Supplementary Information

for

Moving Beyond Crude Trial–Real-World Comparisons Through Population Alignment and Decomposition

Juho Lähteenmaa, Pauliina Ilmonen, Elli-Noora Hännikäinen, Joonas Laaksolilja, Anders Ståhls, Kirsi Santti, Peeter Karihtala, Johanna Mattson

Contents

Supplementary Methods **3**

Supplementary Methods S1: Trial data extraction and harmonization 3

Supplementary Methods S2: HUS cohort construction and variable extraction . . 4

Supplementary Methods S3: Crude pCR estimation and uncertainty propagation . 5

Supplementary Methods S4: MAIC implementation and diagnostics 5

Supplementary Methods S5: Bayesian MRP with raking model specification . . . 5

Supplementary Methods S6: Shapley decomposition and Monte Carlo approximation 8

Supplementary Methods S7: Treatment-regimen definitions and treatment-completion sensitivity 10

Supplementary Tables **12**

Supplementary Table 1: Marginal variables used for baseline covariate alignment . 12

Supplementary Table 2: NeoSphere baseline characteristics (extracted) 14

Supplementary Table 3: KRISTINE baseline characteristics (extracted) 15

Supplementary Table 4: HELEN-006 baseline characteristics (extracted) 16

Supplementary Table 5: Outcome model coefficients (NeoSphere comparison) . . . 17

Supplementary Table 6: Outcome model coefficients with NAT discontinuation indicator (NeoSphere comparison) 18
Supplementary Table 7: Outcome model coefficients (KRISTINE comparison) . . 19
Supplementary Table 8: Outcome model coefficients with NAT discontinuation indicator (KRISTINE comparison) 20
Supplementary Table 9: Outcome model coefficients (HELEN-006 comparison) . . 21
Supplementary Table 10: Outcome model coefficients with NAT discontinuation indicator (HELEN-006 comparison) 22
Supplementary Table 11: NAT regimen group counts 24
Supplementary Table 12: NAT regimen group descriptions 25
Supplementary Table 13: MAIC-adjusted pCR rate summaries 26

Supplementary Figures **27**

Supplementary Fig. 1: MAIC weight distributions NeoSphere 27
Supplementary Fig. 2: MAIC weight distributions KRISTINE 28
Supplementary Fig. 3: MAIC weight distributions HELEN-006 29
Supplementary Fig. 4: Posterior pCR distributions NeoSphere B 30
Supplementary Fig. 5: Posterior pCR distributions KRISTINE B 31
Supplementary Fig. 6: Posterior pCR distributions HELEN-006 B 32
Supplementary Fig. 7: Bayesian MRP with raking model 33
Supplementary Fig. 8: Pertuzumab-trastuzumab cohort construction 34

Supplementary Methods

Supplementary Methods S1: Trial data extraction and harmonization

Clinical variables in the HUS registry were harmonized to match the categorical covariates reported in the target randomized trials.

Age was categorized using trial-specific cutpoints reported in the trial publications (e.g. < 50, 50–70, > 70 years in the HELEN-006 trial).

Hormone receptor status was defined as positive if either estrogen receptor (ER) or progesterone receptor (PR) expression was positive in the pre-treatment biopsy specimen.

Histology was mapped to the categories ductal, lobular, and other. Mixed or rare histological subtypes (including micropapillary and mixed ductal–lobular carcinomas) were grouped into the "other" category.

Tumor size was defined as the maximum preoperative tumor diameter measured from imaging (MRI preferred over ultrasound when both were available) and categorized as T1 (< 20 mm), T2 (20–50 mm), or T3 (> 50 mm).

HER2 status was harmonized using trial-specific definitions combining immunohistochemistry (IHC) and in situ hybridization (ISH) results (e.g. IHC3+ or IHC2+/ISH+).

Missing values in categorical covariates in the HUS registry were coded as a separate "N/A" category. When the published trial marginals did not include a corresponding category, the marginal count for this category was set to zero and the category was excluded from the raking constraints.

Supplementary Methods S2: HUS cohort construction and variable extraction

Data collection was conducted in two stages, see Supplementary Figure 8. First, relevant information was extracted from structured and semi-structured electronic medical records. In this context, *semi-structured* refers to data obtained using rule-based extraction methods, such as regular expressions, to identify relevant variables and clinical events. The majority of eligible cases were identified through this process. A detailed description of the data extraction logic is provided in the ONCOVALUE GitHub repository[1]. Second, to supplement the original cohort, additional cases were manually identified by reviewing records of patients who underwent a breast needle biopsy followed by the administration of pertuzumab and trastuzumab within a three-month window. Manual verification was applied to exclude non-NAT episodes, including treatment of residual disease.

Most key clinical variables were available in structured or semi-structured formats and were extracted using the aforementioned logic. However, some of the variables such as pre-operative tumor size, lymph node status, multifocality and bilaterality assessed by magnetic resonance imaging (MRI) or ultrasound imaging were not available in structured form and were therefore manually extracted from free-text clinical notes.

Among the manually extracted variables, baseline tumor size was defined as the maximum diameter of the largest tumor focus. This information was obtained primarily from magnetic resonance imaging reports and, when unavailable, from ultrasound assessments. Disease was classified as multifocal when explicitly described as such by the radiologist, including cases reported as having satellite lesions.

Baseline lymph node status was determined using radiological assessments, pathological biopsy results, or a combination of both. Among patients classified as lymph node–positive at baseline, 94% also had pathologically confirmed nodal involvement based on core needle or fine-needle aspiration biopsy.

Supplementary Methods S3: Crude pCR estimation and uncertainty propagation

Posterior distributions of the observed pCR rates were estimated for the RWD population, $\hat{\theta}_{\text{RWD}}$, and for each RCT population, $\hat{\theta}_{\text{RCT}}$. These quantities were estimated separately for each population using Bayesian binomial models. In the data application, posterior distributions for pCR rates in the HUS cohort and in each RCT treatment arm were obtained from the observed sample sizes and numbers of patients achieving pCR under a binomial model with a uniform Beta$(1, 1)$ prior. Markov chain Monte Carlo (MCMC) draws were sampled from $\hat{\theta}_{\text{RCT}}$ and $\hat{\theta}_{\text{RWD}}$ and propagated through the subsequent stages of the analysis. Throughout the article, posterior samples are denoted by overhats, as in $\hat{\theta}_{\text{RCT}}$.

Supplementary Methods S4: MAIC implementation and diagnostics

MAIC was implemented using custom Python code developed for this study. However, the open-source `indcomp` package by Cooper[2] was used as a methodological reference, and its core logic is adapted and extended to meet the specific requirements of the present analysis. Uncertainty was quantified using a multiplier bootstrap to construct confidence intervals. Effective sample sizes (ESS) were computed for each MAIC comparison, and weight distributions were examined to assess the stability of the reweighting procedure. Weight diagnostics are shown in Supplementary Figures 1–3.

