohort (eFigure 4). This was likely due to the stringent specification of the exposure contrast that was imposed on the population.

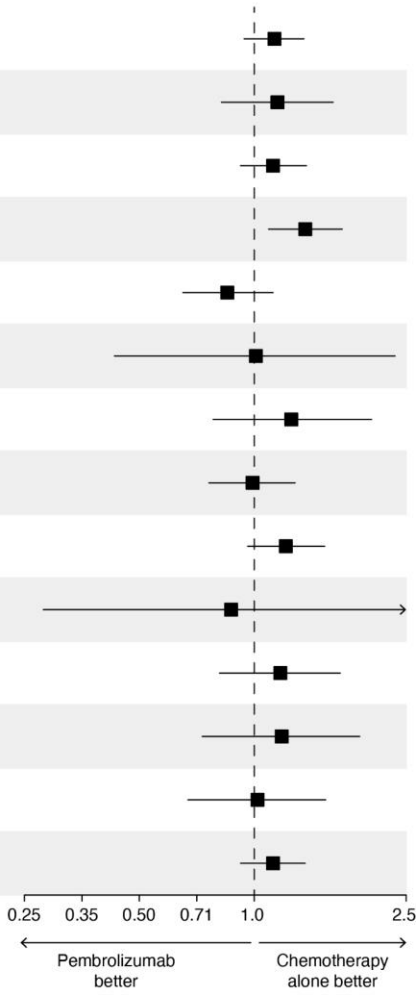
eFigure 1. Subgroup Analysis of the Intention-to-Treat Cohort

Subgroup	N (Exposed)	N (Comparator)	HR (95% CI)
All Patients	567	1277	0.95 (0.78, 1.16)
Age < 65	270	671	0.93 (0.71, 1.23)
Age >= 65	297	606	0.97 (0.73, 1.27)
Male	277	665	1.10 (0.87, 1.40)
Female	290	612	0.84 (0.62, 1.12)
Performance Status – 0	161	332	1.08 (0.66, 1.78)
Performance Status – 1	213	478	1.00 (0.69, 1.46)
Current or Former Smoker	177	434	1.01 (0.74, 1.39)
Never Smoker	391	843	0.93 (0.73, 1.18)
PD–L1 < 1%	175	421	0.81 (0.50, 1.30)
PD–L1 1% to 49%	247	533	1.05 (0.80, 1.37)
PD–L1 >= 50%	146	323	0.95 (0.61, 1.49)
De Novo – Yes	87	381	0.52 (0.28, 0.96)
De Novo – No	480	896	1.03 (0.85, 1.26)



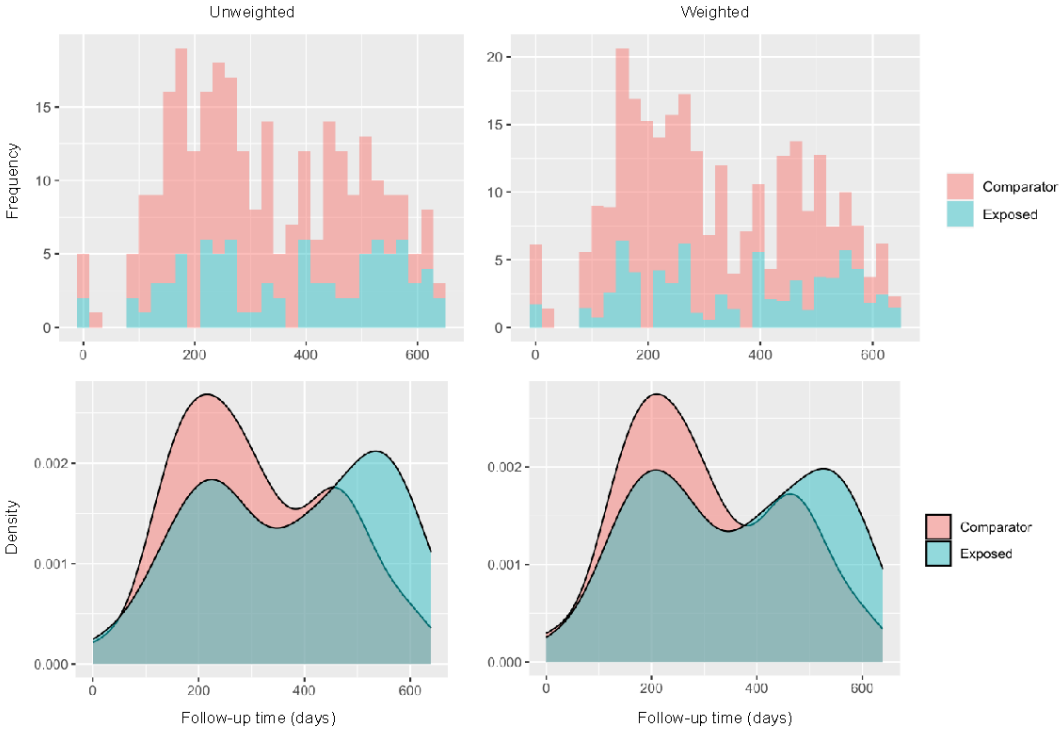
eFigure 2. Subgroup Analysis of the Per-Protocol Cohort

