Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

Juho Lähteenmaa

juho.lahteenmaa@hus.fi

Aalto University

Pauliina Ilmonen

Aalto University

Elli-Noora Hännikäinen

University of Helsinki

Joonas Laaksolilja

Helsinki University Hospital

Anders Ståhls

Helsinki University Hospital

Kirsi Santti

Helsinki University Hospital

Peeter Karihtala

Helsinki University Hospital

Johanna Mattson

Helsinki University Hospital

Research Article

Keywords:

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11551757/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

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

Juho Lähteenmaa[1,2], Pauliina Ilmonen[1], Elli-Noora Hännikäinen[3], Joonas Laaksolilja[2], Anders Ståhls[4], Kirsi Santti[5], Peeter Karihtala[5], and Johanna Mattson[5]

[1]Department of Mathematics and Systems Analysis, Aalto University School of Science, Espoo, Finland

[2]IT Management, Helsinki University Hospital, Helsinki, Finland

[3]University of Helsinki, Helsinki, Finland

[4]HUSLAB, Helsinki University Hospital, Helsinki, Finland

[5]Department of Oncology, Comprehensive Cancer Center, Helsinki University Hospital and University of Helsinki, Helsinki, Finland

Corresponding author: Juho Lähteenmaa, juho.lahteenmaa@hus.fi

Funding: This work was conducted as part of the ONCOVALUE project which has received funding from the European Union's Horizon Europe Research and Innovation Programme under grant agreement No. 101095245. The views expressed are those of the authors and do not necessarily reflect those of the European Union or the granting authority. Neither the European Union nor the granting authority can be held responsible for them.

Abstract

Background: Randomized clinical trials provide high internal validity but may not fully reflect outcomes in routine clinical practice. We propose a population-alignment and decomposition framework for interpreting how measured baseline differences contribute to trial–real-world outcome differences when individual-level real-world data are available but trial information is limited to aggregate baseline summaries.

Methods: The framework was applied to human epidermal growth factor receptor 2–positive breast cancer by comparing a Helsinki University Hospital real-world cohort of 267 patients with three randomized neoadjuvant therapy trial arms. Crude pCR rates and crude differences were estimated using Bayesian binomial models. Population-aligned pCR estimates were obtained using matching-adjusted indirect comparison (MAIC) and Bayesian multilevel regression with post-stratification and raking. Shapley decomposition was used to identify which measured variables contributed most to the adjustment differences.

Results: Across comparisons, baseline alignment shifted the real-world estimates downward, indicating less favorable measured baseline profiles in the trial populations with respect to pathological complete response. Adjustment differences were modest in the NeoSphere comparison and larger in the KRISTINE and HELEN-006 comparisons, where positive residual differences were observed after alignment.

Conclusions: The application illustrates how crude trial–real-world comparisons can obscure the contribution of measured baseline composition. Population alignment and decomposition provide a practical framework for interpreting such comparisons under aggregate trial reporting.

Keywords: Real-world evidence; population alignment; matching-adjusted indirect comparison; Bayesian multilevel regression; Shapley decomposition; clinical trials

Abbreviations: ASCO, American Society of Clinical Oncology; ECOG, Eastern Cooperative Oncology Group; ER, estrogen receptor; ESMO, European Society for Medical Oncology; HDI, highest density interval; HER2, human epidermal growth factor receptor 2;

HEOR, health economics and outcomes research; HR, hormone receptor; HUS, Helsinki University Hospital; IHC, immunohistochemistry; ISH, in situ hybridization; MAIC, matching-adjusted indirect comparison; MRI, magnetic resonance imaging; MRP, multilevel regression with post-stratification; NAT, neoadjuvant therapy; pCR, pathological complete response; PR, progesterone receptor; RCTs, randomized clinical trials; RW, real-world; RWD, real-world data; RWE, real-world evidence; TCHP, docetaxel, carboplatin, pertuzumab, and trastuzumab; THP, docetaxel, pertuzumab, and trastuzumab

1 Background

Real-world evidence (RWE) has become a valuable and widely used source of evidence for evaluating healthcare outcomes. In health economics and outcomes research (HEOR) and health technology assessment (HTA), randomized clinical trial (RCT) evidence and RWE are increasingly interpreted together to support decisions about clinical value, reimbursement, and implementation in routine care. Although RCTs remain the main method for assessing the effects of new therapies, RWE can provide valuable complementary information on outcomes in clinical practice. Without further assumptions, RCT results have external validity only for populations similar to those studied. In practice, RCT populations often differ from real-world (RW) populations not only because of eligibility criteria but also because of selection processes and participation patterns that influence who enrolls in trials[1–3]. In addition to differences between trial and target populations, treatment regimens, dosages, and other factors, such as treatment adherence, may also vary from those of clinical trials, affecting outcomes[4,5].