Supplementary Methods S5: Bayesian MRP with raking model specification

The Bayesian MRP with raking approach comprises two components: a multilevel regression outcome model and a raking model, which are linked through a post-stratification step. Model development followed the Bayesian workflow framework[3] and included prior predic-

tive checks, prior testing using synthetic data designed to reflect key features of the observed data, posterior predictive checks, and sensitivity analyses conducted via self-normalized importance sampling. Modeling decisions were guided by the results of these model-checking steps. Supplementary Figure 7 provides a graphical representation of the Bayesian MRP with raking model.

Throughout this section, vectors are denoted by overbars (e.g. $\bar{p}$) or indexed notation (e.g. $(\theta_c)_{c\in C}$), and matrices by bold uppercase letters (e.g. $\mathbf{L}$). Subscripts $c \in C$ refer to contingency-table cells, indices k refer to covariates and j denotes the level of a covariate. The index i refers to individuals in the HUS cohort.

The proposed Bayesian MRP with raking framework builds on established components of multilevel regression and marginal alignment but differs from existing Bayesian raking approaches in its formulation and application. In contrast to the Bayes-raking framework of Si and Zhou[4], which incorporates an explicit population selection model, the present approach constrains the trial-aligned contingency table to remain close to the observed distribution through a structured multinomial deviation model. This formulation is tailored to comparative effectiveness settings in which individual-level data are available for the RW cohort but only aggregate baseline marginals are reported for trial populations.

The *outcome model* is specified as a hierarchical, analysis-of-variance (ANOVA)-style logistic regression with additive effects. Outcomes are linked to the model through a Bernoulli likelihood $y_i \sim \text{Bern}(\theta_i)$ with a logit link $\text{logit}(\theta_i)$ as described in Supplementary Figure 7. The level-specific coefficients $\beta_{k,j}$ for each covariate are assigned Student's t prior distributions with mean zero, $\nu = 4$ degrees of freedom, and with covariate-specific scale parameters τ_k. Each scale parameter τ_k is in turn assigned a half-normal prior with scale σ_τ, where σ_τ itself follows a half-normal hyperprior with standard deviation 0.7. The Student's t prior can be interpreted as a continuous mixture model that induces weak regularization and partial pooling across levels, while mitigating over-shrinkage in settings with moderate data sparsity.

The trial-aligned contingency table is modeled in the *raking component* that is shown in

the upper part of Supplementary Figure 7. This component further comprises two linked submodels. The first is a *deviation model* that constrains the trial-aligned distribution to remain close to the observed contingency distribution $\bar{q}$. It is specified as a multinomial regression with a softmax link,

$$p_c = \frac{\exp(\eta_c)}{\sum_{c' \in C} \exp(\eta_{c'})},$$

where the latent linear predictor η_c is defined as the sum of $\log(q_c)$ and a deviation term. The deviation term is constructed using additive, ANOVA-style effects based on the covariate levels given by cell c. This ANOVA structure substantially reduces dimensionality relative to a fully saturated specification in which each cell would have its own independent deviation parameter.

The deviation effects are assigned level-specific standard normal priors and are scaled by a global parameter σ_δ, which follows a half-normal prior. This formulation expresses the trial-aligned distribution $\bar{p}$ as a smooth deformation of the observed distribution $\bar{q}$, while allowing moderate departures when required to satisfy the target marginal constraints.

The normalized trial-aligned contingency vector $\bar{p}$ is then scaled by the target population size T_{target} to obtain implied cell counts. These cell counts are linearly mapped to variable-level marginals via the loading matrix $\mathbf{L}$. The resulting marginal counts are then softly aligned with the target marginals $\bar{N}^*$ using a Poisson likelihood defined at the level-wise margins, following an approach similar to that of Si and Zhou[4]. Contingency-table rows corresponding to variable levels with zero counts in the target marginals were excluded from the raking component, because they receive zero weight in the subsequent post-stratification step.

The final step is *post-stratification*, in which the cell-level pCR rate estimates $(\theta_c)_{c \in C}$ are combined with the normalized trial-aligned contingency vector $\bar{p} = (p_c)_{c \in C}$. The cell-level pCR rates are obtained from the outcome model evaluated at each contingency-table cell, as illustrated in the middle panel of Supplementary Figure 7. Post-stratification yields the

posterior distribution of the adjusted pCR rate, $\theta_{\text{RWD}\to\text{RCT}}$.

Sensitivity of this posterior distribution was evaluated with respect to the global scale parameters σ_τ and σ_δ, which represent the most subjective prior choices in the model and primarily govern the outcome model and the raking component, respectively. Sensitivity analysis was performed using importance sampling, in which posterior draws obtained under the original prior specification were reweighted to reflect alternative prior choices, following the approach described by Gelman et al.[5].

Supplementary Methods S6: Shapley decomposition and Monte Carlo approximation

The difference between the adjusted and observed pCR rates can be further decomposed into variable-level contributions, including a component corresponding to missing values in the RWD dataset. Shapley values[6] provide a unique additive decomposition that satisfies four fairness axioms: efficiency (the total effect is fully allocated), symmetry (variables with identical contributions receive equal attribution), the dummy property (irrelevant variables receive zero contribution), and additivity (contributions combine consistently across decompositions). The Shapley value of a variable is defined as its average marginal contribution. That is, it measures how much the overall outcome (in our data application, the adjusted pCR rate) changes when the variable is added to each possible subset of the other variables, averaged over all such subsets. Shapley values have been used widely in empirical literature in varying decomposition tasks (see e.g.[7] and[8]).

In our data application, the Shapley decomposition was carried out in two parts: for contingency table rows included in the raking model and separately for rows not included in the raking model, namely rows associated with zero target marginals, including those with missing values in the raking variables.

In the first part, posterior draws from the Bayesian MRP with raking model were used to estimate the posterior distribution of variable-wise Shapley values. This is possible because

the deviation parameters $\hat{\delta}_{k,j(c)}$, multiplied by the global scale parameter $\hat{\sigma}_\delta$ in the deviation model (see the upper-right corner of Supplementary Figure 7), encode departures from the original log-contingency proportions $\log(q)$.

For each posterior draw, the adjusted overall pCR rate can be evaluated for any subset of raking variables. This can be done by propagating the trial-aligned distribution through the Bayesian MRP with raking model as described in Supplementary Methods S5, but including only the deviation effects associated with the variables in the subset while keeping the remaining deviations fixed at the baseline log-contingency proportions $\log q$. Within each posterior draw, variable-wise Shapley values can then be approximated by averaging marginal contributions across random permutations of the variables. In the present work, the Shapley values were approximated because the number of possible variable orderings grows factorially with the number of variables, thus making the exact computation of Shapley values computationally infeasible. Therefore a Monte Carlo permutation approximation was used following ApproShapley algorithm described in[9]. This procedure was repeated across posterior draws yielding the posterior distribution of the Shapley value for each variable.