Subgroup	N (Exposed)	N (Comparator)	HR (95% CI)
All Patients	4012	4010	1.13 (0.94, 1.35)
Age < 65	2085	2023	1.15 (0.82, 1.61)
Age >= 65	1927	1987	1.12 (0.92, 1.37)
Male	2039	2048	1.36 (1.09, 1.70)
Female	1974	1962	0.85 (0.65, 1.12)
Performance Status – 0	789	792	1.01 (0.43, 2.34)
Performance Status – 1	1824	1810	1.25 (0.78, 2.03)
Current or Former Smoker	1371	1394	0.99 (0.76, 1.28)
Never Smoker	2641	2616	1.21 (0.96, 1.53)
PD–L1 < 1%	597	543	0.87 (0.28, 2.74)
PD–L1 1% to 49%	1848	1973	1.17 (0.81, 1.68)
PD–L1 >= 50%	1568	1494	1.18 (0.73, 1.89)
De Novo – Yes	1017	951	1.02 (0.67, 1.54)
De Novo – No	2995	3059	1.12 (0.92, 1.36)

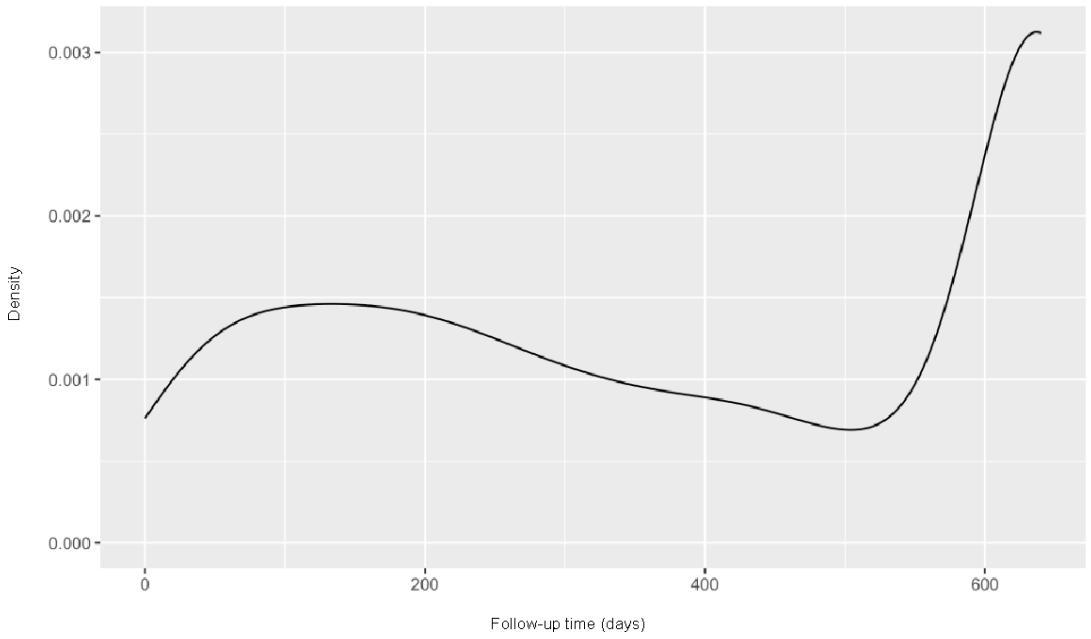


Note: To facilitate convergence of statistical models in sparse subgroups, a lower order time functional was used in the statistical model for the IP weights relative to that in the per-protocol analysis of the larger population.

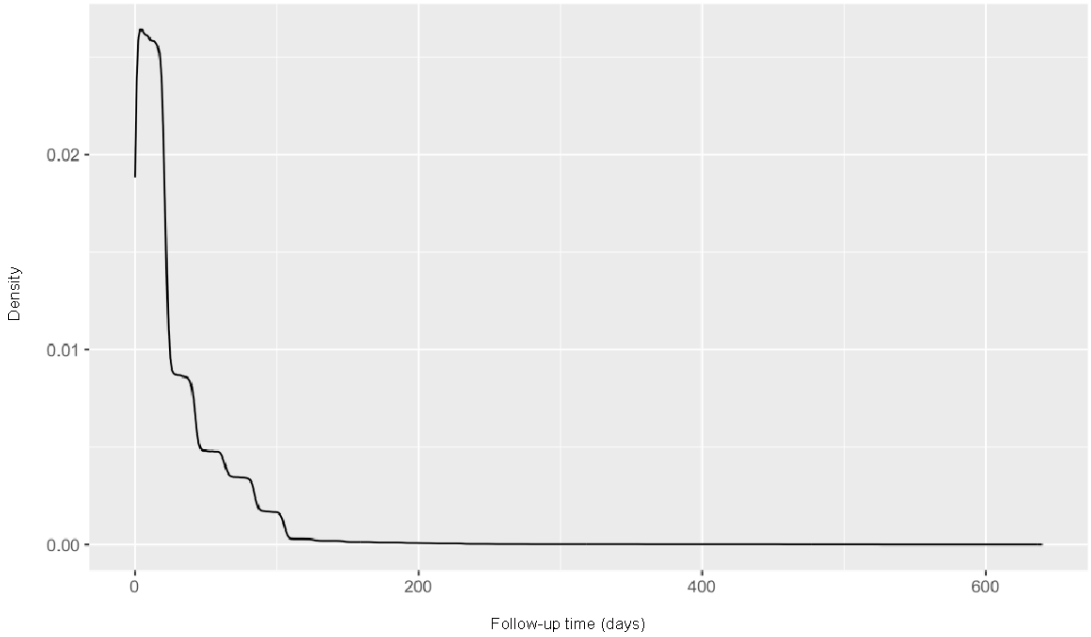
eFigure 3. Censoring Times Histogram and Density Plot in Study Cohort



eFigure 4. Density of Follow-Up Time in the Intention-To-Treat Analyses



eFigure 5. Density of Follow-Up Time in the Per-Protocol Analyses



eReferences

1. 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
2. Bartlett JW, Seaman SR, White IR, Carpenter JR, for the Alzheimer's Disease Neuroimaging Initiative*. 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
3. Moons KGM, Donders RART, Stijnen T, Harrell FE. Using the outcome for imputation of missing predictor values was preferred. *J Clin Epidemiol*. 2006;59(10):1092-1101. doi:10.1016/j.jclinepi.2006.01.009
4. Leyrat C, Carpenter JR, Bailly S, Williamson EJ. Common Methods for Handling Missing Data in Marginal Structural Models: What Works and Why. *Am J Epidemiol*. 2021;190(4):663-672. doi:10.1093/aje/kwaa225
5. 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