Emulation of the MONALEESA-2 (NCT01958021) trial using specialty oncology electronic health records databases

NCT07274709

9th February 2026

1. Title Page

Title	Emulation of the MONALEESA-2 (NCT01958021) trial using specialty oncology electronic health records databases
Research question & Objectives	Emulation of the MONALEESA-2 trial (NCT01958021), which compared ribociclib–letrozole to letrozole alone with progression-free survival (PFS) as the primary endpoint and overall survival (OS) as a secondary endpoint in postmenopausal women with hormone receptor (HR)-positive [estrogen receptor (ER) and/or progesterone receptor (PR)], human epidermal growth factor receptor 2 (HER2)-negative advanced breast cancer.
Protocol version	V2.0
Last update date	February 9, 2026
Contributors	**Primary investigators:** Denys Shay, Shirley V. Wang
Study registration	**Site:** clinicaltrials.gov **Identifier:** NCT07274709
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-scm.partners.org/drugepi/encore/monaleesa2-nct-01958021 (access within Mass General Brigham network only)
Quarto study report (including annotated code and output)	https://gitlab-scm.partners.org/drugepi/encore/monaleesa2-nct-01958021/-/tree/main/public?ref_type=heads (access within Mass General Brigham network only)
***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 packages 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 MONALEESA-2 trial (NCT01958021) using real-world specialty oncology electronic health records data and to investigate the concordance between the trial's original and the emulated treatment effect estimate on overall survival (OS). MONALEESA-2 was a Phase III, randomized, double-blind study evaluating the efficacy and safety of ribociclib (600 mg orally once daily on days 1–21 of a 28-day cycle, followed by 7 days off) in combination with letrozole (2.5 mg orally once daily, continuously) versus placebo plus letrozole in postmenopausal women with hormone receptor–positive [estrogen receptor (ER) and/or progesterone receptor (PR)], human epidermal growth factor receptor 2–negative (HER2−) advanced breast cancer who had not received prior systemic treatment for their advanced disease.

3. Amendments and updates

Version date	Version number	Section of protocol	Amendment or update	Reason
September 07, 2025	V1.0	NA	Initial version	NA
January 25, 2026	V1.1	Section 6.7 Data analysis	Clarified analytic approach, including: (1) exclusion of patients with negative or implausible follow-up times due to death date imputation close to index date; (2) use of cluster-robust standard errors after propensity score matching; (3) clarification that the causal estimand corresponds to the average treatment effect in the control (ATC); and (4) explicit description of combining database-specific estimates using Rubin's rules followed by a fixed-effects meta-analysis	Improve analytic rigor, clarity of estimand interpretation, and transparency of inference procedures
		Section 6.7 / Table 7 (Missing data methods)	Expanded description of the multiple imputation strategy, specifying use of multiple imputation by chained equations with a random forest model and inclusion of auxiliary covariates	Improve reproducibility and transparency of missing data handling
		Sensitivity analyses (Table 8)	Expanded and clarified sensitivity analysis specifications, including prognostic score-based matching and rationale for alternative analytic choices	Strengthen robustness assessment and interpretability of findings

		Appendices – Covariate balance figures	Documented removal of the covariate c_granulocytes_leukocytes_ratio_cont from EDB4 analyses due to excessive missingness relative to post-inclusion/exclusion sample size	Prevent model instability and improve covariate balance assessment in small cohorts
		Appendices – Additional Figures and Tables	Added treatment initiation trends by calendar year for EDB1, EDB3, and EDB4 (Figures 12-14), and finalized expanded tables for laboratory plausibility thresholds, vital sign plausibility thresholds, and state-to-region mapping	Enhance descriptive context, data cleaning transparency, and analytic reproducibility
February 9, 2026	V2.0	6.1.1	Additional reason for the exclusion of EDB4 from the primary analysis was added	Since the study identification period (10/01/2018-09/30/2023) began after the 2017 approval of ribociclib, patients who initiated ribociclib before 10/01/2018 were only included if they survived until the start of the study period, which may introduce immortal time and selection bias due to left truncation.

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 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 **MONALEESA-2 trial**. MONALEESA-2 was a Phase III, double-blind, randomized study assessing the efficacy and safety of **ribociclib (600 mg orally once daily on days 1–21 of a 28-day cycle, followed by 7 days off)** in combination with **letrozole (2.5 mg orally once daily, continuously)** versus **letrozole alone** in **postmenopausal women with hormone receptor–positive [estrogen receptor (ER) and/or progesterone receptor (PR)], human epidermal growth factor receptor 2–negative (HER2–) advanced breast cancer** who had not received prior systemic treatment for their advanced disease.

The **primary trial endpoint** was **progression-free survival (PFS)**, with a hazard ratio (HR) for progression or death of **0.56 (95% CI, 0.43–0.72; P < .001)**. The **median follow-up was not reached** (95% CI, 19.3 months to not estimable) in the ribociclib-plus-letrozole group compared with **14.7 months (95% CI, 13.0–16.5)** in the letrozole-alone group. The FDA granted ribociclib approval in **March 2017**, in combination with letrozole for the treatment of ER-positive, HER2-negative advanced breast cancer as initial endocrine based therapy in postmenopausal women.

Overall survival (OS) was a **secondary endpoint**. At the final OS analysis, a significant overall survival benefit was observed, with a median overall survival of 63.9 months (95% CI, 52.4 to 71.0) in the ribociclib group and 51.4 months (95% CI, 47.2 to 59.7) in the placebo group (**HR for death, 0.76; 95% CI, 0.63 to 0.93**; two-sided P=0.008).

The PFS endpoint was published in the *NEJM* on November 3, 2016 (PMID: 27717303).[5]

Updated results were made available online in Annals of Oncology on January 6, 2020 (PMID: 29718092).[6]

The final OS endpoint was published in the NEJM on March 9, 2022 (PMID 35263519).[7]

5. Research question and objectives

The primary and secondary research question is 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] in patients who initiated ribociclib plus letrozole versus patients who initiated letrozole alone.
Hypothesis:	Initiation of ribociclib plus letrozole improves OS time as compared to initiation of letrozole alone.
Population ***(mention key inclusion-exclusion criteria):***	• Age ≥18 years • Postmenopausal women with ER/PR-positive, HER2-negative locally advanced or metastatic breast cancer who receive treatment without curative intent

	• No prior systemic treatment for advanced/metastatic disease • ECOG 0 or 1
Exposure:	Initiation of ribociclib plus letrozole
Comparator:	Initiation of letrozole
Outcome:	Primary: Time to all-cause mortality/overall survival (OS)
Time *(when follow up begins and ends):*	After exposure assessment window until outcome, death, last observed clinical activity/last sign of the patient being alive, or data cut-off, whichever occurred earliest
Setting:	1L HR/PR-positive, HER2-negative locally advanced or metastatic breast cancer
Main measure of effect:	Hazard ratio (95% CI)

The emulation of the main protocol elements of the MONALEESA-2 is illustrated side by side in **Table 2**.

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

Protocol component	MONALEESA-2 RCT	Emulation	Comments
Eligibility criteria	• Women aged ≥18 years with diagnosis of breast cancer	• Female aged ≥18 years at treatment initiation with a diagnosis of breast cancer	
	• Locoregionally recurrent or metastatic disease not amenable to curative resection/radiotherapy	• Evidence of metastatic or recurrent disease	
	• ER/PR-positive status	• Documentation of ER/PR-positive (or -missing) status	• If ER/PR is missing and a patient received the exposures of interest, then ER/PR is likely to be positive given the alignment with the indication for the exposures of interest
	• HER2-negative status	• Documentation of HER2-negative (or -missing) status	• If HER2 is missing and a patient received the exposures of interest, then HER2 is likely to be negative given the alignment with the indication for the exposures of interest
	• Postmenopausal status	• N/A	• Although not directly captured in RWD, it is likely to be fulfilled given the alignment with the indication for the exposures of interest

	No prior systemic therapy for advanced/metastatic disease	• No systemic anti-cancer therapy[a] following initial record indicating metastatic disease and prior to index date	
	• Measurable disease per RECIST v1.1 or bone-only disease confirmed by imaging	• N/A	• It is reasonable to assume that all patients in RWD had measurable disease if they received treatment
	• WHO/ECOG performance status 0 or 1	• WHO/ECOG performance status of 0 or 1 within 90 days of index date	
	• No prior treatment with CDK4/6 inhibitors	• Exclude patients with prior CDK4/6 inhibitor exposure	
	• No patients with locally advanced (unresectable) or inflammatory breast cancer	• Exclude patients with locally advanced (unresectable) or inflammatory breast cancer	• "Locally advanced" will be defined as Stage 3B or 3C
	• No patients with adjuvant therapy within 12 months prior to index (adjuvant treatment with letrozole will be allowed if it occurred at least 12 months prior to index)	• Exclude patients with adjuvant therapy within 12 months prior to index (adjuvant treatment with letrozole will be allowed if it occurred at least 12 months prior to index)	
	• No other malignancy within 3 years	• No prior record of non-breast cancer malignancy within 3 years	• Prior non-breast cancer malignancy within 3 years is approximated with advanced prior treatments based on recommendation by the data vendor for EDB4[b]
	• N/A	• Restriction to the time period 2017-2023 for identification of initiators	• 2017 was the approval year for ribociclib
	• N/A	• Record of systemic anti-cancer therapy other than the trial treatments[c] during the exposure assessment window based on proprietary business rules	
Treatment strategies	Ribociclib+ letrozole vs. placebo + letrozole	Patients initiating ribociclib + letrozole vs. letrozole alone, using an exposure assessment window to capture combination treatments, based on each data vendor's proprietary business rules	Use treatment start dates to define exposure
Assignment procedures	Randomized 1:1 to ribociclib+ letrozole or placebo + letrozole	Propensity score–based matching or weighting to emulate randomization	Balance baseline covariates to reduce confounding
Follow-up period	Time from randomization to death or censoring	After exposure assessment window until death or censoring	The purpose of an exposure assessment window is a rule-based identification of combination therapies
Outcome	Primary: Progression-free survival (PFS) per investigator assessment Secondary: Overall survival (OS)	Primary: OS	Lack of good measurement of progression, so inferred by initiation of next treatment as a secondary endpoint rather than primary
Causal contrast	Intent-to-treat effect	Effect of initiating ribociclib+ letrozole versus letrozole	Analogous to ITT; emulates initiation rather than adherence

[a]**Includes the following (same as antineoplastic drugs in 6.6.4) :** abemaciclib, alpelisib, anastrozole, atezolizumab, bevacizumab, capecitabine, capivasertib, carboplatin, cisplatin, cyclophosphamide, docetaxel, doxorubicin, elacestrant, entrectinib, epirubicin, eribulin, etoposide, everolimus, exemestane, fluorouracil, fulvestrant, gemcitabine, goserelin, inavolisib, ixabepilone, larotrectinib, letrozole, methotrexate, nab-paclitaxel, olaparib, paclitaxel, palbociclib, pembrolizumab, pertuzumab, ribociclib, sacituzumab, talazoparib, tamoxifen, toremifene, trastuzumab, and vinorelbine.

[b]**Includes the following:** adagrasib, afatinib, alectinib, amivantamab, atezolizumab, bevacizumab, belantamab mafodotin, bendamustine, binimetinib, bortezomib, brigatinib, cabozantinib, capmatinib, carfilzomib, cemiplimab, ceritinib, cetuximab, ciltacabtagene autoleucel, cisplatin, crizotinib, dabrafenib, dacomitinib, daratumumab, datopotomab, dexamethasone, dostarlimab, durvalumab, elotuzumab, elranatamab, encorafenib, ensartinib, entrectinib, erdafitinib, erlotinib, etoposide, fam-trastuzumab deruxtecan, fluorouracil, fruquintinib, gefitinib, gemcitabine, idecabtagene vicleucel, ipilimumab, irinotecan, isatuximab, ixazomib, lapatinib, larotrectinib, lazertinib, lenalidomide, leucovorin, levoleucovorin, linvoseltamab, lorlatinib, melphalan, mobocertinib, nivolumab, osimertinib, oxaliplatin, panitumumab, panobinostat, pembrolizumab, pemetrexed, pomalidomide, pralsetinib, ramucirumab, regorafenib, repotrectinib, retifanlimab, selinexor, selpercatinib, sotorasib, sunvozertinib, taletrectinib, talquetamab, teclistamab, telisotuzumab vedotin, tepotinib, thalidomide, tislelizumab, toripalimab, trametinib, tremelimumab, vandetanib, venetoclax, vemurafenib, vinorelbine, zenocutuzumab, ziv-aflibercept, zongertinib.