In the second part, the Shapley decomposition for the remaining component corresponding to observations not included in the raking model was estimated separately. For this component, posterior distributions of the pCR rate were obtained using binomial models fitted to each subset of variables. Importantly, this part of the decomposition was constructed by combining posterior summaries from distinct models and therefore the corresponding posterior draws do not correspond to a single joint posterior distribution, in contrast to the Shapley values derived in the first part.

The contributions from these two components together sum to the total baseline-adjustment difference defined in the main article.

Supplementary Methods S7: Treatment-regimen definitions and treatment-completion sensitivity

NAT regimens used in the HUS cohort and their frequencies are summarized in Supplementary Table 11, with detailed regimen descriptions in Supplementary Table 12. Dual anti-HER2 therapy with pertuzumab and trastuzumab has been used in NAT at HUS since 2015. However, a major change in clinical practice occurred in 2019, when national guidelines were updated to allow neoadjuvant therapy also for operable HER2-positive breast cancer.

From 2020 onwards, the combination of pertuzumab and trastuzumab was established as the default NAT backbone for HER2-positive disease at HUS. During the period 2020–2022, anthracycline-based chemotherapy was routinely incorporated into neoadjuvant regimens and was typically administered as 3–4 cycles of THP, followed by 3–4 cycles of anthracycline-based chemotherapy consisting of cyclophosphamide combined with either epirubicin or doxorubicin, with or without fluorouracil (C[E/A][F]). During this period, omission of anthracyclines was generally reserved for patients demonstrating an excellent early treatment response. From 2022 onwards, HER2-positive NAT was primarily based on THP alone (typically 6–8 cycles), with anthracyclines reserved for cases with poor treatment response. Carboplatin was administered to fewer than three patients in the HUS cohort.

Overall, 80% of patients in the HUS cohort completed a full NAT course according to contemporaneous institutional guidelines, defined as either THP $\times \geq 3$ followed by C[E/A][F] $\times \geq 3$ under the earlier treatment protocol, or THP $\times \geq 6$ under the newer protocol. When the definition of a completed neoadjuvant course was relaxed to correspond to the treatment intensity used in the NeoSphere THP arm—namely THP $\times \geq 3$ followed by C[E/A][F] $\times \geq 3$, or THP $\times \geq 4$—this proportion increased to 90%.

Docetaxel was substituted by paclitaxel in patients who were unable to tolerate docetaxel. Drug dosing levels and dose-level changes were not examined in the present analysis.

These variations in regimen composition and treatment intensity reflect routine clinical practice and are not captured by baseline covariate adjustment; they are therefore expected

to contribute primarily to the residual difference.

Supplementary Tables

Supplementary Table 1: Marginal variables used for baseline covariate alignment

Table 1: Adjusted marginal variables and level definitions by trial arm.

Variable	**NeoSphere B (THP)**	**KRISTINE B (TCHP)**	**HELEN-006 B (TCHP)**
Age group	Under 50 50–77 Over 77	Under 49 Over 49	Under 50 50–70 Over 70
Hormone receptor status	Negative Positive	Negative Positive	Negative Positive
Baseline nodal status	Negative Positive N/A	Negative Positive [a]	Negative Positive N/A
Tumor size category	T1 T2 T3	T1 T2–T3	T1–T2 T3
Bilaterality	True False	True False	True False
Histologic grade			1–2 3 N/A
Histology subtype			Ductal Lobular Other
Ki-67 group (30% cutoff)[b]			Under 30% Over 30%
HER2 IHC/ISH category			IHC2+ and ISH+ IHC3+
Clinical stage			II–III [c]

[a] KRISTINE study excluded the patients if they had a positive sentinel lymph node biopsy at baseline.
[b] Ki-67 index and MIB-1 index are used interchangeably throughout in this paper.
[c] Based on the combination of the tumor size (T1 or T2-T3) and the lymph node status (Positive or Negative).

Supplementary Table 2: NeoSphere baseline characteristics (extracted)

Characteristic	N	Summary
Median age	107	50 (Range 28-77)
Median tumour size [mm]	107	55 (Range 20-150)
N1 + N2 + N3	75	71%
N0	31	29%
Bilateral: False	107	100%
ER or PR positive	50	47%
ER and PR negative	57	53%
Operable[a1]	65	61%
Locally advanced[a2]	32	30%
Inflammatory[a3]	10	9%

[a1] T2-3, N0-1; IIA, IIB, IIIA
[a2] T2-3, N2-3 or T4a-c, N0-3; IIIA, IIIB, IIIC
[a3] T4d, N0-3; IIIB-IIIC

Table 2: Baseline characteristics for the NeoSphere arm B (THP) population, restricted to variables used for adjustment and staging.

Supplementary Table 3: KRISTINE baseline characteristics (extracted)

Characteristic	N	Summary
Median age	221	49 (IQR 41-57)
Negative sentinel lymph node biopsy at baseline	221	100%
Bilateral: False	221	100%
ER or PR positive	138	62%
ER and PR negative	83	38%
Clinical stage IIA - IIIA	183	83%
Clinical stage IIIB - IIIC	38	17%

Table 3: Baseline characteristics for the KRISTINE arm B (TCHP) population, restricted to variables used for adjustment and staging.

Supplementary Table 4: HELEN-006 baseline characteristics (extracted)

Characteristic	N	Summary
Median age	337	50 (IQR 44-56)
T stage: T1 to T2	281	83%
T stage: T3 to T4	56	17%
Nodal status: Positive	247	73%
Nodal status: Negative	90	27%
Bilateral: False	337	100%
Tumor grade: 1	1	$< 1\%$
Tumor grade: 2	142	42%
Tumor grade: 3	194	58%
Histological type: Ductal	317	94%
Histological type: Lobular	4	1%
Histological type: Other	16	5%
ER or PR positive	193	57%
ER and PR negative	144	43%
IHC: 2+ and ISH	104	31%
IHC: 3+	233	69%
Ki-67 index ≤ 30 %	75	22%
Ki-67 index > 30 %	262	78%
Disease stage: II	212	63%
Disease stage: III	125	37%

Table 4: Baseline characteristics for the HELEN-006 arm B (TCHP) population, restricted to variables used for adjustment and staging.