Crude comparisons between RCT and RW outcomes can be difficult to interpret. Similar crude outcome rates may mask differences in measured baseline composition, whereas crude outcome differences may partly reflect measured case-mix differences rather than differences in treatment performance or implementation. This is a practical challenge in HEOR and

HTA settings, where decision makers often need to interpret trial evidence alongside RW outcomes despite having access only to published aggregate trial summaries. A useful first step is therefore to estimate how measured baseline composition changes the RW outcome estimate and to examine the residual difference not accounted for by measured alignment. These challenges are particularly relevant in oncology, where novel systemic therapies may receive new indications based on limited randomized evidence, sometimes from a single RCT[6], while routine-practice populations and treatment pathways may differ from those studied in trials.

HER2-positive early breast cancer treated with preoperative neoadjuvant therapy (NAT) provides a useful setting for this approach, because treatment response is measured using pathological complete response (pCR), defined as no residual invasive tumor in breast or lymph nodes at surgery, a clinically meaningful short-term endpoint associated with disease-free and overall survival[7–9]. Key prognostic and treatment-guiding characteristics, including hormone receptor status, HER2 expression, Ki-67 proliferation rate, histology, grade, and disease stage, are routinely reported in trials and real-world cohorts[10].

Previous retrospective studies have reported real-world pCR rates for HER2-positive breast cancer treated with pertuzumab and trastuzumab-based NAT, typically ranging from approximately 46% to 66%[11–16]. These studies suggest broad consistency between routine-practice outcomes and randomized trials, but crude comparisons do not clarify how measured baseline composition contributes to apparent similarities or differences.

The aim of this study is to propose a population-alignment and decomposition framework for interpreting trial–RW outcome differences when individual-level RWD are available but trial information is limited to published aggregate baseline summaries. We apply the framework to RWD from the Helsinki University Hospital (HUS) HER2-positive NAT cohort and compare pCR outcomes with three pertuzumab and trastuzumab–based RCT arms: NeoSphere B, KRISTINE B, and HELEN-006 B[17–19]. For each comparison, crude pCR differences between the HUS and trial cohorts are first estimated. Adjusted pCR rates are then

constructed by aligning RW baseline covariate distributions to published trial summaries while preserving outcome–covariate relationships estimated from the RWD. We implement the framework using matching-adjusted indirect comparison (MAIC) and Bayesian multilevel regression with post-stratification (MRP) with raking. Shapley decomposition is used to identify which measured variables contribute most to the adjustment differences. The framework is intended for descriptive evidence interpretation rather than causal effect estimation.

2 Methods

2.1 Modeling approach

Our approach proceeds in three stages: estimation of crude outcome rates, baseline population alignment, and decomposition of the baseline-adjustment difference. In the first stage, crude pCR rates are estimated separately for the RWD population, $\hat{\theta}_{\mathrm{RWD}}$, and for the RCT population, $\hat{\theta}_{\mathrm{RCT}}$. The observed difference in pCR rates is defined as

$$\Delta\hat{\theta}_{\mathrm{obs}} = \hat{\theta}_{\mathrm{RCT}} - \hat{\theta}_{\mathrm{RWD}}.$$

Uncertainty in these estimates was propagated through the subsequent stages of the analysis; details of the Bayesian binomial models used for crude pCR estimation are provided in Supplementary Methods S3 (Additional file 1).

In the second stage, the pCR rate is estimated for a population-aligned version of the RW cohort, constructed by aligning the marginal distribution of baseline characteristics to that of the target RCT population while preserving the dependence between pCR and baseline variables observed in the row-level RWD. The baseline-adjustment difference captured by this alignment is defined as

$$\Delta\hat{\theta}_{\mathrm{adj}} = \hat{\theta}_{\mathrm{RWD}\to\mathrm{RCT}} - \hat{\theta}_{\mathrm{RWD}}, \tag{1}$$

where $\hat{\theta}_{\mathrm{RWD}\to\mathrm{RCT}}$ denotes the pCR estimate obtained after modeling the RW cohort to match the RCT baseline margins.

The residual difference is then defined as

$$\Delta\hat{\theta}_{\mathrm{res}} = \hat{\theta}_{\mathrm{RCT}} - \hat{\theta}_{\mathrm{RWD}\to\mathrm{RCT}},$$

which reflects discrepancies not accounted for by baseline covariate alignment and may include contributions from unmeasured patient characteristics and differences in treatment implementation.