[c]**Includes the following:** same as in a, except for ribociclib and letrozole.

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, breast cancer-specific data are available for EDB1, EDB3 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[8,9]).

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 MONALEESA-2 trial, EDB3 was found insufficient to be included in the main analysis for the following reasons.

  - Rationale for excluding of EDB3 from primary analysis: After applying all I/E criteria, EDB3 results in a cohort with the very small sample size even before matching (Figure 6) and imbalanced covariates after matching (Figure 9).
  - Rationale for excluding of EDB4 from primary analysis: After applying all I/E criteria, EDB4 results in a cohort with the very small sample size even before matching (Figure 7) and imbalanced covariates after matching (Figure 10). In addition, because the study identification period (10/01/2018–09/30/2023) began after the 2017 approval of ribociclib, patients who initiated ribociclib before 10/01/2018 were only included if they survived until the start of the study period, which may introduce selection bias due to left truncation.[10]

- For these reasons, only EDB1 will be used for the primary analysis. However, EDB3 and EDB4 will be considered as part of a sensitivity analysis in which all databases are individually analyzed (see sensitivity analysis #11 in Table 8).

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	EDB3	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	Follow-up information through June 2023 (there is no specific time period restrictions for patient eligibility)	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:	Anytime at start of study drug initiation	Anytime at start of study drug initiation	Anytime at start of study drug initiation
Data Version (or date of last update):	Delivery: Jul 11, 2024	Delivery: Jun 16, 2023	Delivery: Oct 24, 2023 Updated (demographics): Feb 29, 2024
Data sampling/extraction criteria:	Patients are sampled if they have a confirmed diagnosis of metastatic breast cancer via abstraction on or after 1 Jan 2011, and at least 2 EHR visits on or after 1 Jan 2011. Both ICD-9 (174.x) and ICD-10 (C50.x) codes are used for the initial selection, and advanced diagnosis are then confirmed via abstraction (since ICD codes do not specify advanced diseases).	EDB3 identifies patients for curation using a structured ICD-10 diagnosis code (ICD-10 C50*), corresponding to the indication of interest, along with at least the year of diagnosis. Once this initial screening list is generated, patients are randomly selected for further review. Curation begins with confirmation of the diagnosis and diagnosis date, primarily based on pathology reports and other unstructured data sources. All patients must be over 18 years of age at the time of their first diagnosis. Certain breast cancer cases are excluded from curation. Specifically, in-situ breast cancers such as ductal carcinoma in situ (DCIS) or lobular carcinoma in situ (LCIS) without an invasive component are not eligible. However, DCIS with microinvasion or DCIS associated with Paget's disease qualifies for curation. Additionally, in-situ breast cancers that later progress to invasive carcinoma are eligible; in these cases, the diagnosis date	Breast cancer 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 breast cancer 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.

		should reflect the date of the original in-situ diagnosis, not the later invasive diagnosis. Patients are ultimately selected for inclusion in data products through quality control processes that assess consistency and potential conflicts in their records. These evaluations may rely on structured data alone or a combination of structured and curated data, depending on the specific data product. Importantly, CAI does not exclude patients based on data completeness, in order to prevent the introduction of selection bias.	
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 derived from structured EHR data, manual curation, and third-party sources including obituary data and the Social Security Death Index. De-identified tokens link patients across datasets using hashed PII. Curated data is prioritized, followed by EHR and then third-party sources. Reported death dates are shifted to the nearest Sunday within four days to enhance privacy. Curated death information follows a source hierarchy: death certificate, obituary, or provider-reported date.	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	Yes	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*

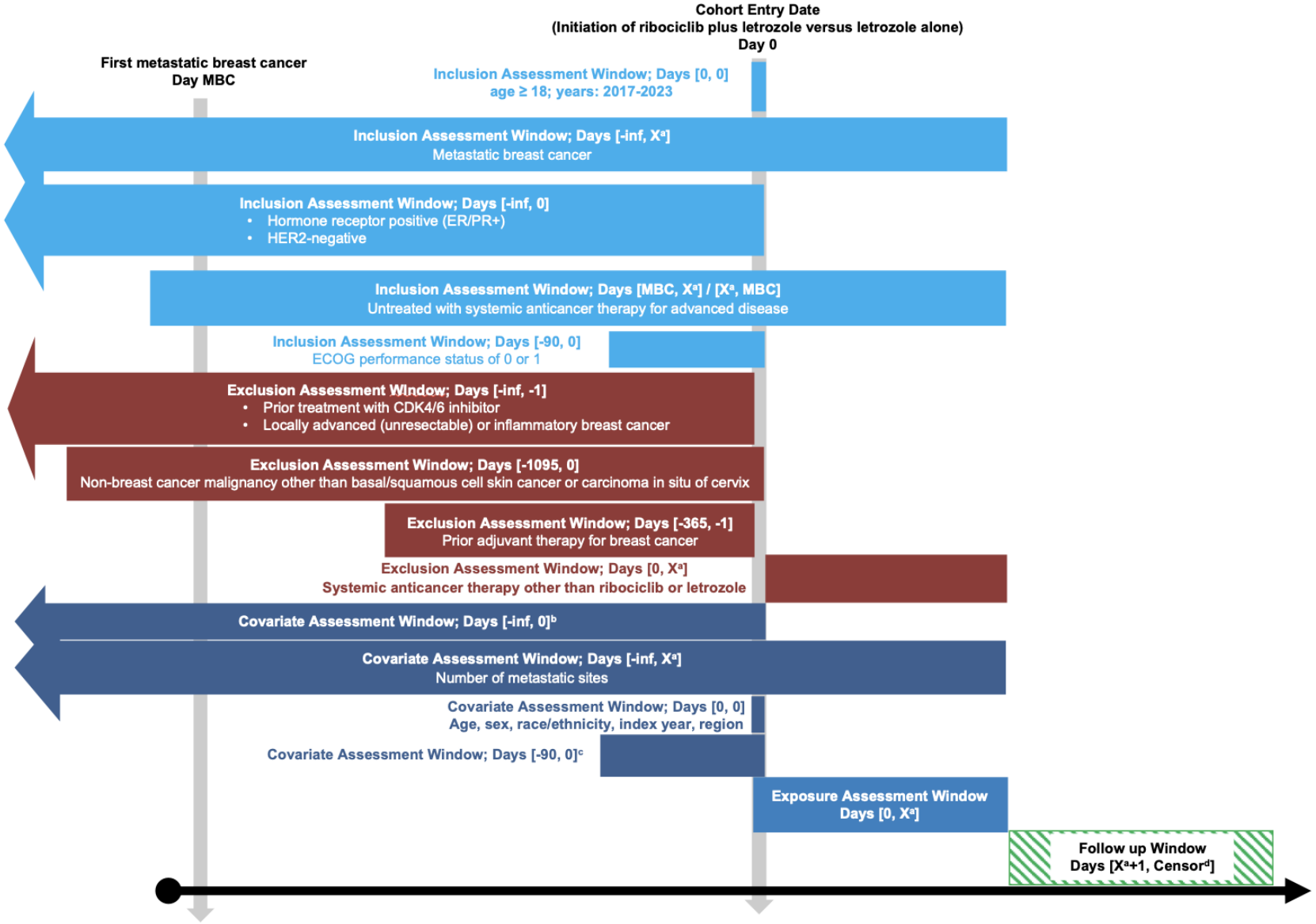
Research design (e.g. cohort, case-control, etc.): Cohort study

Rationale for study design choice: Resembles the principles of the (target) trial emulation framework.[11]

6.5. *Study design diagram*

Figure 1 depicts study design and variable measurement considerations for the emulation of the MONALEESA-2 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.[12-16]

Figure 1. Study design illustration for MONALEESA-2 trial emulation.



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

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

c. Labs (albumin, hemoglobin, etc.) and vitals (BMI, etc.) that are part of the ROPRO prognostic score[12]

d. 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)

No observability criterion was applied because measures like continuous enrollment periods (claims data) are not available in electronic health records.

Abbreviations: MBC = metastatic breast cancer; ER = estrogen receptor; PR = progesterone receptor; HER2 = Factor Epidermal Growth Receptor 2; ECOG = Eastern Cooperative Oncology Group; CNS = Central Nervous System; CDK4/6 = Cyclin-Dependent Kinase 4/6; ROPRO = Real-wOrld PROgnostic score

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 ribociclib plus letrozole (exposure) or letrozole alone (comparator) as part of their 1L systemic antineoplastic treatment for advanced/metastatic adenocarcinoma of the breast. 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 report 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. In addition, the eligible time period (years) was based on the approval year of the trial's intervention drug as well as the numbers of treatment initiation. A flowchart of the study cohort assembly is provided in the appendix (10.1).

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 the appendix (10.1).

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

The exposure and comparator were defined to emulate the agents compared for the trial, i.e., initiation of ribociclib plus letrozole versus letrozole alone in a 1L metastatic 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 therapy, based on a proprietary business rule with an exposure assessment window. 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 are identified who received ribociclib-plus-letrozole or letrozole-alone treatment regimen by their generic names (string match) in 1L, using the

exposure assessment window; follow-up begins after the end of the exposure assessment window. 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 EDB1* Quarto report (access within MGB network only).

- **EDB3**: Exposure is derived using a manually curated LOT table provided by the data partner that programmatically categorizes treatment regimens into a coherent line of therapy, based on a proprietary business rule with an exposure assessment window. That is, each patient is represented with one row per curated line of therapy **and** drug name with corresponding information online number, regimens as well as start and end dates. Based on this table, patients are identified who received only ribociclib-and-letrozole within the first line of therapy, and letrozole alone within the first line of therapy by their generic names (string match), using the exposure assessment window; follow-up begins after the end of the exposure assessment window. The LOT implicitly only considers regimens that were given as part of a metastatic disease setting.
- **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 breast cancer (see list below*). Only these are considered.
    - Identify patients who received any of the MONALEESA-2 treatments within the exposure assessment window as the first antineoplastic treatment on or after the metastasis date.
    - Identify and exclude patients who received ribociclib before the metastasis date.
    - Follow-up begins after the end of the exposure assessment window.

***Antineoplastic drugs considered:** abemaciclib, alpelisib, anastrozole, atezolizumab, bevacizumab, capecitabine, capivasertib, carboplatin, cisplatin, cyclophosphamide, docetaxel, doxorubicin, elacestrant, entrectinib, epirubicin, eribulin, etoposide, everolimus, exemestane, fluorouracil, fulvestrant, gemcitabine, goserelin, inavolisib, ixabepilone, larotrectinib, letrozole, methotrexate, nab-paclitaxel, olaparib, paclitaxel, palbociclib, pembrolizumab, pertuzumab, ribociclib, sacituzumab, talazoparib, tamoxifen, toremifene, trastuzumab, and vinorelbine.

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

The primary outcome for the database study was defined to emulate the OS outcome for the trial, time from end of exposure assessment window to death due to any reason (OS). Operational definitions:

- **EDB1**: Time in [days, months and years] from end of exposure assessment window 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. In 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 OS endpoint is operationalized using a parameterized R function edb1_get_os() and more details can be found in the attached pdf documentation.

- EDB3: Time in [days, months and years] from end of exposure assessment 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. Patients without evidence of death are censored at the earlier of the last recorded activity date or the data cut-off date. Activity dates are defined as documented in Table 4. Death dates are compiled from Electronic Medical Records (EMR), manual curation, and third-party death data, linked via de-identified tokens generated from hashed personally identifiable information (PII). Sources of death dates are prioritized as follows: (1) manual curation, (2) EMR, and (3) third-party sources. For privacy reasons, the final reported death date is shifted to the nearest Sunday within four days of the actual date. Manually curated death dates follow a prioritization hierarchy of source documentation: death certificate first, followed by scanned obituaries or death announcements, and lastly exact dates reported by providers without other supporting documentation. The OS endpoint calculation uses a parameterized R function (*edb3_get_os()*), with detailed documentation provided in the attached PDF.
- EDB4: Time in [days, months and years] from end of exposure assessment 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 (as per data vendor recommendation). The OS endpoint is operationalized using a parameterized R function *edb4_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

6.6.6. Context and rationale for follow up

Only intention-to-treat (ITT) analyses will be conducted. Although cross-over from the exposure to the comparator can be expected which biases the exposure treatment effect more towards the null, this also applies to the RCT. In the final overall survival (OS) analysis reported by Hortobagyi et al., 304 patients (87.8%) in the ribociclib-plus-letrozole arm and 317 patients (90.2%) in the placebo-plus-letrozole arm had discontinued study treatment. After discontinuation, 66 patients (21.7%) in the ribociclib arm and 109 patients (34.4%) in the placebo arm received a CDK4/6 inhibitor as part of subsequent therapy, which effectively constituted crossover.[7]

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.[12] 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 9** and **Table 10**.