Supplementary Table 5: Outcome model coefficients (NeoSphere comparison)

	mean	sd	hdi_3%	hdi_97%
Age group[Under 50]	0.09	0.19	-0.26	0.46
Age group[50 - 77]	-0.18	0.22	-0.70	0.12
Age group[Over 77]	0.09	0.33	-0.48	0.78
Baseline nodal status[N/A]	-0.27	0.29	-0.87	0.14
Baseline nodal status[neg]	0.06	0.18	-0.26	0.43
Baseline nodal status[pos]	0.21	0.20	-0.09	0.60
Tumor size category[T1]	-0.07	0.18	-0.45	0.24
Tumor size category[T2]	-0.16	0.18	-0.52	0.11
Tumor size category[T3]	0.04	0.22	-0.38	0.51
Tumor size category[N/A]	0.19	0.22	-0.14	0.64
Hormone receptor status[Negative]	0.71	0.15	0.42	0.99
Hormone receptor status[Positive]	-0.71	0.15	-0.99	-0.42
Bilaterality[False]	0.08	0.23	-0.32	0.61
Bilaterality[True]	-0.08	0.23	-0.61	0.32

Table 5: Posterior summaries of coefficients for the NeoSphere outcome model.

Supplementary Table 6: Outcome model coefficients with NAT discontinuation indicator (NeoSphere comparison)

	mean	sd	hdi_3%	hdi_97%
Age group[Under 50]	0.06	0.20	-0.28	0.48
Age group[50 - 77]	-0.17	0.22	-0.61	0.19
Age group[Over 77]	0.11	0.34	-0.46	0.84
Baseline nodal status[N/A]	-0.28	0.30	-0.89	0.15
Baseline nodal status[neg]	0.07	0.18	-0.25	0.44
Baseline nodal status[pos]	0.21	0.20	-0.10	0.60
Tumor size category[T1]	-0.07	0.18	-0.46	0.25
Tumor size category[T2]	-0.17	0.18	-0.51	0.14
Tumor size category[T3]	0.04	0.22	-0.37	0.53
Tumor size category[N/A]	0.20	0.23	-0.14	0.67
Hormone receptor status[Negative]	0.72	0.15	0.44	0.99
Hormone receptor status[Positive]	-0.72	0.15	-0.99	-0.44
Bilaterality[False]	0.07	0.23	-0.32	0.60
Bilaterality[True]	-0.07	0.23	-0.60	0.32
NAT discontinuation[False]	0.17	0.19	-0.12	0.54
NAT discontinuation[True]	-0.17	0.19	-0.54	0.12

Table 6: Posterior summaries of coefficients for the NeoSphere outcome model with the treatment discontinuation indicator.

Supplementary Table 7: Outcome model coefficients (KRISTINE comparison)

	mean	sd	hdi_3%	hdi_97%
Age group[Under 49]	0.15	0.13	-0.07	0.38
Age group[Over 49]	-0.15	0.13	-0.38	0.07
Baseline nodal status[False]	-0.08	0.11	-0.31	0.09
Baseline nodal status[True]	0.08	0.11	-0.09	0.31
Tumor size category[N/A]	0.20	0.22	-0.15	0.63
Tumor size category[T1]	-0.06	0.17	-0.45	0.21
Tumor size category[T2-T3]	-0.13	0.16	-0.44	0.12
Hormone receptor status[Negative]	0.72	0.15	0.45	1.00
Hormone receptor status[Positive]	-0.72	0.15	-1.00	-0.45
Bilaterality[False]	0.10	0.22	-0.27	0.59
Bilaterality[True]	-0.10	0.22	-0.59	0.27

Table 7: Posterior summaries of coefficients for the KRISTINE outcome model.

Supplementary Table 8: Outcome model coefficients with NAT discontinuation indicator (KRISTINE comparison)

	mean	sd	hdi_3%	hdi_97%
Age group[Under 49]	0.14	0.13	-0.07	0.38
Age group[Over 49]	-0.14	0.13	-0.38	0.07
Baseline nodal status[False]	-0.08	0.11	-0.30	0.10
Baseline nodal status[True]	0.08	0.11	-0.10	0.30
Tumor size category[N/A]	0.18	0.20	-0.11	0.59
Tumor size category[T1]	-0.06	0.16	-0.39	0.23
Tumor size category[T2-T3]	-0.12	0.15	-0.43	0.10
Hormone receptor status[Negative]	0.71	0.15	0.45	0.99
Hormone receptor status[Positive]	-0.71	0.15	-0.99	-0.45
Bilaterality[False]	0.09	0.23	-0.30	0.60
Bilaterality[True]	-0.09	0.23	-0.60	0.30
NAT discontinuation[False]	0.01	0.11	-0.21	0.24
NAT discontinuation[True]	-0.01	0.11	-0.24	0.21

Table 8: Posterior summaries of coefficients for the KRISTINE outcome model with the treatment discontinuation indicator.

Supplementary Table 9: Outcome model coefficients (HELEN-006 comparison)

	mean	sd	hdi_3%	hdi_97%
Age group[Under 50]	0.08	0.15	-0.16	0.41
Age group[50 - 70]	-0.10	0.15	-0.41	0.15
Age group[Over 70]	0.02	0.17	-0.35	0.35
Histologic grade[1 or 2]	-0.16	0.19	-0.54	0.13
Histologic grade[3]	0.04	0.13	-0.19	0.34
Histologic grade[N/A]	0.12	0.18	-0.17	0.51
Histology subtype[Ductal carcinoma]	0.40	0.26	-0.07	0.86
Histology subtype[Lobular carcinoma]	-0.20	0.31	-0.84	0.35
Histology subtype[other]	-0.19	0.23	-0.67	0.21
Ki-67 group (30% cutoff)[under 30%]	0.02	0.11	-0.19	0.23
Ki-67 group (30% cutoff)[over 30%]	-0.02	0.11	-0.23	0.19
Baseline nodal status[N/A]	-0.21	0.27	-0.75	0.19
Baseline nodal status[neg]	0.03	0.16	-0.27	0.37
Baseline nodal status[pos]	0.18	0.19	-0.12	0.58
Tumor size category[T1-T2]	-0.20	0.21	-0.62	0.11
Tumor size category[T3]	0.07	0.23	-0.36	0.54
Tumor size category[other]	0.14	0.23	-0.21	0.69
Hormone receptor status[Negative]	0.64	0.16	0.35	0.93
Hormone receptor status[Positive]	-0.64	0.16	-0.93	-0.35
HER2 IHC/ISH category[IHC2+ and ISH+]	-0.65	0.25	-1.11	-0.19
HER2 IHC/ISH category[IHC3+]	0.73	0.23	0.31	1.15
HER2 IHC/ISH category[other]	-0.08	0.33	-0.73	0.51
Bilaterality[False]	0.02	0.22	-0.41	0.47
Bilaterality[True]	-0.02	0.22	-0.47	0.41
Clinical stage[II-III]	-0.08	0.16	-0.40	0.20
Clinical stage[other]	0.08	0.16	-0.20	0.40

Table 9: Posterior summaries of coefficients for the HELEN-006 outcome model.