We implement population alignment using MAIC and Bayesian MRP with raking.

2.2 Matching-adjusted indirect comparison (MAIC)

MAIC reweights individual-level data so that selected baseline covariate distributions match target marginals[20]. Since its introduction in 2010, MAIC has been widely adopted in oncology comparative effectiveness studies[20,21]. The weights are obtained as the solution to a Kullback–Leibler divergence minimization problem subject to moment-matching constraints.

MAIC implementation details and diagnostics are provided in Supplementary Methods S4. Weight diagnostics are shown in Supplementary Figures 1–3.

2.3 Bayesian multilevel regression with post-stratification and raking

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. The outcome model was specified as a hierarchical logistic regression with additive covariate effects and partially pooled level-specific coefficients. The raking component modeled the trial-aligned joint covariate distribution as a regularized, structured deviation from the observed RW contingency-table distribution, favoring distributions close to the observed RW

distribution while using a Poisson likelihood to align marginal counts with published RCT baseline summaries. Finally, post-stratification combined the estimated cell-specific pCR probabilities with the trial-aligned cell proportions to obtain the adjusted RW pCR rate.

The approach builds on established components of multilevel regression, post-stratification, and marginal alignment, but is tailored to settings in which individual-level data are available for the RW cohort while trial information is limited to aggregate baseline summaries. Full model specification, notation, model checks, sensitivity analyses, and the graphical representation of the model are provided in Supplementary Methods S5 and Supplementary Figure 7.

2.4 Comparison of MAIC and Bayesian MRP with raking

In our approach, both MAIC and Bayesian MRP with raking are applied to assess the robustness of population-aligned estimates under different modeling assumptions. Because the adjusted pCR rate is not directly observed, its estimation depends on assumptions about the outcome–covariate relationships in the RWD and about how baseline distributions are aligned to the trial summaries. Applying complementary methods therefore provides an important sensitivity check.

MAIC is a widely used frequentist approach that reweights individual-level data to match target marginal covariate distributions, making minimal assumptions about the outcome model and facilitating comparison with existing oncology literature. However, because MAIC relies entirely on reweighting observed individuals, it can suffer from reduced effective sample size and instability when covariate overlap between populations is limited.

Bayesian MRP with raking introduces additional structure through an explicit outcome model and a probabilistic formulation of the raking constraints. This requires modeling assumptions but allows more efficient use of row-level data, partial pooling across covariate levels, and extrapolation beyond observed covariate combinations. Uncertainty arising from both outcome modeling and population alignment is propagated jointly within a coherent

posterior framework. For the subsequent variable-level decomposition, Shapley values were calculated from posterior draws of the Bayesian MRP with raking model, allowing posterior uncertainty to be propagated to the estimated contributions.

Taken together, MAIC serves as a low-assumption, field-standard benchmark, while Bayesian MRP with raking provides a flexible and efficient modeling framework. Agreement between the two approaches increases confidence in the findings, whereas discrepancies would highlight sensitivity to modeling assumptions.

2.5 Shapley decomposition

The difference between the adjusted and observed pCR rates in Equation (1) can be further decomposed into variable-level contributions, including a component corresponding to missing values in the RWD dataset. Shapley decomposition was used to identify which measured variables contributed most to the adjustment difference. In this context, a Shapley contribution represents the average change in the adjusted pCR estimate associated with adding a variable across possible subsets of the other alignment variables. Full computational details are provided in Supplementary Methods S6.

2.6 Application to HER2-positive breast cancer

The HUS cohort was compared with three randomized trial arms involving pertuzumab and trastuzumab–based neoadjuvant regimens. NeoSphere B included docetaxel, pertuzumab, and trastuzumab (THP), whereas KRISTINE B and HELEN-006 B included docetaxel, carboplatin, pertuzumab, and trastuzumab (TCHP)[17–19]. The NeoSphere THP arm was selected because THP has constituted the backbone of NAT for HER2-positive breast cancer[22]. The TCHP arms of KRISTINE and HELEN-006 were selected to represent more contemporary trial comparators. Current American Society of Clinical Oncology (ASCO) guidelines recommend carboplatin as a preferred companion to pertuzumab, trastuzumab, and docetaxel, while European Society for Medical Oncology (ESMO) guidelines list it as

an optional component[22,23].