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 index date	dem_age_index	edbx_get_demographics()	Age measured at index date	Binary (<65, 65+); modelled continuously in ROPRO[12]	[0;0]
Year of index date	c_year_index	De novo derived from dt_index	Calendar year in which patient initiated study treatment	Nominal (<2020, 2020+)	[0;0]
Family history	dem_family_history	edbx_get_demographics()	Family history of cancer. Not available in EDB1.	Logical (TRUE, FALSE)	[0;0] (no specific date is associated)
Race	dem_race	edbx_get_demographics()	Race categorized as in the original RCT	Nominal (White, Asian, Black, Others)	[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 11**)	Nominal (Northeast, South, West, Midwest)	[0;0]
Practice type	dem_practice	edbx_get_demographics()	Setting patient is receiving care at. Not available in EDB3 and EDB4.	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. Not available in EDB3 and EDB4.	Nominal (from '1 - Lowest SES' through '5 - Highest SES)	[-inf;0]
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. In EDB1, "smoking history" is unavailable for breast cancer.	Binary logical (TRUE, FALSE)	[-inf;0]
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[12]; due to I/E criteria ECOG is modelled as a binary (0, 1) covariate	[-90;0]

Characteristic	Harmonized analysis variable name	R function to derive covariate (see pdf in appendix)	Details[1]	Variable encoding	Assessment window
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)	[-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
Number of metastatic sites	c_number_met_sites	edbx_get_diagnosis_solid() / edbx_get_number_met_sites_solid()	Number of metastatic sites for a given patient before/on index date	Integer	[-inf;X]; Proprietary business-rule-based covariate assessment window
Time between initial diagnosis to index date	c_time_dx_to_index_quartiles	edbx_get_diagnosis_solid()	Time in days between initial diagnosis to index date. Quartiles will be used because there may be "negative" times intervals due to date imprecision of the lines of therapy.	Categorical (quartiles)	[-initial dx;0]
Time between earliest evidence of a metastatic and index date	c_time_met_dx_to_index_quartiles	edbx_get_diagnosis_solid()	Time in days between earliest evidence of a metastatic diagnosis and index date. Quartiles will be used because there may be "negative" times intervals due to date imprecision of the lines of therapy.	Categorical (quartiles)	[met dx;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 index date in U/L. In case of ties, the lower is selected. Only used to compute AST-ALT ratio.	Continuous	[-90;0]

Characteristic	Harmonized analysis variable name	R function to derive covariate (see pdf in appendix)	Details[1]	Variable encoding	Assessment window
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]
Estrogen receptor status	c_er_status	edbx_get_biomarker()	Evidence of estrogen receptor status.	binary (positive/ negative)	[-inf; 0]
Glucose[2]	c_glucose_mg_dl_cont	edbx_get_labs()	Closest glucose measurement (in serum/plasma) relative to index date in mg/dL. In case of ties, the lower is selected	Continuous	[-90;0]
Granulocyte/leukocyte ratio [2]	c_granulocytes_leukocytes_ratio	edbx_get_labs()	Closest granulocyte/leukocyte ratio measurement (in blood) relative to index date. In case of ties, the lower is selected. Used to compute granulocyte/lymphocyte ratio	Continuous	[-90;0]
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. Only	Continuous	[-90;0]

Characteristic	Harmonized analysis variable name	R function to derive covariate (see pdf in appendix)	Details[1]	Variable encoding	Assessment window
			used to compute neutrophil/lymphocyte ratio.		
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]
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. Only 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]
Progesterone receptor status	c_pr_status	edbx_get_biomarker()	Evidence of <u>any</u> PR mutation present. If patient has multiple measurements, any evidence of a mutation is prioritized. In case of ties, the closest measurement relative to index date is selected	Binary (positive/ negative)	[-inf;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/dL. 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]
Diastolic blood pressure	c_dbp_cont	edbx_get_vitals()	Closest diastolic 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.	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 oxygen saturation measurement (in bpm) relative to index date. In case of ties, the lower is selected	Continuous	[-90;0]

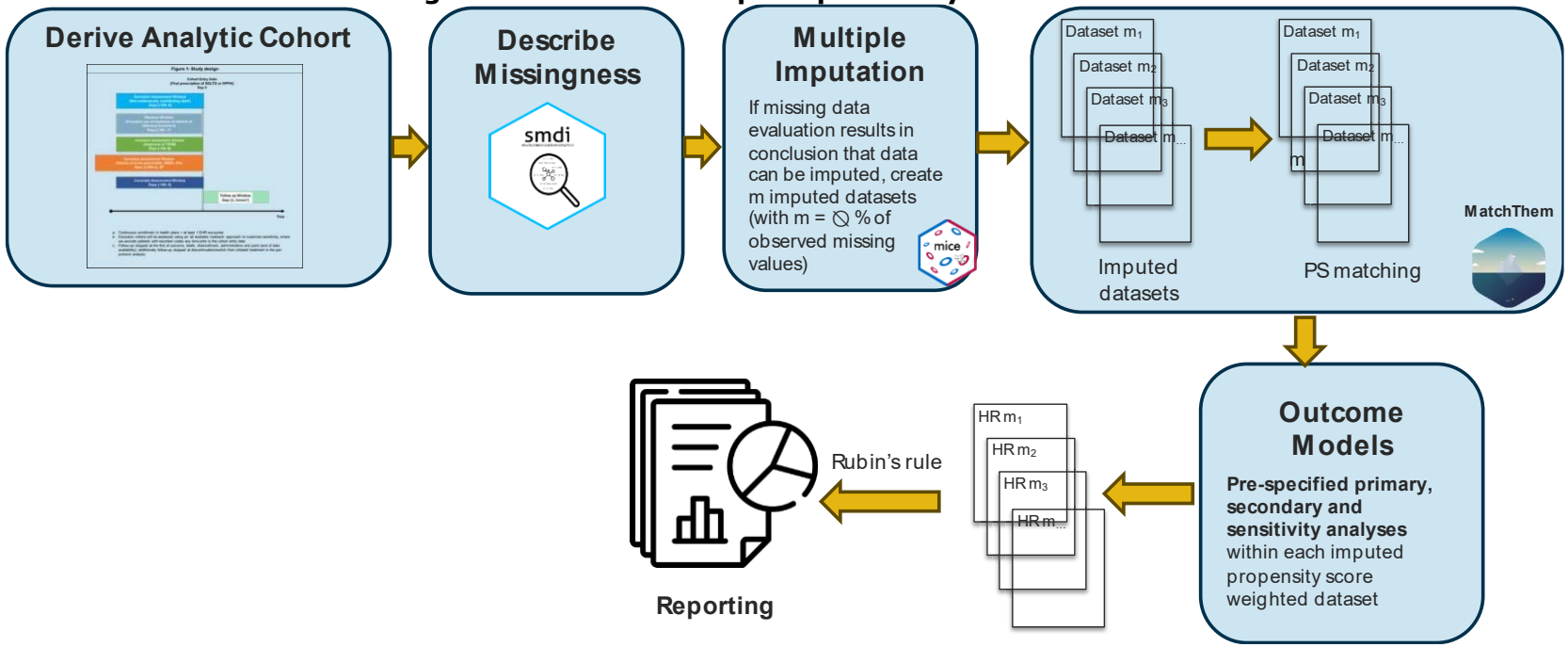
[1] x stands for the pseudonymized number of the respective database, i.e., EDB1, EDB3 or EDB4
[2] For calculation of ROPRO prognostic score[12], this variable is log transformed.
[3] For calculation of ROPRO prognostic score[12], this variable is log-log transformed.

6.7. Data analysis

6.7.1. Context and rationale for analysis plan

To emulate the MONALEESA-2 trial, the following analytical workflow will be used (Figure 2). 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).[17] The analytical cohort will be derived by first identifying a metastatic breast cancer inception cohort of initiators of ribociclib plus letrozole or letrozole alone 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]).[18] 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).[19,20] 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, Hotelling's[21] and Little's[22] 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[23] 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 Hotelling[21] Little[22]	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). >0.1[a]: imbalances in observed patient characteristics; missingness may be likely at random (~MAR).	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 covariates and missingness (~MAR).	No association in either univariate or adjusted model and no meaningful difference in the log HR after full adjustment (~MCAR). Association in univariate but not fully adjusted model (~MAR). Meaningful difference in the log HR also after full adjustment (~MNAR).

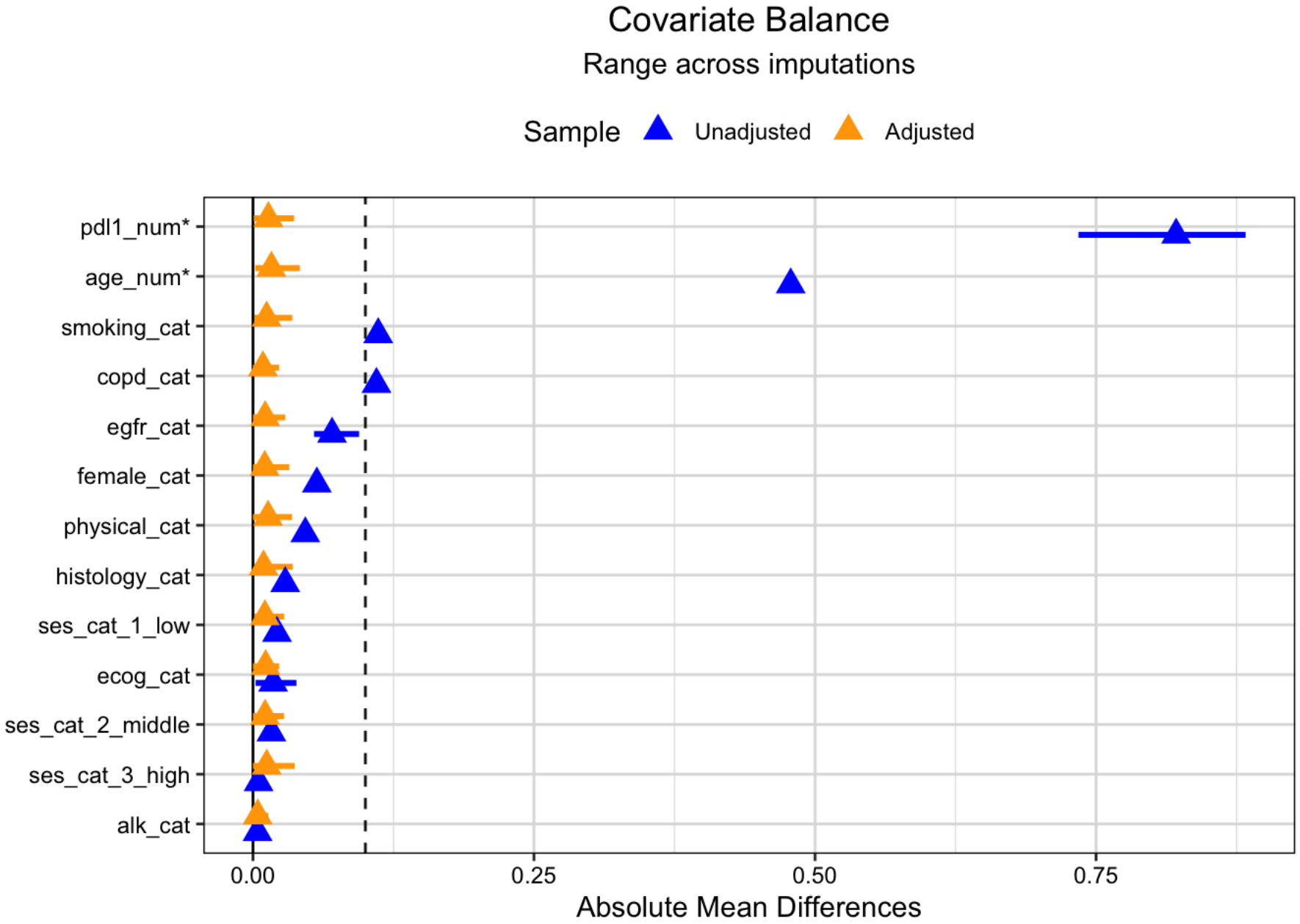
Abbreviations: ASMD = Median absolute standardized mean difference across all covariates, AUC = Area under the curve, CI = Confidence interval, MAR = Missing at random mechanism in which the missingness probability depends on observed covariates, MCAR = Missing completely at random mechanism in which 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.[24]