Supplementary Table 10: Outcome model coefficients with NAT discontinuation indicator (HELEN-006 comparison)

	mean	sd	hdi_3%	hdi_97%
Age group[Under 50]	0.08	0.14	-0.17	0.38
Age group[50 - 70]	-0.10	0.15	-0.41	0.17
Age group[Over 70]	0.02	0.17	-0.31	0.35
Histologic grade[1 or 2]	-0.15	0.19	-0.53	0.14
Histologic grade[3]	0.04	0.14	-0.21	0.33
Histologic grade[N/A]	0.11	0.18	-0.19	0.50
Histology subtype[Ductal carcinoma]	0.38	0.25	-0.04	0.85
Histology subtype[Lobular carcinoma]	-0.19	0.30	-0.77	0.36
Histology subtype[other]	-0.19	0.23	-0.65	0.20
Ki-67 group (30% cutoff)[under 30%]	0.02	0.11	-0.19	0.24
Ki-67 group (30% cutoff)[over 30%]	-0.02	0.11	-0.24	0.19
Baseline nodal status[N/A]	-0.22	0.27	-0.76	0.19
Baseline nodal status[neg]	0.04	0.17	-0.27	0.38
Baseline nodal status[pos]	0.18	0.20	-0.10	0.59
Tumor size category[T1-T2]	-0.21	0.21	-0.62	0.12
Tumor size category[T3]	0.08	0.23	-0.33	0.58
Tumor size category[other]	0.13	0.23	-0.26	0.62
Hormone receptor status[Negative]	0.63	0.16	0.33	0.93
Hormone receptor status[Positive]	-0.63	0.16	-0.93	-0.33
HER2 IHC/ISH category[IHC2+ and ISH+]	-0.65	0.25	-1.15	-0.21
HER2 IHC/ISH category[IHC3+]	0.72	0.23	0.30	1.17
HER2 IHC/ISH category[other]	-0.07	0.33	-0.70	0.54
Bilaterality[False]	0.02	0.20	-0.39	0.43
Bilaterality[True]	-0.02	0.20	-0.43	0.39
Clinical stage[II-III]	-0.08	0.15	-0.40	0.19
Clinical stage[other]	0.08	0.15	-0.19	0.40
NAT discontinuation[False]	0.02	0.12	-0.22	0.26
NAT discontinuation[True]	-0.02	0.12	-0.26	0.22

Table 10: Posterior summaries of coefficients for the HELEN-006 outcome model with the treatment discontinuation indicator.

Supplementary Table 11: NAT regimen group counts

NAT group	Patients (N)
THP → AC	109
THP	143
THP → TH	6
THP → AC ± F → THP	< 3
THP → T	< 3
THP → AC → HP	< 3
THP → TH → AC ± F	4
TCHP → TCH → AC	< 3

Table 11: Number of patients in each NAT group in the HUS cohort. Counts less than three are masked with < 3 for privacy. An arrow indicates a regimen change during NAT. For example, notation THP → AC means that the treatment started with THP and continued with AC. For the regimen descriptions, see Supplementary Table 12.

Abbreviations: THP = taxane + pertuzumab + trastuzumab; TH = taxane + trastuzumab; AC = anthracycline + cyclophosphamide; F = fluorouracil; T = taxane; HP = pertuzumab + trastuzumab; TCHP = taxane + carboplatin + pertuzumab + trastuzumab; TCH = taxane + carboplatin + trastuzumab.

Supplementary Table 12: NAT regimen group descriptions

Group	Regimens
THP → AC	DTHP → EC DTHP → PTHP → AC DTHP → FEC DTHP → AC DTHP → PTHP → FEC PTHP → EC DTHP → PTHP → EC DTHP → AC → AC DTHP → FEC → AC
THP	PTHP DTHP DTHP → PTHP PTHP → DTHP DTHP → PTH → PTHP
THP → TH	DTHP → PTH DTHP → DTH PTHP → PTH
THP → AC ± F → THP	DTHP → AC → DTHP DTHP → FEC → DTHP
THP → T	PTHP → P
THP → AC → HP	DTHP → AC → PTH
THP → TH → AC ± F	DTHP → DTH → FEC DTHP → PTH → EC DTHP → DTH → EC
TCHP → TCH → AC	TCHP → TCH → AC

Abbreviations: DTHP = Docetaxel + Trastuzumab + Pertuzumab; PTHP = Paclitaxel + Trastuzumab + Pertuzumab; DTH = Docetaxel + Trastuzumab; PTH = Paclitaxel + Trastuzumab; AC = Doxorubicin (+/- hydrochloride) + Cyclophosphamide; EC = Epirubicin + Cyclophosphamide; FEC = Fluorouracil + Epirubicin + Cyclophosphamide; TCHP = Docetaxel + Carboplatin + Trastuzumab + Pertuzumab; TCH = Docetaxel + Carboplatin + Trastuzumab; P = Paclitaxel.

Table 12: NAT group descriptions. An arrow indicates a regimen change during NAT.

Supplementary Table 13: MAIC-adjusted pCR rate summaries

Trial comparison	Median pCR	95% bootstrap interval
NeoSphere	0.58	0.46–0.68
KRISTINE	0.41	0.30–0.53
HELEN-006	0.49	0.40–0.58

Table 13: MAIC-adjusted pCR rate summaries. Confidence intervals were obtained using a multiplier bootstrap.

Supplementary Figures

Supplementary Fig. 1: MAIC weight distributions NeoSphere

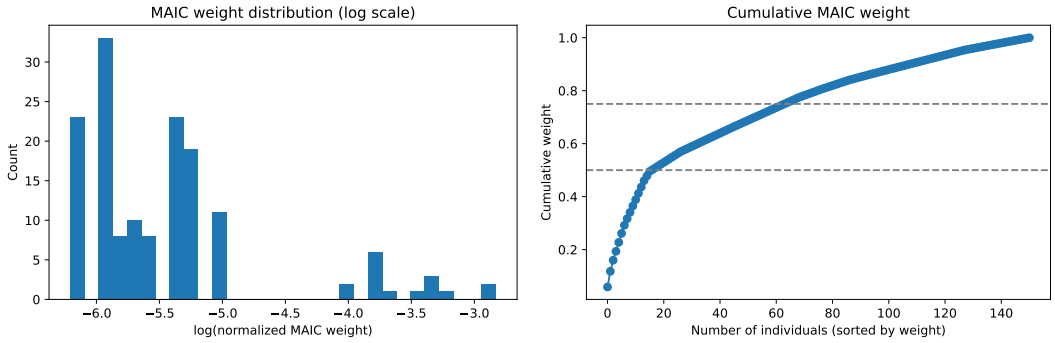


Figure 1: MAIC weight diagnostics for the NeoSphere comparison. The effective sample size (ESS) had median 31 with 95% bootstrapped confidence interval (CI) from 23 to 38. Top panel: Histogram of individual log-weights from the MAIC model, illustrating a multimodal distribution with a small number of observations receiving relatively large weights. Bottom panel: Cumulative weight proportion as a function of the number of individuals, showing that the 20 highest-weighted individuals account for approximately 50% of the total weight.