2.6.1 Data sources and extraction

The study cohort comprised breast cancer patients treated at HUS who received at least one cycle of NAT with pertuzumab and trastuzumab. Patients who did not receive pertuzumab and trastuzumab during the first two cycles were excluded because their treatment did not represent current clinical guidelines. Other exclusion criteria were metastatic diagnosis (ICD-10 C78–C79) before the start of NAT and drug administrations outside HUS. For each patient, the treatment period began at diagnostic needle biopsy and ended at surgical resection, documented cancer progression when surgery was not performed, or death. Patients were included if both the biopsy and endpoint occurred between January 1, 2019, and September 1, 2024. In total, the final HUS cohort comprised 268 cancers from 267 patients, including one patient with bilateral HER2-positive breast cancer. Clinical data were collected until September 1, 2024.

Data collection was conducted in two stages. Detailed extraction logic and cohort-construction details are provided in Supplementary Methods S2, and cohort construction is illustrated in Supplementary Figure 8.

2.6.2 Variables

Baseline characteristics of the HUS cohort are summarized in Table 1. For privacy protection, cell counts between 1 and 2 were masked and reported as "< 3". Patient age was defined at the time of diagnostic core needle biopsy. Pre-treatment pathological information was primarily derived from breast biopsy specimens; when unavailable, corresponding lymph node biopsy results were used. Bilaterality was determined based on pathological diagnosis.

Additional variable definitions and manual extraction details are provided in Supplementary Methods S2.

Less than three of the 267 patients in the HUS cohort were male. Baseline Eastern Coop-

erative Oncology Group (ECOG) performance status was missing for 51% of the patients and was therefore not considered comparable with the RCT populations. Clinical TNM stage at baseline was not explicitly recorded for HUS patients and could not be reliably derived.

2.6.3 Treatment regimens and treatment completion

NAT regimens used in the HUS cohort and their frequencies are summarized in Supplementary Table 11, with detailed regimen descriptions in Supplementary Table 12. Details of HUS regimen evolution, treatment-completion definitions, and drug substitutions are provided in Supplementary Methods S7. 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; under a relaxed NeoSphere-aligned definition, this proportion increased to 90%.

2.6.4 Target trials and extracted aggregate data

Baseline covariate marginals for each RCT arm were extracted from published baseline-characteristics tables and eligibility criteria[17–19]. Alignment was restricted to variables available in the HUS cohort and harmonizable with published trial categories. The variables and RCT marginal distributions used for alignment are reported in Supplementary Tables 1–4.

HUS variables were harmonized to trial definitions using rule-based recoding, including discretization of continuous measurements and grouping of histology and receptor-status categories. Harmonization rules are described in Supplementary Methods S1, and the implementation is available in the code repository[24].

Use of generative AI tools

Large language models (ChatGPT, OpenAI) were used for language editing and to assist with code development, including drafting helper functions and troubleshooting implementation. All study design, modeling decisions, final code, and interpretation of results were determined

by the authors.

3 Results

We evaluated trial–real-world outcome comparisons by estimating crude HUS pCR rates and adjusted HUS pCR rates after baseline covariate alignment to each RCT arm. We first summarize the main associations from the multilevel outcome models used in the Bayesian MRP with raking analyses. We then present crude comparisons, population-aligned comparisons, Shapley decompositions of the adjustment differences, and a sensitivity analysis evaluating whether NAT discontinuation explained residual differences. Additional diagnostics, extended results, and reproducibility materials are provided in the digital repository[24].

3.1 Results of the outcome models

Six multilevel logistic models were fitted across the three comparisons, with and without treatment-discontinuation indicators. These models estimated associations between baseline characteristics and pCR in the HUS cohort, and were then used to construct population-aligned HUS estimates. Outcome model specifications are detailed in Subsection 2.3. Full coefficient summaries are reported in Supplementary Tables 5–10.

Across specifications, several baseline variables showed consistent associations with pCR. Hormone receptor-positive cases had lower probability of achieving pCR compared to hormone receptor-negative cases. Higher HER2 expression was associated with higher pCR rates in the HUS cohort.

3.2 Crude comparisons

Among the three RCT arms, only NeoSphere showed a lower pCR rate than the HUS cohort, with posterior probability of 94%. Unadjusted pCR rate in the HUS cohort had mean 0.54 with 95% highest density interval (HDI) 0.49–0.60. In comparison, posterior mean of the

pCR rate in NeoSphere was 0.46 with 95% HDI 0.37–0.55.

The pCR rate in KRISTINE was similar to HUS (mean 0.56, 95% HDI 0.50–0.63). The pCR rate in HELEN-006 was slightly higher than in HUS (mean 0.57, 95% HDI 0.52–0.63), with a posterior probability of 0.77 that the HELEN-006 rate exceeded that of HUS.