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).[25,26] 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.[27,28] Hence, multiple imputation with flexible, non-parametric random forest imputation algorithms[29] (mice R package[30]) 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 ribociclib plus letrozole using propensity score matching across imputed datasets we will apply the "within" approach using the "MatchThem" R package.[31,32] 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 ribociclib plus letrozole 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 3. 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 ribociclib-plus-letrozole 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.[33] As a secondary endpoint, we will additionally estimate the median OS survival time difference between the two exposure 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.[30,34] To arrive at a single estimate across databases, the final estimates will be summarized through a meta-analytic fixed effects model.[35] 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. Analyses

Table 7. Primary, secondary, and subgroup analysis specification

Hypothesis:	Initiation of ribociclib plus letrozole decreases the hazard of all-cause mortality as compared to initiation of letrozole alone
Exposure contrast:	Initiation of ribociclib plus letrozole vs letrozole alone
Primary outcome:	Time to all-cause mortality/ overall survival (OS)
Databases used:	EDB1 (primary analysis), and EDB3 and EDB4 (sensitivity analysis)
Time period:	From 2017 (approval year for ribociclib) - 2023
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 25, 2024.
Model(s): *(provide details or code)*	See example code here. The annotated code for the trial emulation will be hosted at https://drugepi.gitlab-pages.partners.org/encore/monaleesa2-nct-01958021 / (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, multiple imputation, 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, ≥65) 2. Race (Asian vs. Non-Asian) 3. ECOG (0, 1) 4. Hormone receptor status (ER and PR-positive vs. Other)

Table 8. 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	Average treatment effect in the overlap population (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 asynchronously. To account for the potential lag of updated mortality information (*ghost-time bias*[36]), the censoring 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.[37]	Approach implements a more conservative censoring rule	Approach addresses ghost-time bias by censoring patients without a recorded death event earlier
Sensitivity #5	Delta imputation models for MNAR (tipping point analysis)	Primary multiple imputation analysis assumes MAR which	Estimates impact of deviations from MAR assumption on final	Delta parameters must be assumed and results are complex to interpret in

		may not hold for every covariate	treatment effect estimates for key covariates	multivariate missingness settings; just most 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 MONALEESA-2	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 MONALEESA-2 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	Using all available calendar time	This analysis explores potential confounding related to calendar time, evolving clinical practice patterns, and access to therapies.	Aims to assess the extent of confounding introduced by changes in time, treatment practices, and access to therapies.	We expect the results to be more confounded due to calendar time.
Sensitivity #9	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 #10	Exclusion of patients with >1 year between metastasis diagnosis and index date	This sensitivity analysis varies the inclusion criteria by excluding patients whose index date (treatment start) is >1 year after their metastatic diagnosis date which is	Helps ensure that treatment reflects initial first-line treatment of metastatic disease, not late-line therapy.	Exclusion of further patients results in a smaller cohort size which may reduce statistical power.

		clinically unrealistic or implausible for a first-line treatment		
Sensitivity #11	Addition of EDB3 and EDB4	EDB3 and EDB4 are not considered in primary analyses for reasons given in section 6.1.1.	Increased sample size and potentially broader coverage of general US cancer population	See limitations listed in section 6.1.1 regarding potential immortal time and selection bias due to left truncation as well as small cohort size leaving residual imbalance on prognostic factors.
Sensitivity #12	Descriptive analysis of switching/crossover patterns in the trial and its emulation	This sensitivity analysis assesses crossover patterns comparatively in the trial and its emulation	Help to understand the potential concordance/discordance of treatment effects between the trial and its emulation	Analysis is on the description level, which may limit conclusions

7. Limitation 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 enrollment periods in administrative claims data) do not exist which may lead to measurement error in key covariates and exposure misclassification. Sensitivity analysis #7 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.

8. Protection of human subjects

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

9. References

1. Franklin JM, Pawar A, Martin D, et al. Nonrandomized Real-World Evidence to Support Regulatory Decision Making: Process for a Randomized Trial Replication Project. *Clin Pharmacol Ther*. Apr 2020;107(4):817–826. 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*. Mar 9 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*. Apr 2020;107(4):735–737. doi:10.1002/cpt.1793
4. Wang SV, Schneeweiss S, Initiative R-D, et al. Emulation of Randomized Clinical Trials With Nonrandomized Database Analyses: Results of 32 Clinical Trials. *JAMA*. Apr 25 2023;329(16):1376–1385. doi:10.1001/jama.2023.4221
5. Hortobagyi GN, Stemmer SM, Burris HA, et al. Ribociclib as First-Line Therapy for HR-Positive, Advanced Breast Cancer. *N Engl J Med*. Nov 3 2016;375(18):1738–1748. doi:10.1056/NEJMoa1609709
6. Hortobagyi GN, Stemmer SM, Burris HA, et al. Updated results from MONALEESA-2, a phase III trial of first-line ribociclib plus letrozole versus placebo plus letrozole in hormone receptor-positive, HER2-negative advanced breast cancer. *Ann Oncol*. Nov 1 2019;30(11):1842. doi:10.1093/annonc/mdz215
7. Hortobagyi GN, Stemmer SM, Burris HA, et al. Overall Survival with Ribociclib plus Letrozole in Advanced Breast Cancer. *N Engl J Med*. Mar 10 2022;386(10):942–950. doi:10.1056/NEJMoa2114663
8. Curtis MD, Griffith SD, Tucker M, et al. Development and Validation of a High-Quality Composite Real-World Mortality Endpoint. *Health Serv Res*. Dec 2018;53(6):4460–4476. doi:10.1111/1475-6773.12872
9. Dong S, Kansagra A, Kaur G, et al. Validation of a Composite Real-World Mortality Variable Among Patients With Hematologic Malignancies Treated in the United States. *JCO Clin Cancer Inform*. Apr 2025;9:e2400233. doi:10.1200/CCI-24-00233
10. Applebaum KM, Malloy EJ, Eisen EA. Left truncation, susceptibility, and bias in occupational cohort studies. *Epidemiology*. Jul 2011;22(4):599–606. doi:10.1097/EDE.0b013e31821d0879
11. Hernan MA, Wang W, Leaf DE. Target Trial Emulation: A Framework for Causal Inference From Observational Data. *JAMA*. Dec 27 2022;328(24):2446–2447. doi:10.1001/jama.2022.21383
12. 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*. Nov 2020;31(11):1561–1568. doi:10.1016/j.annonc.2020.07.013
13. 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. *Medical Research Archives*. 2023;11(4)
14. 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*. Jun 1 2009;27(16):2692–6. doi:10.1200/JCO.2008.19.5081
15. 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
16. Loureiro H, Roller A, Schneider M, Talavera-Lopez C, Becker T, Bauer-Mehren A. Matching by OS Prognostic Score to Construct External Controls in Lung Cancer Clinical Trials. *Clin Pharmacol Ther*. Feb 2024;115(2):333–341. doi:10.1002/cpt.3109
17. 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*. Jan 2024;33(1):e5740. doi:10.1002/pds.5740
18. Rubin DB. Inference and missing data. *Biometrika*. 1976;63(3):581–592.
19. 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
20. 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 Pharmacometrics Syst Pharmacol*. Sep 2023;12(9):1201–1212. doi:10.1002/psp4.12998
21. Hotelling H. The generalization of Student's ratio. 1931;
22. Little RJ. A test of missing completely at random for multivariate data with missing values. *Journal of the American statistical Association*. 1988;83(404):1198–1202.

23. 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*. Apr 2024;7(1):ooae008. doi:10.1093/jamiaopen/ooae008
24. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behav Res*. May 2011;46(3):399–424. doi:10.1080/00273171.2011.568786
25. 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*. Dec 1 2018;187(12):2705–2715. doi:10.1093/aje/kwy173
26. Moreno-Betancur M, Lee KJ, Leacy FP, Simpson JA, Carlin JB. Correction to: "Canonical causal diagrams to guide the treatment of missing data in epidemiologic studies". *Am J Epidemiol*. Mar 4 2025;194(3):877–880. doi:10.1093/aje/kwae406
27. Bartlett JW, Seaman SR, White IR, Carpenter JR, Alzheimer's Disease Neuroimaging I. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Stat Methods Med Res*. Aug 2015;24(4):462–87. doi:10.1177/0962280214521348
28. Weberpals J, Shaw PA, Lin KJ, et al. High-dimensional multiple imputation (HDMI) for partially observed confounders including natural language processing-derived auxiliary covariates. *Am J Epidemiol*. Jan 22 2025;doi:10.1093/aje/kwaf017
29. 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*. Mar 15 2014;179(6):764–74. doi:10.1093/aje/kwt312
30. Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. *Journal of statistical software*. 2011;45:1–67.
31. 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*. Jan 2019;28(1):3–19. doi:10.1177/0962280217713032
32. Pishgar F, Greifer N, Leyrat C, Stuart E. MatchThem:: matching and weighting after multiple imputation. *arXiv preprint arXiv:200911772*. 2020;
33. 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*. Mar 30 2014;33(7):1242–58. doi:10.1002/sim.5984
34. Rubin DB. Multiple imputation. *Flexible imputation of missing data, second edition*. Chapman and Hall/CRC; 2018:29–62.
35. Schwarzer G CJ, Rücker G. *Meta-Analysis with R*. vol 4784. Springer; 2015.
36. Jacobs EJ, Newton CC, Wang Y, Campbell PT, Flanders WD, Gapstur SM. Ghost-time bias from imperfect mortality ascertainment in aging cohorts. *Ann Epidemiol*. Oct 2018;28(10):691–696 e3. doi:10.1016/j.annepidem.2018.06.002
37. Chen L, Fajardo O, Huntley M, Meyer A-M, Taylor M. Use of last clinical activity date in overall survival analysis with real world data. WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA; 2021:116–116.
38. Schoenfeld DA. Sample-size formula for the proportional-hazards regression model. *Biometrics*. 1983:499–503.

10. Appendices

10.1. CONSORT diagrams

The following CONSORT attrition diagrams depict the process to select eligible MONALEESA-2-like populations in EDB1, EDB3 and EDB4 for the main analysis, respectively.

Figure 4. CONSORT attrition to select eligible MONALEESA-2-like populations in EDB1.

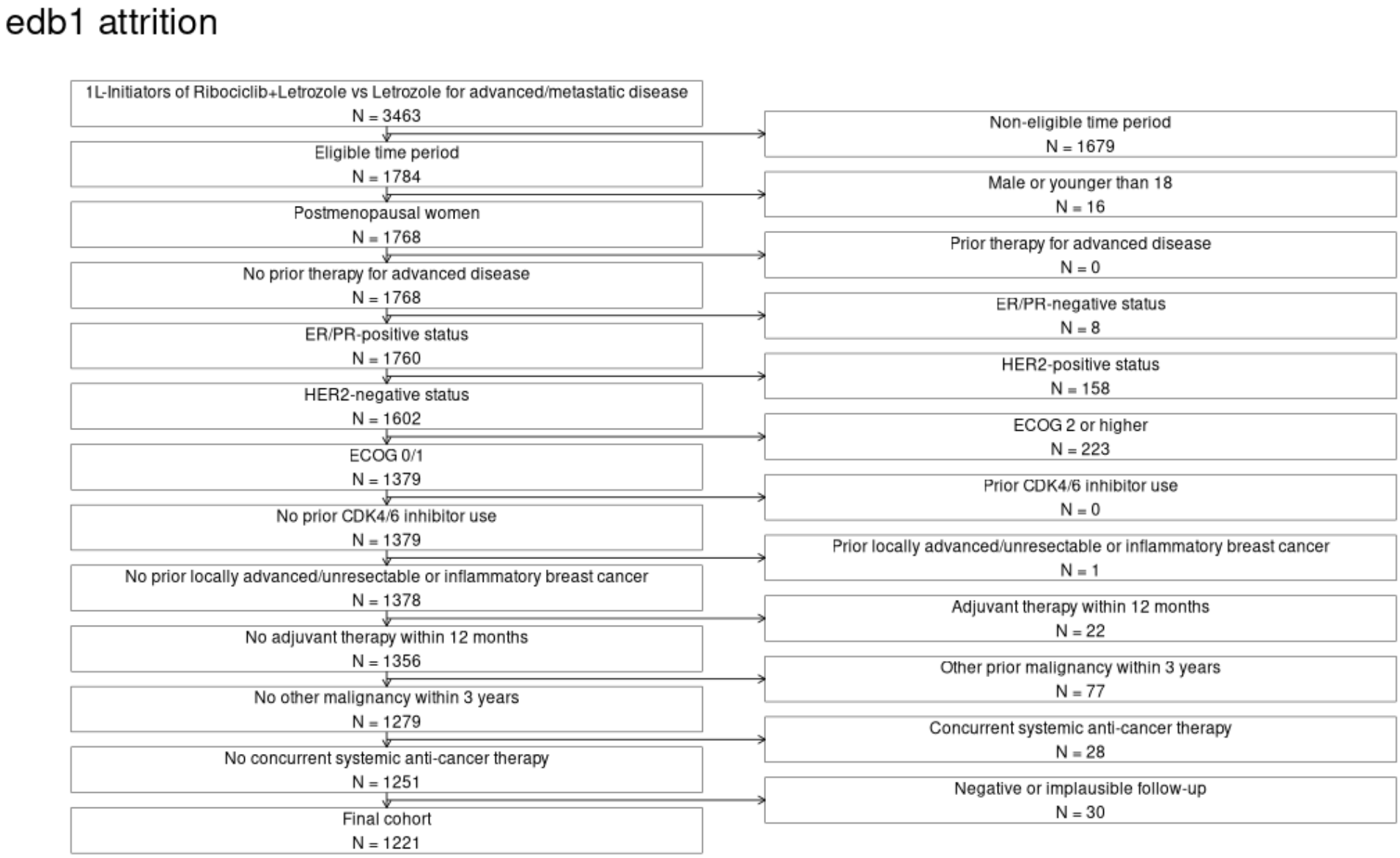


Figure 5. CONSORT attrition to select eligible MONALEESA-2-like populations in EDB3.