Supplementary Fig. 2: MAIC weight distributions KRISTINE

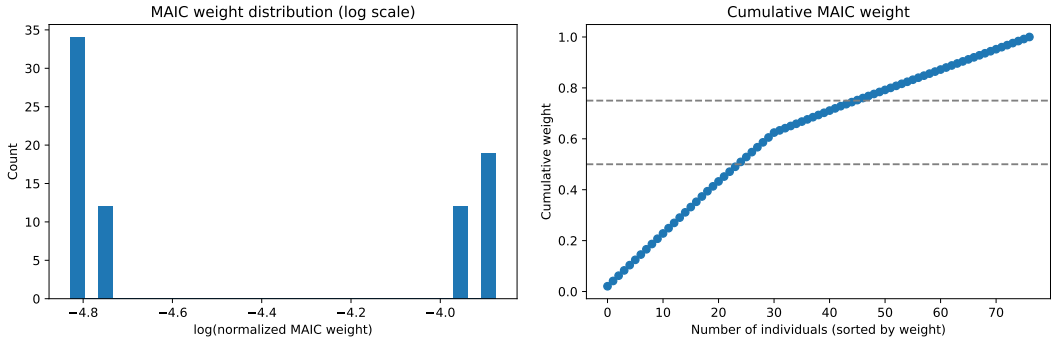


Figure 2: MAIC weight diagnostics for the KRISTINE comparison. The ESS had median 34 with 95% bootstrapped CI from 27 to 40. Top panel: Distribution of individual log-weights from the MAIC model. Bottom panel: Cumulative weight proportion as a function of the number of individuals.

Supplementary Fig. 3: MAIC weight distributions HELEN-006

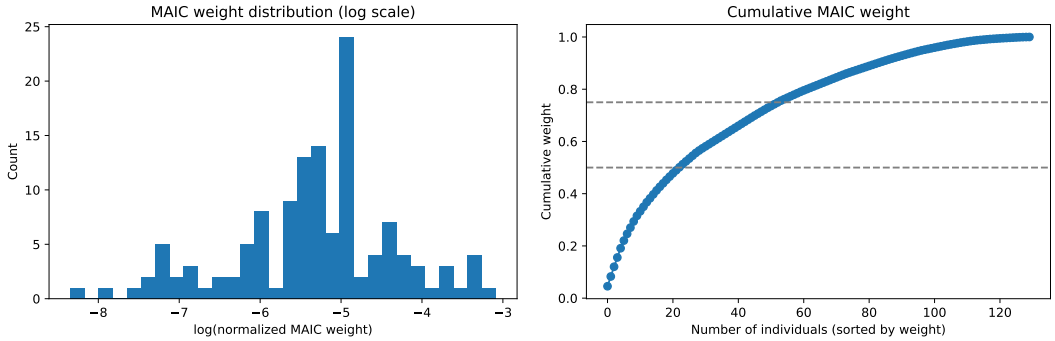


Figure 3: MAIC weight diagnostics for the HELEN-006 comparison. The ESS had median 38 with 95% bootstrapped CI from 31 to 44. Top panel: Distribution of individual log-weights from the MAIC model. Bottom panel: Cumulative weight proportion as a function of the number of individuals.

Supplementary Fig. 4: Posterior pCR distributions NeoSphere B

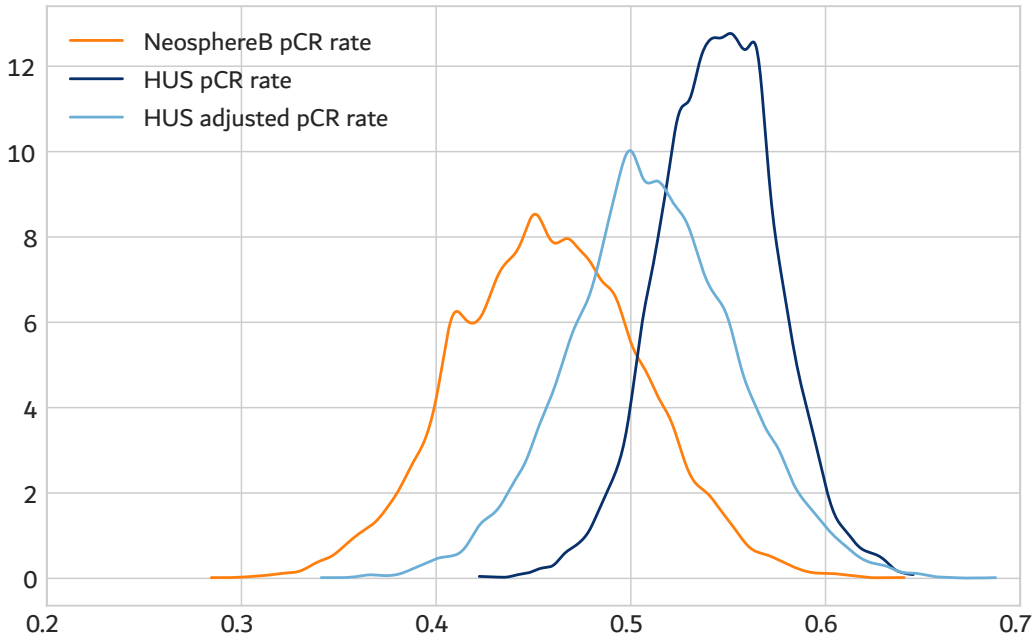


Figure 4: Posterior distributions of estimated pCR rates for the NeoSphere B comparison. The figure shows the pCR rate for the NeoSphere B arm, the crude HUS pCR rate, and the adjusted HUS pCR rate after baseline covariate alignment.