3.3 Population-aligned comparisons

Across trials, baseline alignment consistently shifted the HUS estimates downward, reflecting differences in measured baseline characteristics. The residual difference, defined as the difference between the RCT estimate and the baseline-aligned HUS estimate, showed different patterns across trials. Posterior summaries of pCR rates for the RCT arms, crude HUS cohort, and adjusted HUS cohort across all three trial comparisons are shown in Figure 1.

In the NeoSphere comparison, after aligning baseline distributions, the adjusted HUS pCR rate under Bayesian MRP with raking was 0.51 (95% HDI 0.42–0.59). This is closer to the NeoSphere estimate than the unadjusted HUS pCR rate. The adjustment difference was modest with mean -0.03, and the residual difference was centered near zero with substantial uncertainty. MAIC estimates were broadly compatible, albeit uncertain. The bootstrap MAIC results are reported in Supplementary Table 13.

Larger adjustment differences were observed in the KRISTINE and HELEN-006 comparisons. In the KRISTINE comparison, the adjusted HUS pCR rate had mean 0.42 (95% HDI 0.33–0.51), corresponding to an adjustment difference of -0.12.

In the HELEN-006 comparison, the adjusted HUS pCR rate had mean 0.48 (95% HDI 0.40–0.56), with an adjustment difference of -0.07. MAIC yielded similar estimates in both KRISTINE and HELEN-006 comparisons.

The residual differences were positive in both the KRISTINE comparison with mean 0.13 (95% HDI 0.02–0.24) and the HELEN-006 comparison with mean 0.10 (95% HDI 0.03–0.19).

3.4 Shapley decompositions

To examine which baseline characteristics contributed to the adjustment difference, we decomposed it into variable-level contributions using the Shapley framework. Trial-specific decompositions are shown in Figure 2.

In the NeoSphere comparison, the overall adjustment difference was small and uncertain, and this uncertainty was reflected in the Shapley decomposition. Most variable-level contributions had wide 95% HDIs relative to their means. The only variable showing a consistently negative contribution, with mean contribution -0.03 (95% HDI -0.06 to 0.00), was hormone receptor (HR) status, reflecting the higher proportion of HR-positive patients in NeoSphere (47%) compared to HUS (33%).

In the KRISTINE comparison, the higher prevalence of HR-positive disease (62% vs. 33% in HUS) was the main contributor to the adjustment difference. Some of this contribution may overlap with the nodal-biopsy exclusion criterion in the KRISTINE study, as in HUS data, patients excluded from the adjusted comparisons due to nodal-positive biopsy had a lower proportion of HR-positive tumors than those retained.

In the HELEN-006 comparison, HR status was again the dominant contributor (mean -0.09, 95% HDI -0.13 to -0.05), accounting for a substantial portion of the total absolute contributions. Pre-treatment histology contributed positively (mean 0.04, 95% HDI 0.02–0.06), largely reflecting the higher proportion of non-ductal histological subtypes (lobular, micropapillary, and apocrine) in the HUS cohort. Differences in HER2 immunohistochemistry/in situ hybridization (IHC/ISH) distribution contributed modestly (mean -0.03, 95% HDI -0.05 to -0.01): see the IHC/ISH groups in Table 1.

3.5 Sensitivity to NAT discontinuation

Because incomplete NAT could plausibly influence pCR rates, we examined whether higher discontinuation in the HUS cohort contributed to residual differences by adding discontinuation as a baseline indicator to the alignment model.

Including discontinuation had negligible impact: posterior means of residual differences changed by less than one percentage point in all comparisons, suggesting that treatment completion alone is unlikely to explain them.

4 Discussion

This study applied a population-alignment and decomposition framework to compare pCR outcomes between a real-world HER2-positive breast cancer NAT cohort and three randomized trial arms. The application shows how crude trial–RW pCR comparisons can obscure differences in measured baseline composition.

Across all comparisons, baseline alignment shifted the RW estimates downward, indicating that the RCT populations had less favorable measured baseline profiles with respect to pCR than the HUS cohort. In KRISTINE and HELEN-006, crude outcomes were comparable between cohorts, yet baseline alignment revealed sizeable negative adjustment differences and positive residual differences. The NeoSphere comparison followed a different and more uncertain pattern, with alignment narrowing the negative observed difference.

The large positive residual differences in the KRISTINE and HELEN-006 comparisons were not meaningfully reduced by adjusting for the higher NAT discontinuation proportion in HUS. This suggests that the measured variables available for alignment were insufficient to account for the difference between the adjusted HUS estimates and the corresponding RCT estimates.