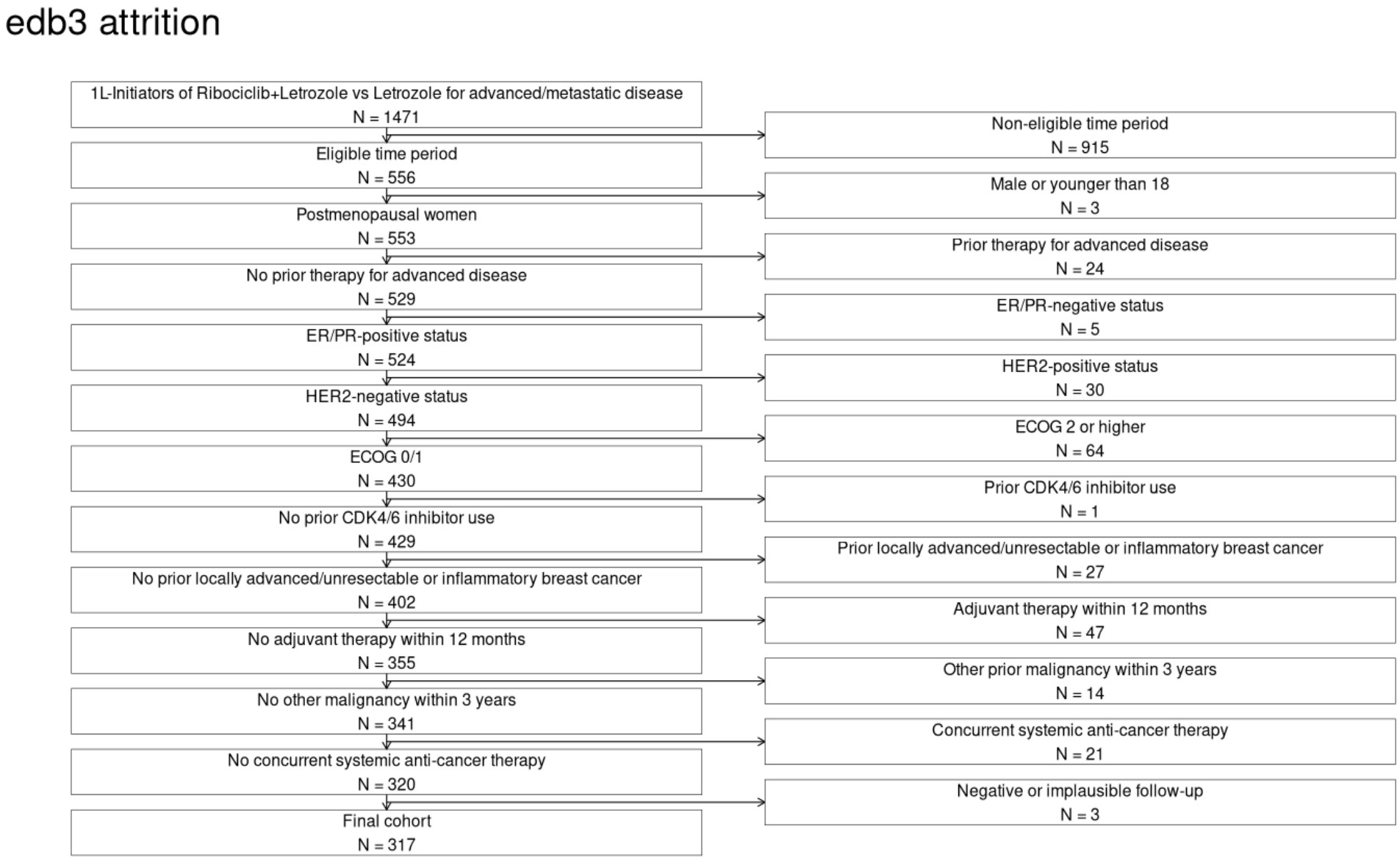
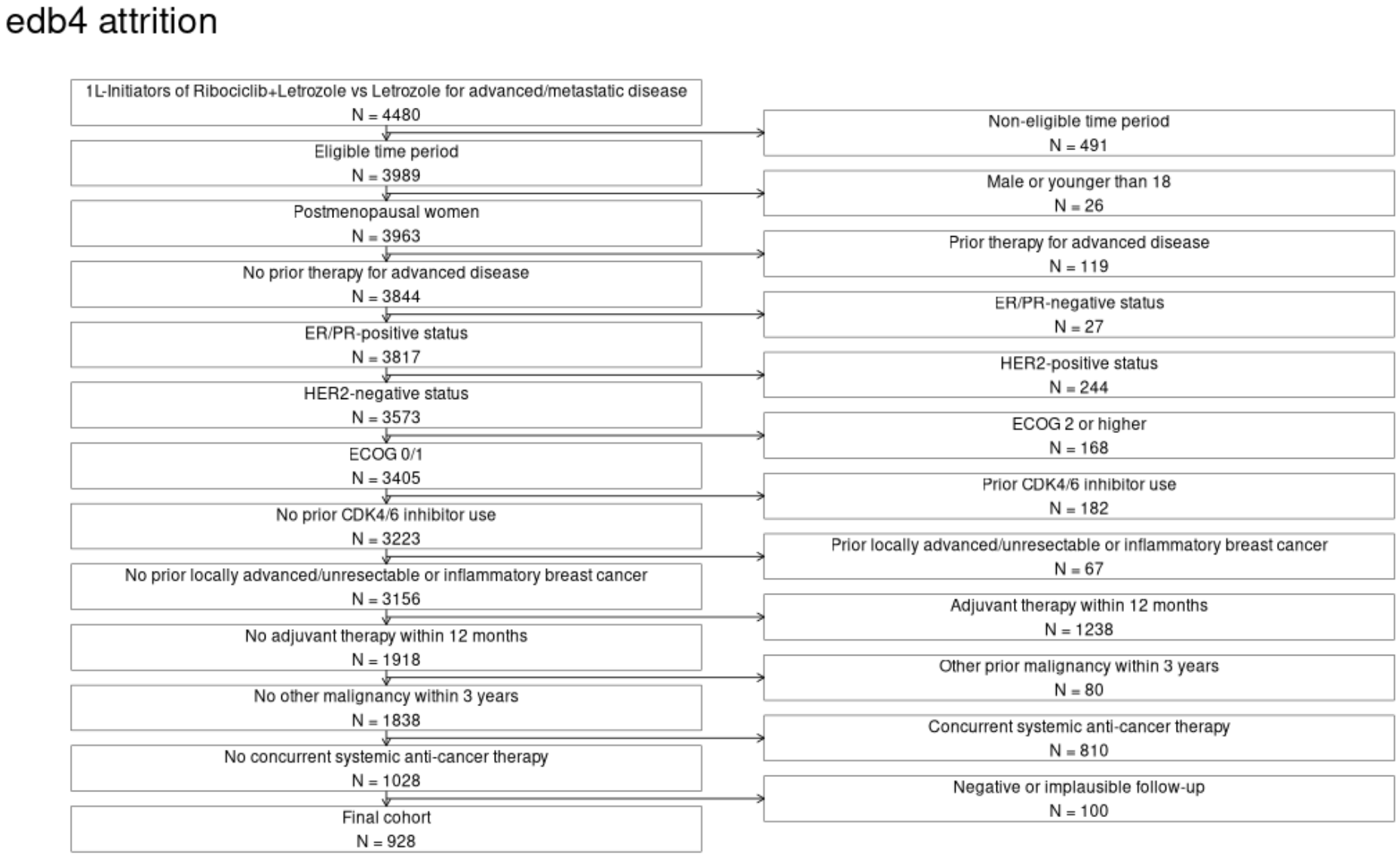


Figure 6. CONSORT attrition to select eligible MONALEESA-2-like populations in EDB4.



10.2. Covariate balance figures

The following figures illustrate the balance of key covariates included in propensity score models among eligible MONALEESA-2-like populations in EDB1, EDB3 and EDB4, respectively.

Figure 7. EDB1 covariate balance of covariates included in propensity score model before and after matching.

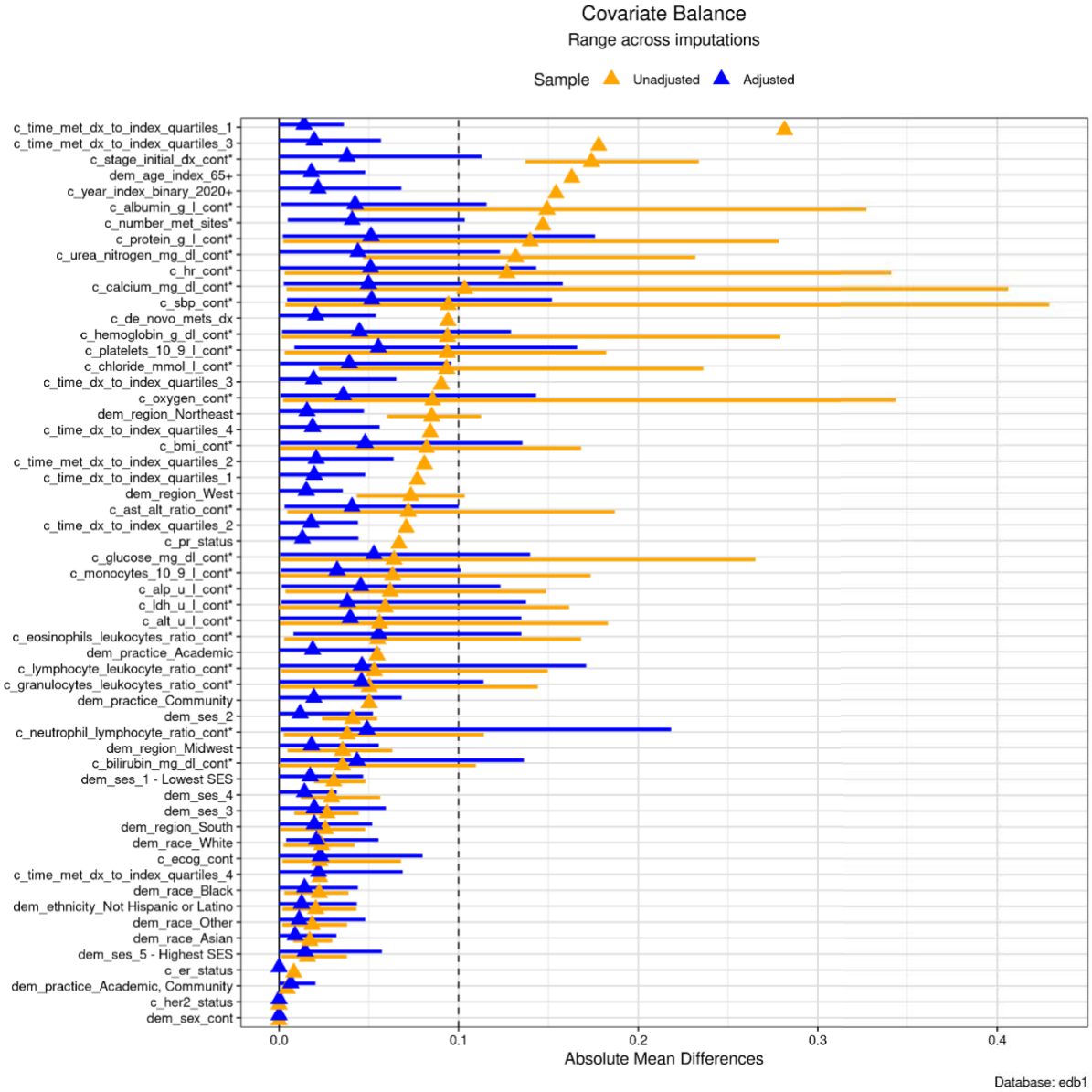


Figure 8. EDB3 covariate balance of covariates included in propensity score model before and after matching.