Supplementary Fig. 5: Posterior pCR distributions KRISTINE B

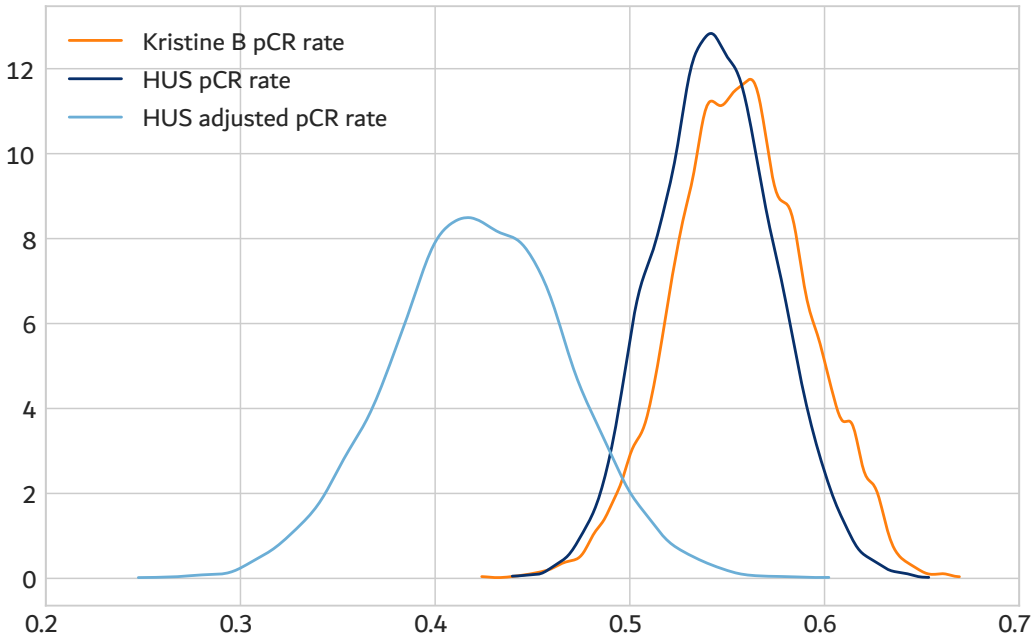


Figure 5: Posterior distributions of estimated pathological complete response (pCR) rates for the KRISTINE B comparison. The figure shows the pCR rate for the KRISTINE B arm, the crude HUS pCR rate, and the adjusted HUS pCR rate after baseline covariate alignment.

Supplementary Fig. 6: Posterior pCR distributions HELEN-006 B

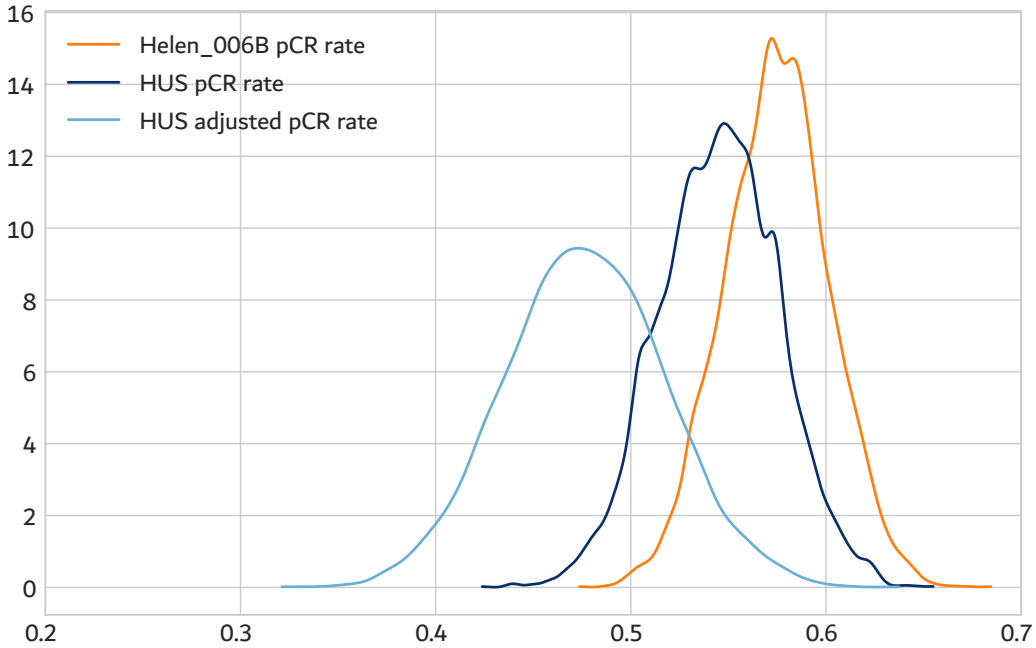


Figure 6: Posterior distributions of estimated pathological complete response (pCR) rates for the HELEN-006 B comparison. The figure shows the pCR rate for the HELEN-006 B arm, the crude HUS pCR rate, and the adjusted HUS pCR rate after baseline covariate alignment.