One possible hypothesis could be that the inclusion of carboplatin in the TCHP regimens used in KRISTINE and HELEN-006 contributed to the residual differences. However, the randomized phase III neoCARHP trial reported only a modest difference in pCR rates between THP and TCHP arms (-1.84 percentage points with 95% confidence interval between -8.5 and 5.0)[25], suggesting that differences in chemotherapy backbone alone are unlikely to account for the magnitude of the residual differences observed here.

The variables available for alignment do not capture all baseline factors that may be relevant when interpreting trial–RW outcome differences. In particular, disease severity can be aligned only partially using the overlapping variables available across datasets. For example, detailed staging information comparable to that reported in NeoSphere (Supplementary Table 1) is not available in the same form for the other cohorts. Differences in the prevalence of advanced T or N stages across trials could therefore contribute to the residual differences.

The residual differences may also reflect unmeasured aspects of baseline health status, such as performance status or comorbidity burden, which were not available for alignment. One possible interpretation is that participants in the KRISTINE and HELEN-006 trials were, on average, clinically fitter than patients in the HUS cohort. This interpretation more plausibly reflects differences in baseline health status or other contextual factors than differences in chemotherapy backbone alone.

The possibility that RCTs over-represent patients with more severe disease profiles but relatively favorable baseline health status is supported by the literature. Clinical trials may apply enrichment strategies designed to increase the likelihood of detecting treatment effects, which can lead to preferential enrollment of higher-risk disease populations[26,27]. In line with this, the Shapley decomposition suggests that the higher prevalence of hormone receptor–positive disease in the RCT cohorts—particularly in KRISTINE and HELEN-006—contributes to the negative adjustment differences and is consistent with RCT populations having, on average, less favorable baseline characteristics with respect to treatment response.

At the same time, differences in baseline health and functional status between trial participants and RW patients, and their implications for external validity, have been widely discussed[28–30]. These considerations provide a plausible context for the patterns observed here, while also underscoring the inherent challenges of interpreting trial evidence alongside real-world outcomes.

Methodologically, this study addresses a common HEOR constraint: interpreting trial

evidence alongside RW outcomes when individual-level trial data are unavailable. By combining MAIC, Bayesian MRP with raking, and Shapley decomposition, the framework estimates adjusted HUS pCR rates, assesses sensitivity to modeling assumptions, and identifies measured baseline variables contributing to adjustment differences. This provides a flexible approach for interpreting trial–RW outcome differences under aggregate trial reporting.

A key limitation of this analysis is the lack of detailed data on patients' baseline health status. Although marginal Eastern Cooperative Oncology Group (ECOG) distributions were available for NeoSphere and KRISTINE, ECOG is coarse and shows limited variation in the relatively fit NAT population, reducing its informativeness. Moreover, ECOG was missing for over 50% of the HUS cohort (see Subsection 2.6.2), and thus it was not included in the alignment models. Additional data limitations arise from the use of broad variable classes in the RCTs, which, after harmonization with the HUS data, further constrained detailed predictive performance. Apart from the relatively coarse analysis of NAT discontinuation, treatment details were outside the scope of this study. For instance, information on cumulative dosages could help explain differences between treatment settings.

5 Conclusions

Taken together, these findings show that crude comparisons between trial and RW outcomes may obscure important differences in baseline population structure. Population alignment and decomposition provide a practical framework for making such comparisons more informative when only aggregate trial summaries are available. This is increasingly relevant in HEOR and oncology, where RWE is often interpreted alongside limited randomized evidence for novel systemic therapies. Beyond the present application, the proposed framework may also serve an exploratory and hypothesis-generating role by identifying measured factors that contribute to trial–RW outcome differences. The same general approach could be extended to time-to-event outcomes.

Declarations

Funding

This work was conducted as part of the ONCOVALUE project, which has received funding from the European Union's Horizon Europe Research and Innovation Programme under grant agreement No. 101095245. This work was also supported by the Research Council of Finland via the Finnish Centre of Excellence in Randomness and Structures.

The funders had no role in the preparation, review, or approval of the manuscript or in the decision to submit the manuscript for publication. Publication of the study results was not contingent on funder or sponsor approval or censorship. The views expressed are those of the authors and do not necessarily reflect those of the European Union or the granting authority. Neither the European Union nor the granting authority can be held responsible for them.

Authors' contributions

J. Lähteenmaa conceived the study, developed the methodology, performed the statistical analyses, curated the data, and drafted the manuscript. P.I. supervised the methodological development, contributed to study design, and critically revised the manuscript. P.K. contributed to clinical study design, interpretation of clinical findings, and manuscript revision. J.M. provided clinical supervision and contributed to study design and interpretation. E.-N.H. contributed to data collection. J. Laaksolilja contributed to data acquisition and data processing. A.S. provided pathological expertise and manuscript review. K.S. provided clinical expertise and manuscript review. All authors reviewed and approved the final manuscript. P.I., P.K., and J.M. jointly supervised the study.