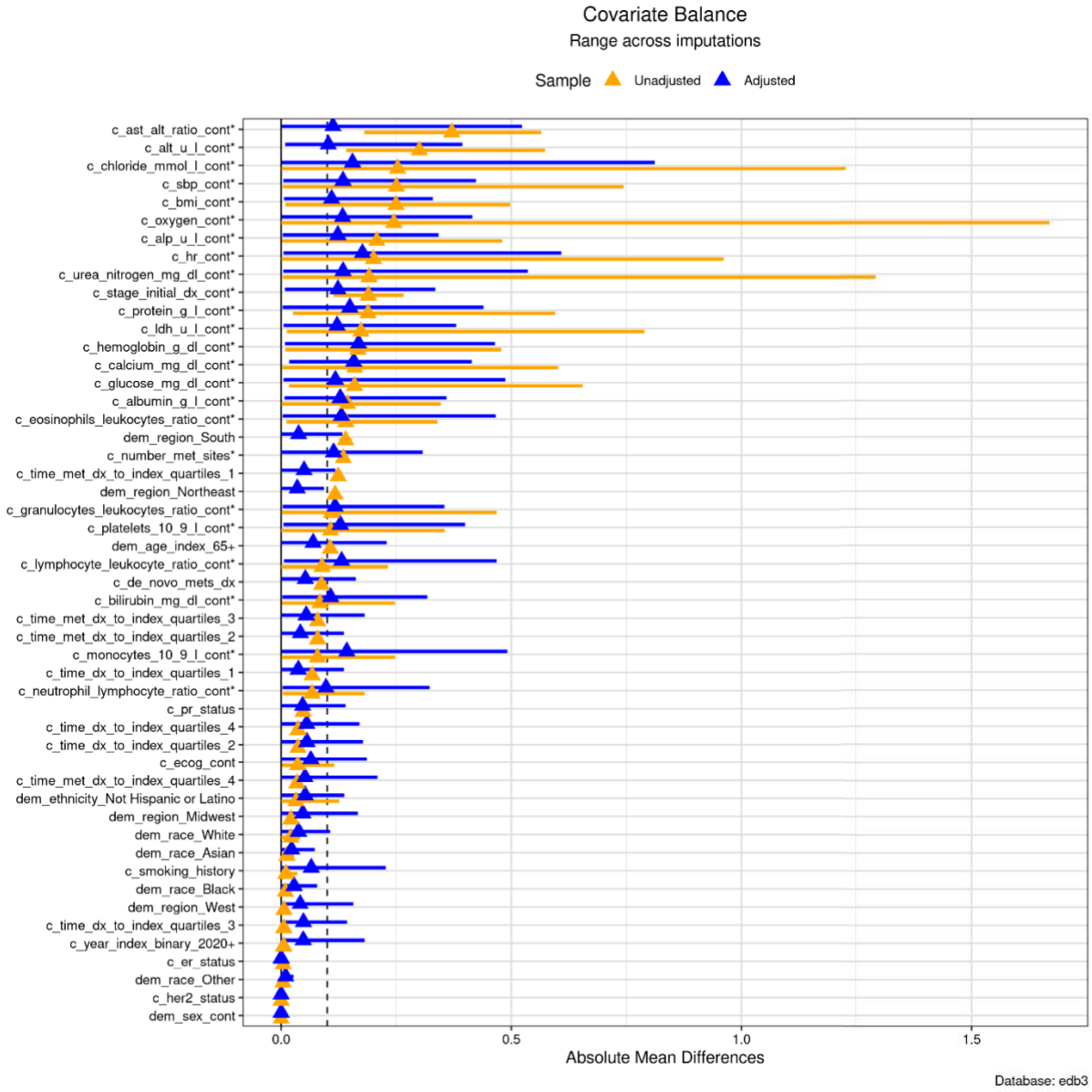
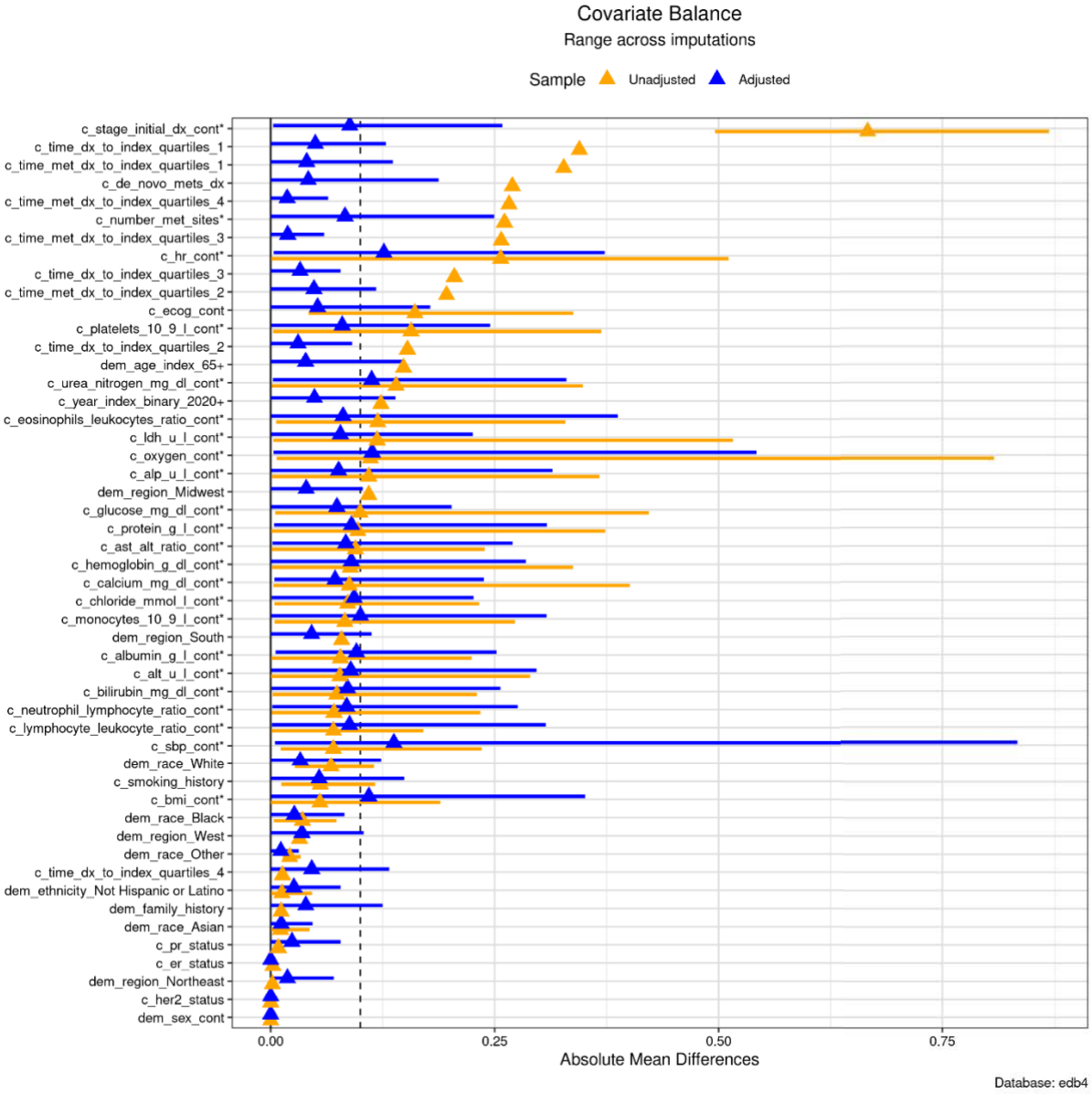


Figure 9. EDB4 covariate balance of covariates included in propensity score model before and after matching.

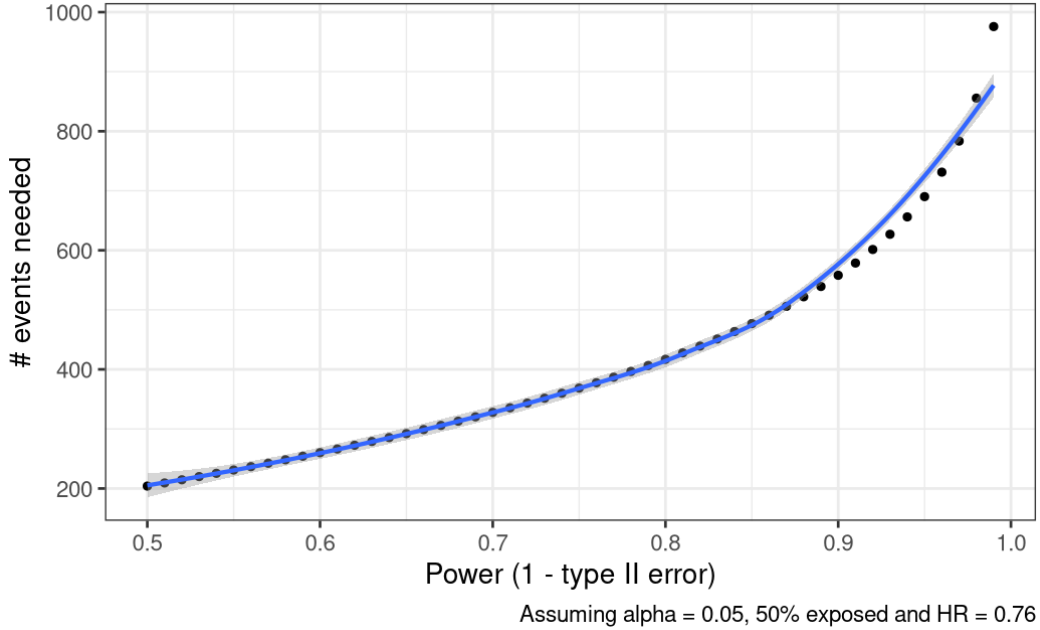


In EDB4, following variable was removed due to too many missings relative to number of covariates and study size after applying the inclusion/ exclusion criteria: c_granulocytes_leukocytes_ratio_cont

10.3. Sample size/power calculations

Power estimations are computed based on the average number of observed events across imputed and matched datasets according to the methodology described by Schoenfeld.[38] The trial showed a modest but statistically significant OS benefit with HR = 0.76 (95% CI 0.63–0.93). Because our 1:1-nearest-neighbor--matched samples are smaller than the number of trial participants (668 in the trial vs 500 in EDB1 [primary analysis], and 144 in EDB4 and 94 in EDB3 [both sensitivity analysis]), we expect our confidence intervals to be wider.

Figure 10 Number of events needed to achieve X power for overall-survival outcome



Primary analysis:

- Based on a total of 191 events in the 1:1-matched set in EDB1, we would achieve a power of 0.47 to detect an effect as large or larger than HR = 0.76.

Thus, our study will be underpowered for the OS outcome, akin to the original trial that was not powered for OS.

Sensitivity analysis:

- Based on a total of 50 events in the 1:1-matched set in EDB3, we would achieve a power of 0.16 to detect an effect as large or larger than HR = 0.76.
- Based on a total of 36 events in the 1:1-matched set in EDB4, we would achieve a power of 0.13 to detect an effect as large or larger than HR = 0.76.

10.4. Additional Figures and Tables

Table 9. 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 10. 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 11. 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 11. Treatment initiation trends by calendar year and treatment in EDB1.