Supplementary Fig. 7: Bayesian MRP with raking model

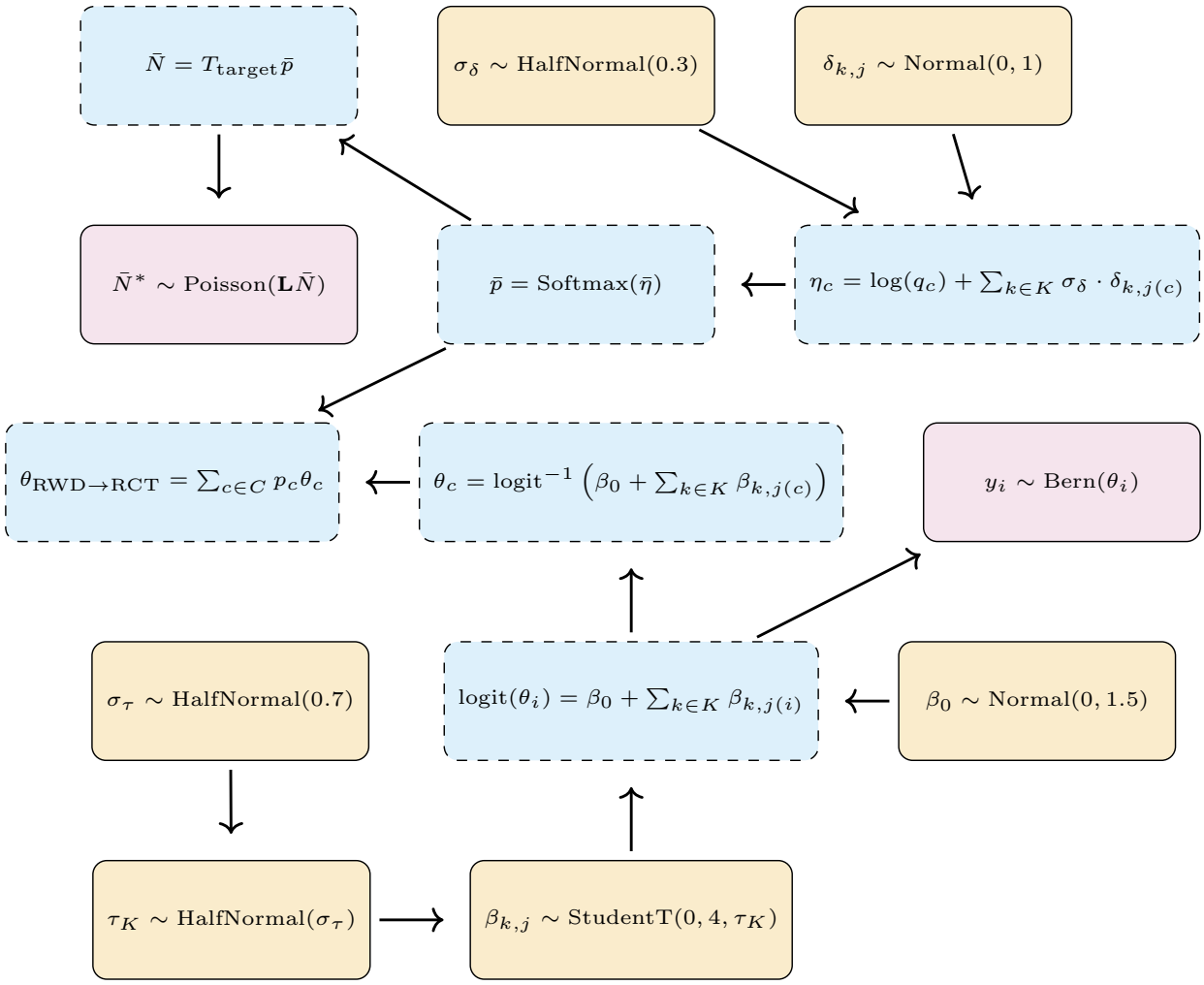


Figure 7: Bayesian MRP with raking. Blue boxes denote deterministic functions, orange boxes denote prior distributions, and purple boxes denote likelihood components. Arrows indicate conditional dependence in the generative model.

Supplementary Fig. 8: Pertuzumab-trastuzumab cohort construction

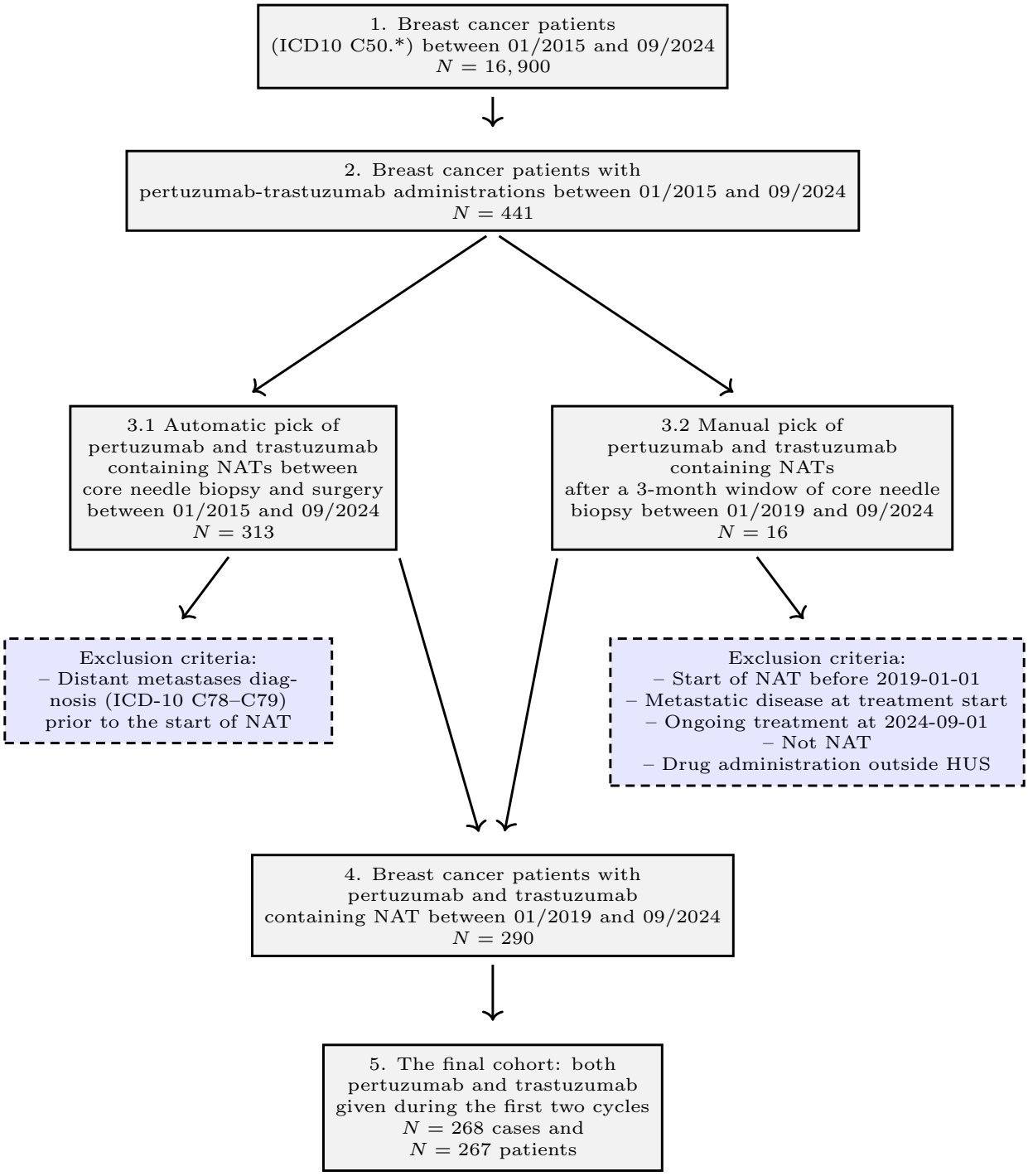


Figure 8: Flowchart of pertuzumab-trastuzumab cohort construction

References

[1] ONCOVALUE Consortium. Helsinki University Hospital Neoadjuvant Breast Cancer Cohort; 2025. GitHub repository. https://github.com/ONCOVALUE/neoadjuvant_cohort_hus.

[2] Cooper A. indcomp: Individual-level comparison methods; 2023. GitHub repository. Available from: https://github.com/AidanCooper/indcomp.

[3] Gelman A, Vehtari A, Simpson D, Margossian CC, Carpenter B, Yao Y, et al. Bayesian workflow. arXiv preprint arXiv:201101808. 2020.

[4] Si Y, Zhou P. Bayes-Raking: Bayesian Finite Population Inference with Known Margins. Journal of Survey Statistics and Methodology. 2020 Jun;9(4):833-55.

[5] Gelman A, Carlin J, Stern H, Dunson D, Aki V, Rubin D. Bayesian data analysis. J Chem Inf Model. 2013;53:1689-99.

[6] Shapley LS. A value for n-person games. In: Kuhn HW, Tucker AW, editors. Contributions to the Theory of Games II. Princeton, NJ: Princeton University Press; 1953. p. 307-17.

[7] Owen AB. Sobol' Indices and Shapley Value. SIAM/ASA Journal on Uncertainty Quantification. 2014;2(1):245-51.

[8] Shorrocks AF. Decomposition procedures for distributional analysis: a unified framework based on the Shapley value. The Journal of Economic Inequality. 2013;11(1):99-126.

[9] Castro J, Gómez D, Tejada J. Polynomial calculation of the Shapley value based on sampling. Computers & Operations Research. 2009;36(5):1726-30. Selected papers presented at the Tenth International Symposium on Locational Decisions (ISOLDE X).