Acknowledgments

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Software and reproducibility

All analyses were conducted in Python (version 3.10.18). Key packages include NumPy (v1.26.4), pandas (v2.3.1), PyMC (v5.23.0), ArviZ (v0.21.0), PyTensor (v2.31.7), and scikit-learn (v1.7.0). Random number generation was controlled using a fixed seed via NumPy's `default_rng`.

Ethics approval and consent to participate

This retrospective registry-based study was conducted using existing patient data in accordance with the Finnish Act on the Secondary Use of Health and Social Data (552/2019) and the General Data Protection Regulation (EU 2016/679). The research and data permit (HUS/44/2021) was granted by Helsinki University Hospital (HUS).

The study was not submitted to an ethics committee because prior ethics committee review was not required under the applicable Finnish regulations for this type of retrospective registry-based research. No prospective recruitment or research interventions involving patients were undertaken.

Individual informed consent to participate was not obtained, as consent was not required for the secondary use of patient registry data under the Finnish Act on the Secondary Use of Health and Social Data (552/2019). The study was conducted in accordance with the applicable institutional and legal requirements for the protection of patient data.

Clinical trial registration

Not applicable. This study is a retrospective observational and methodological analysis of existing clinical data and published aggregate results from previously conducted randomized clinical trials. No participants were prospectively assigned to interventions as part of this study.

Patient engagement

Patients were not directly engaged in the design, conduct, reporting, or dissemination of this research.

Availability of data and materials

The patient-level data used in this study are subject to institutional and legal restrictions and are not publicly available. All analyses were conducted within the secure HUS Acamedic computing environment, and no personally identifiable information was exported. Aggregated data are available in the public repository[24]. Requests for access to the underlying data may be directed to the corresponding author and will be considered subject to institutional and legal approval.

All analysis code used for data harmonization, model fitting, population-aligned estimation, and figure generation is available in the public repository[24]. The repository also contains extended diagnostic outputs and reproducibility materials generated during the analysis.

Supplementary information

Additional file 1. Supplementary methods, tables, and figures (PDF). Contains additional methodological details, data harmonization procedures, model specifications, diagnostics, sensitivity analyses, and supporting results.

References

[1] Hall PS. Real-world data for efficient health technology assessment. European Journal of Cancer. 2017 Jul;79:235-7.

[2] Slattery J, Kurz X. Assessing strength of evidence for regulatory decision making in licensing: What proof do we need for observational studies of effectiveness? Pharmacoepidemiology and Drug Safety. 2020 Oct;29:1336-40.

[3] Yang W, Zilov A, Soewondo P, Bech OM, Sekkal F, Home PD. Observational studies: going beyond the boundaries of randomized controlled trials. Diabetes Research and Clinical Practice. 2010;88:S3-9.

[4] Nordon C, Karcher H, Groenwold RH, Ankarfeldt MZ, Pichler F, Chevrou-Severac H, et al. The "efficacy-effectiveness gap": historical background and current conceptualization. Value in Health. 2016;19(1):75-81.

[5] Sargent D. What constitutes reasonable evidence of efficacy and effectiveness to guide oncology treatment decisions? The Oncologist. 2010;15(Suppl 1):19-23.

[6] Wilson BE, Booth CM. Real-world data: bridging the gap between clinical trials and practice. EClinicalMedicine. 2024 Dec;78:102915.

[7] Broglio KR, Quintana M, Foster M, Olinger M, McGlothlin A, Berry SM, et al. Association of Pathologic Complete Response to Neoadjuvant Therapy in HER2-Positive Breast Cancer With Long-Term Outcomes: A Meta-Analysis. JAMA Oncology. 2016 Jun;2(6):751-60. Available from: https://doi.org/10.1001/jamaoncol.2015.6113.

[8] Antonini M, Mattar A, Pereira TM, Oliveira LL, Teixeira MD, Amorim AG, et al. Pathologic complete response and breast cancer survival post-neoadjuvant chemotherapy: A systematic review and meta-analysis of real-world data. Heliyon. 2025;11(10):e43069.

[9] Geyer CEJ, Untch M, Huang CS, Mano MS, Mamounas EP, Wolmark N, et al. Survival with Trastuzumab Emtansine in Residual HER2-Positive Breast Cancer. The New England Journal of Medicine. 2025 Jan;392:249-57.