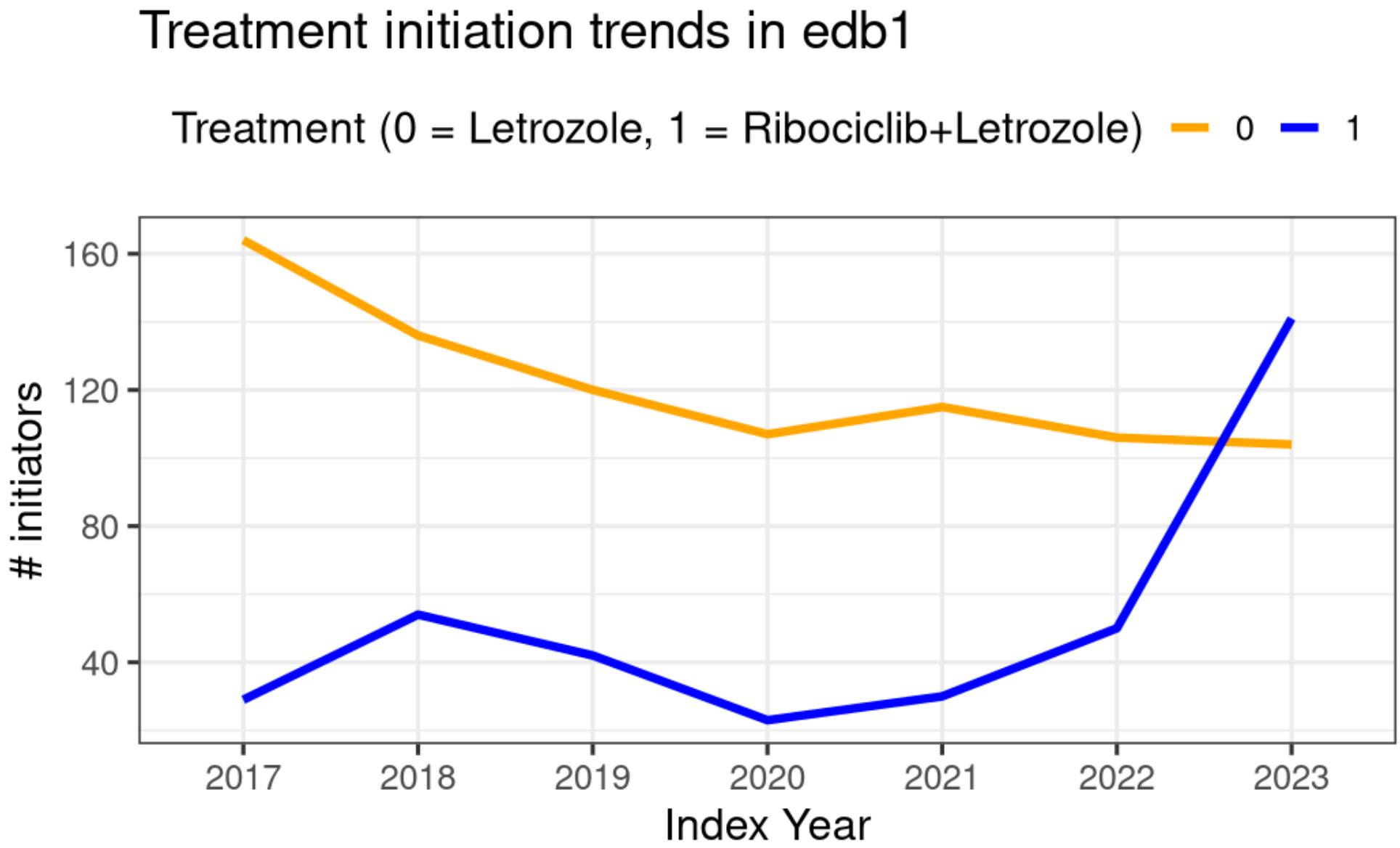


Figure 12. Treatment initiation trends by calendar year and treatment in EDB3.

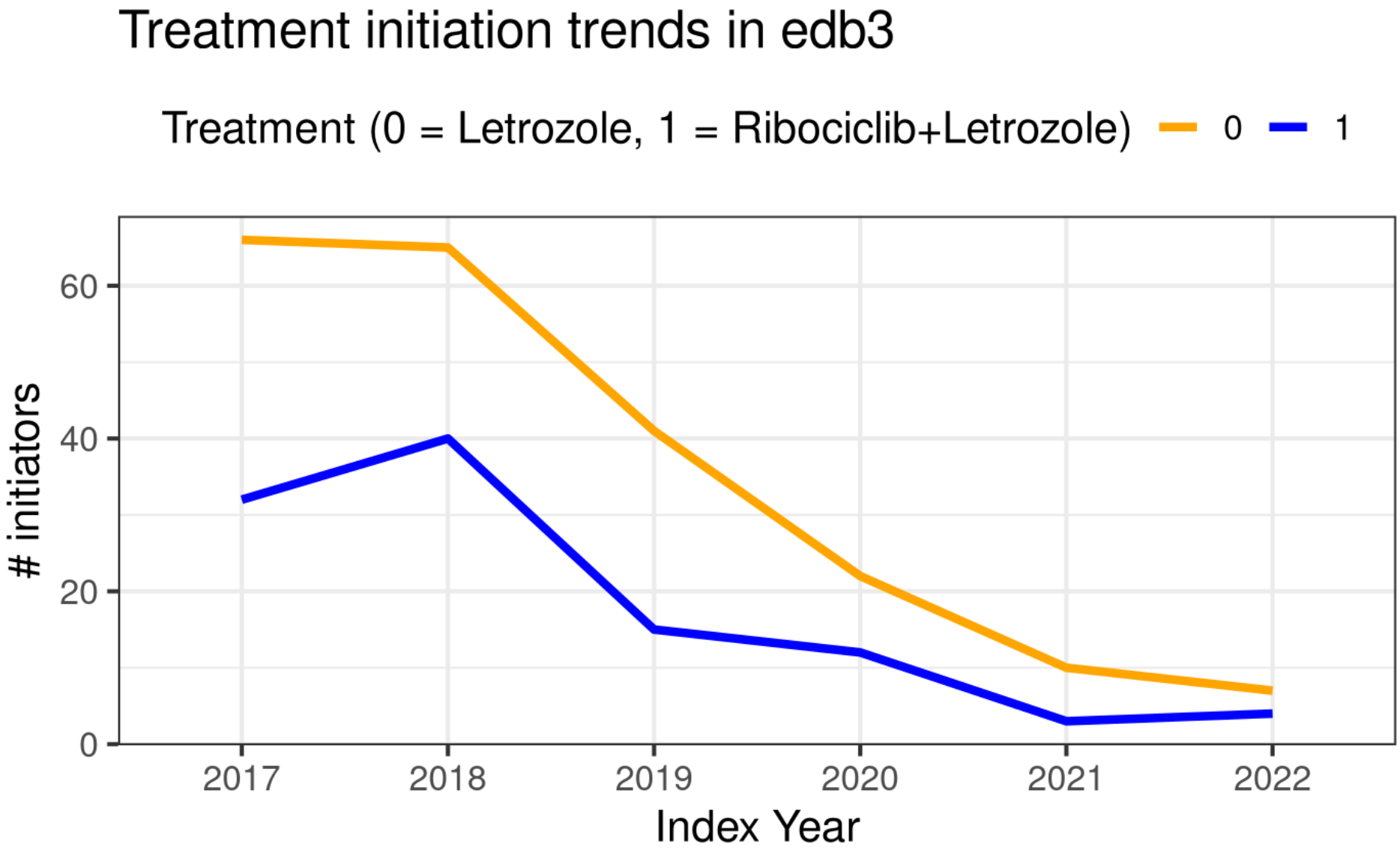
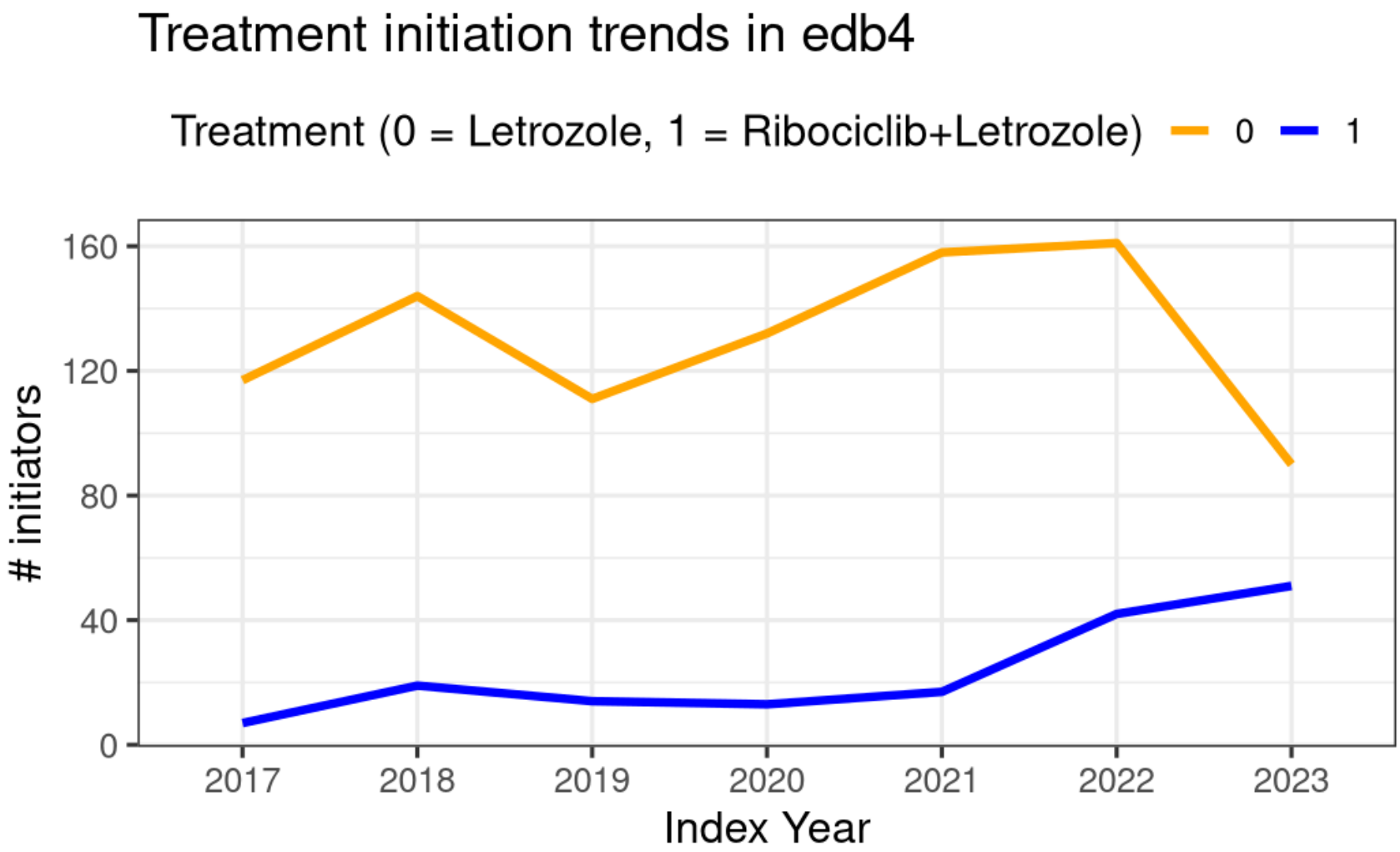


Figure 13. Treatment initiation trends by calendar year and treatment in EDB4.



NCTID	NCT01958021
Acronym	**MONALEESA2**
Protocol	https://www.dropbox.com/scl/fi/bc1c5xtfkso1emt3dczzd/nejmoa1609709_protocol.pdf?rlkey=xeyvwausi1ey6k8u01605ascp&st=rzj54y1r&dl=0
PMID	27717303
Indication	Breast cancer
Line of Therapy	Metastatic (first-line)
Exposure	Ribociclib plus letrozole - ribociclib 600 mg orally once daily on days 1–21 of every 28-day cycle, followed by 7 days off treatment, in combination with letrozole 2.5 mg orally once daily (continuously).
Comparisons	Placebo plus letrozole - placebo orally once daily on days 1–21 of every 28-day cycle, followed by 7 days off treatment, in combination with letrozole 2.5 mg orally once daily (continuously).
Emulated outcome	Overall survival (secondary endpoint in original trial)

Key Inclusion/Exclusion Criteria						Comment
Criteria	**RCT eligibility criteria (taken from the original trial protocol)**	**Clinical relevance**	**Emulation [EDB1]**	**Emulation [EDB3]**	**Emulation [EDB4]**	
Inclusion	Women with advanced (recurrent or metastatic) breast cancer who received no prior therapy for advanced disease	Relevant	Possible	Possible	Possible	
Inclusion	Patient is postmenopausal. Postmenopausal status is defined either by: - Prior bilateral oophorectomy - Age ≥60 - Age <60 and amenorrhea for 12 or more months (in the absence of chemotherapy, tamoxifen, toremifen, or ovarian suppression) and FSH and estradiol in the postmenopausal range per local normal range Note: For women with therapy-induced amenorrhea, serial measurements of FSH and/or estradiol are needed to ensure postmenopausal status. Ovarian radiation or treatment with a luteinizing hormone-releasing hormone agonist (LH-RHa) (goserelin acetate or leuprolide acetate) is not permitted for induction of ovarian suppression in this trial.	Relevant	Limited	Limited	Limited	To protect privacy, most databases only provide month- or year-level granularity of dates; post-menopausal status can be defined based on age and inferred by the fact that ribociclib was administered (indication in postmenopausal women)
Inclusion	Patient has a histologically and/or cytologically confirmed diagnosis of estrogen-receptor positive and/or progesterone receptor positive breast cancer by local laboratory.	Relevant	Possible	Possible	Possible	Keep if HR+ (ER+ or PR+) or missing because: - Ribociclib + letrozole is indicated for HR+/HER2-negative metastatic breast cancer.
Inclusion	Patient has HER2-negative breast cancer defined as a negative in situ hybridization test or an IHC status of 0, 1+ or 2+. If IHC is 2+, a negative in situ hybridization (FISH, CISH, or SISH) test is required by local laboratory testing.	Relevant	Possible	Possible	Possible	Keep if HER2-negative or HER2 status unknown, because: - Ribociclib + letrozole is indicated for HR+/HER2-negative metastatic breast cancer. - Letrozole monotherapy was historically the standard first-line treatment for HR+/HER2-negative metastatic breast cancer before CDK4/6 inhibitors became standard of care - In HER2+ HR+ metastatic breast cancer, letrozole is not typically given alone; instead, it is usually combined with anti-HER2 therapy (e.g., trastuzumab ± pertuzumab).

Inclusion	Patient must have either: - Measurable disease, i.e., at least one measurable lesion as per RECIST 1.1 criteria or, - At least one predominantly lytic bone lesion	Relevant	Limited	Limited	Limited	We are assuming that all subjects in RWD have measurable disease if they are receiving treatment
Inclusion	Patient has an Eastern Cooperative Oncology Group (ECOG) performance status 0 or 1	Relevant	Possible	Possible	Possible	ECOG implementation possible; high % missingness likely
Inclusion	Restriction to period 2017-2023 for initiator identification	Relevant	Possible	Possible	Possible	2017 was the approval year for ribociclib.
Exclusion	Patient who received any CDK4/6 inhibitor.	Relevant	Possible	Possible	Possible	
Exclusion	Patients with locally advanced (unresectable) or inflammatory breast cancer. - Patients who received adjuvant therapy for breast cancer are eligible. The treatment free interval since the last adjuvant treatment must be at least 12 months prior to randomization. - Prior therapy with letrozole in the adjuvant setting is permitted if the patient did not have progressive disease while on letrozole	Relevant	Limited	Limited	Limited	Patients were excluded if there was evidence of locally advanced (defined as stage 3B or 3C) or inflammatory breast cancer or patients received adjuvant therapy within 12 months prior to index (adjuvant treatment with letrozole will be allowed if it occurred at least 12 months prior to index)
Exclusion	Patient has a concurrent malignancy or malignancy within 3 years of randomization, with the exception of adequately treated, basal or squamous cell carcinoma, non-melanomatous skin cancer or curatively resected cervical cancer.	Relevant	Limited	Limited	Limited	EDB4 does not contain records on diagnoses of other malignancies besides breast cancer. Prior non-breast cancer malignancy within 3 years is approximated with advanced prior treatments based on recommendation by the data vendor for EDB4
Exclusion	Patient has active cardiac disease or a history of cardiac dysfunction including any of the following: - History of angina pectoris, symptomatic pericarditis, or myocardial infarction within 12 months prior to study entry - History of documented congestive heart failure (New York Heart Association functional classification III-IV) - Documented cardiomyopathy - Patient has a Left Ventricular Ejection Fraction (LVEF) < 50% as determined by Multiple Gated acquisition (MUGA) scan or echocardiogram (ECHO) - History of any cardiac arrhythmias, e.g., ventricular, supraventricular, nodal arrhythmias, or conduction abnormality in the previous 12 months. - On screening, any of the following cardiac parameters: bradycardia (heart rate < 50 at rest), tachycardia (heart rate > 90 at rest), PR interval > 220 msec, QRS interval >109 msec, or QTcF >450 msec. - Systolic blood pressure >160 or <90 mmHg	Relevant	Not implementable	Not implementable	Not implementable	Not captured well

Exclusion	Patient is currently receiving any of the following medications: - That are known strong inducers of CYP3A4: avasimibe, carbamazepine, mitotane, phenobarbital, phenytoin, rifabutin, rifampin (rifampicin), St. John's wort (hypericum perforatum) - That are known strong inhibitors of CYP3A4: boceprevir, clarithromycin, cobicistat, conivaptan, elvitegravir, indinavir, itraconazole, ketoconazole, lopinavir, mibefradil, nefazodone, nelfinavir, posaconazole, ritonavir, saquinavir, telaprevir, telithromycin, tipranavir, troleandomycin, voriconazole - That have a known risk to prolong the QT interval or induce Torsades de Pointes: Amiodarone, arsenic trioxide, astemizole, azithromycin, bepridil, chloroquine, chlorpromazine, cisapride, citalopram, clarithromycin, disopyramide, dofetilide, domperidone, droperidol, erythromycin, flecainide, halofantrine, haloperidol, ibutilide, levomethadyl, mesoridazine, methadone, moxifloxacin, pentamidine, pimozide, probucol, procainamide, quinidine, sotalol, sparfloxacin, terfenadine, thioridazine, vavdetanib - That have a narrow therapeutic window and are predominantly metabolized through CYP3A4: alfentanil, astemizole, cisapride, cyclosporine, diergotamine (dihydroergotamine), ergotamine, fentanyl, pimozide, quinidine, sirolimus, tacrolimus, terfenadine	Relevant	Not implementable	Not implementable	Not implementable	Not captured well
Exclusion	Record of systemic anticancer therapy other than exposure/comparator during exposure assessment window	Relevant	Possible	Possible	Possible	Not an explicit RCT criterion, added for database study. Exposure assessment window is based on proprietary business rules

NCTID	NCT01958021
Acronym	**MONALEESA2**
Protocol	https://www.dropbox.com/scl/fi/bc1c5xtfkso1emt3dczzd/nejmoa1609709_protocol.pdf?rlkey=xeyvwausi1ey6k8u01605ascp&st=rzj54y1r&dl=0
PMID	27717303
Indication	Breast cancer
Line of Therapy	Metastatic (first-line)
Exposure	Ribociclib plus letrozole - ribociclib 600 mg orally once daily on days 1–21 of every 28-day cycle, followed by 7 days off treatment, in combination with letrozole 2.5 mg orally once daily (continuously).
Comparisons	Placebo plus letrozole - placebo orally once daily on days 1–21 of every 28-day cycle, followed by 7 days off treatment, in combination with letrozole 2.5 mg orally once daily (continuously).
Emulated outcome	Overall survival (secondary endpoint in original trial)

Measurement eligibility criteria						Comment	encore.io function
Criteria	**Criteria rule as defined in original protocol**	**Time point/period of emulated measurement [days]**	**Emulation [EDB1]**	**Emulation [EDB3]**	**Emulation [EDB4]**		
Inclusion 1	Women with advanced (recurrent or metastatic) breast cancer who received no prior therapy for advanced disease	Advanced breast cancer: [-inf; 0] No prior therapy for adbanced disease: [MBC; -1]	First line of therapy needs to be for "Advanced" setting (LoT table)	First line of therapy needs to be for "Advanced" 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. First therapy for advanced disease after date of first metastasis	EDB4 does not have a well-curated line of therapy variable	EDB4: Derived from edbx_get_diagnosis_solid()
Inclusion 2	Patient is postmenopausal. Postmenopausal status is defined either by: - Prior bilateral oophorectomy - Age ≥60 - Age <60 and amenorrhea for 12 or more months (in the absence of chemotherapy, tamoxifen, toremifen, or ovarian suppression) and FSH and estradiol in the postmenopausal range per local normal range Note: For women with therapy-induced amenorrhea, serial measurements of FSH and/or estradiol are needed to ensure postmenopausal status. Ovarian radiation or treatment with a luteinizing hormone-releasing hormone agonist (LH-RHa) (goserelin acetate or leuprolide acetate) is not permitted for induction of ovarian suppression in this trial.	NA	NA	NA	NA	Although not directly captured in RWD, it is likely to be fulfilled given the alignment with the indication for the exposures of interest	
Inclusion 3	Patient has a histologically and/or cytologically confirmed diagnosis of estrogen-receptor positive and/or progesterone receptor positive breast cancer by local laboratory.	[-inf; 0]	Any hormone-positive or hormone-missing status to or on index date	Any hormone-positive or hormone-missing status to or on index date	Any hormone-positive or hormone-missing status to or on index date	If estrogen receptor or progestorone receptor missing and a patient received the exposures of interest, then is likely to be positive given the alignment with the indication for the exposures of interest	edbx_get_biomarker()
Inclusion 4	Patient has HER2-negative breast cancer defined as a negative in situ hybridization test or an IHC status of 0, 1+ or 2+. If IHC is 2+, a negative in situ hybridization (FISH, CISH, or SISH) test is required by local laboratory testing.	[-inf; 0]	Any HER2-negative or HER2-missing status prior to or on index date	Any HER2-negative or HER2-missing status prior to or on index date	Any HER2-negative or HER2-missing status prior to or on index date	If HER2 is missing and a patient received the exposures of interest, then HER2 is likely to be negative given the alignment with the indication for the exposures of interest	edbx_get_her2()
Inclusion 5	Patient must have either: - Measurable disease, i.e., at least one measurable lesion as per RECIST 1.1 criteria or, - At least one predominantly lytic bone lesion	NA	NA	NA	NA	It is reasonable to assume that all patients in RWD had measurable disease if they received treatment	

Inclusion 6	Patient has an Eastern Cooperative Oncology Group (ECOG) performance status 0 or 1	[-90; 0]	ECOG = 0 | 1	ECOG = 0 | 1	ECOG = 0 | 1		edbx_get_ecog()
Inclusion 7	Restriction to period 2017-2023	[0; 0]	Years 2017-2023	Years 2017-2023	Years 2017-2023		
Exclusion 1	Patient who received any CDK4/6 inhibitor.	[-inf; -1]	Prior treatment with ribociclib, palbociclib, abemaciclib	Prior treatment with ribociclib, palbociclib, abemaciclib	Prior treatment with ribociclib, palbociclib, abemaciclib		
Exclusion 2	Patients with locally advanced (unresectable) or inflammatory breast cancer. - Patients who received adjuvant therapy for breast cancer are eligible. The treatment free interval since the last adjuvant treatment must be at least 12 months prior to randomization. - Prior therapy with letrozole in the adjuvant setting is permitted if the patient did not have progressive disease while on letrozole	[-inf; 0] for prior locally advanced or inflammatory breast cancer [-365; -1] for adjuvant therapy, including letrozole	Prior locally advanced or inflammatory breast cancer or patients received adjuvant therapy within 12 months prior to index (adjuvant treatment with letrozole will be allowed if it occurred at least 12 months prior to index)	Prior locally advanced or inflammatory breast cancer or patients received adjuvant therapy within 12 months prior to index (adjuvant treatment with letrozole will be allowed if it occurred at least 12 months prior to index)	Prior locally advanced or inflammatory breast cancer or patients received adjuvant therapy within 12 months prior to index (adjuvant treatment with letrozole will be allowed if it occurred at least 12 months prior to index)	"Locally advanced (unresectable)" will be defined as stage 3B or 3C	
Exclusion 3	Patient has a concurrent malignancy or malignancy within 3 years of randomization, with the exception of adequately treated, basal or squamous cell carcinoma, non-melanomatous skin cancer or curatively resected cervical cancer.	[-1095; 0]	Record of cancer diagnosis within 1095 days prior to or on index day	Record of cancer diagnosis within 1095 days prior to or on index day	Record of non-index cancer diagnosis within 1095 days prior to or on index day	No diagnosis table for EDB4 available	edbx_get_diagnosis_solid()
Exclusion 4	Record of systemic anticancer therapy other than exposure/comparator during exposure assessment window	[proprietary business rules]	Exclusion if treatment is any other than ribociblic-plus-letrozole or letrozole alone	Exclusion if treatment is any other than ribociblic-plus-letrozole or letrozole alone	Exclusion if treatment is any other than ribociblic-plus-letrozole or letrozole alone	Exposure assessment window is based on proprietary business rules	