[10] Loibl S, Poortmans P, Morrow M, Denkert C, Curigliano G. Breast cancer. The Lancet. 2021;397(10286):1750-69.

[11] González-Santiago S, Saura C, Ciruelos E, Alonso JL, de la Morena P, Santisteban Eslava M, et al. Real-world effectiveness of dual HER2 blockade with pertuzumab and trastuzumab for neoadjuvant treatment of HER2-positive early breast cancer (The NEOPETRA Study). Breast Cancer Research and Treatment. 2020;184(2):469-79.

[12] Zhou M, Wang S, Wan N, Yuan S, Hu X, Zhou W, et al. Efficacy and safety of neoadjuvant pertuzumab plus trastuzumab in combination with chemotherapy regimen in Chinese patients with HER2-positive early breast cancer: a real-world retrospective multi-center cohort study. Annals of Translational Medicine. 2022 Dec;10:1387.

[13] Kim JY, Nam SJ, Lee JE, Yu J, Chae BJ, Lee SK, et al. Real World Evidence of Neoadjuvant Docetaxel/Carboplatin/Trastuzumab/Pertuzumab (TCHP) in Patients with HER2-Positive Early or Locally Advanced Breast Cancer: A Single-Institutional Clinical Experience. Cancer Research and Treatment. 2022 Jan;54(4):1091-8.

[14] Falcón González A, Cruz Jurado J, Llabrés Valenti E, Urbano Cubero R, Álamo de la Gala MC, Martínez Guisado MA, et al. Real-world experience with pertuzumab and trastuzumab combined with chemotherapy in neoadjuvant treatment for patients with early-stage HER2-positive breast cancer: the NEOPERSUR study. Clinical and Translational Oncology. 2024;26(9):2217-26.

[15] Canino F, Barbolini M, De Giorgi U, Fontana T, Gaspari V, Gianni C, et al. Safety and efficacy analysis of neoadjuvant pertuzumab, trastuzumab and standard chemotherapy

for HER2-positive early breast cancer: real-world data from NeoPowER study. BMC Cancer. 2024;24(1):735.

[16] Hall BJ, Bhojwani AA, Wong H, Law A, Flint H, Ahmed E, et al. Neoadjuvant Trastuzumab and Pertuzumab for Early HER2-Positive Breast Cancer: A Real-World Experience. The Breast Journal. 2022 Jan;2022(1):7146172.

[17] Gianni L, Pienkowski T, Im YH, Roman L, Tseng LM, Liu MC, et al. Efficacy and safety of neoadjuvant pertuzumab and trastuzumab in women with locally advanced, inflammatory, or early HER2-positive breast cancer (NeoSphere): a randomised multicentre, open-label, phase 2 trial. The Lancet Oncology. 2012;13(1):25-32.

[18] Hurvitz SA, Martin M, Symmans WF, Jung KH, Huang CS, Thompson AM, et al. Neoadjuvant trastuzumab, pertuzumab, and chemotherapy versus trastuzumab emtansine plus pertuzumab in patients with HER2-positive breast cancer (KRISTINE): a randomised, open-label, multicentre, phase 3 trial. The Lancet Oncology. 2018;19(1):115-26.

[19] Chen XC, Jiao DC, Qiao JH, Wang CZ, Sun XF, Lu ZD, et al. De-escalated neoadjuvant weekly nab-paclitaxel with trastuzumab and pertuzumab versus docetaxel, carboplatin, trastuzumab, and pertuzumab in patients with HER2-positive early breast cancer (HELEN-006): a multicentre, randomised, phase 3 trial. The Lancet Oncology. 2025 Jan;26:27-36.

[20] Signorovitch JE, Wu EQ, Yu AP, Gerrits CM, Kantor E, Bao Y, et al. Comparative Effectiveness Without Head-to-Head Trials. PharmacoEconomics. 2010;28(10):935-45.

[21] Farinasso CM, Ferreira VL, Medeiros FC, da Rocha AP, do Carmo Silva Parreira P, Oliveira LA, et al. Matching-Adjusted Indirect Comparison Studies in Oncology: A Scoping Review Focused on Reporting Quality. Value in Health Regional Issues. 2025;47:101088.

[22] Loibl S, André F, Bachelot T, Barrios CH, Bergh J, Burstein HJ, et al. Early breast cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. Annals of Oncology. 2024 Feb;35(2):159-82.

[23] Korde LA, Somerfield MR, Carey LA, Crews JR, Denduluri N, Hwang ES, et al. Neoadjuvant Chemotherapy, Endocrine Therapy, and Targeted Therapy for Breast Cancer: ASCO Guideline. Journal of Clinical Oncology. 2021;39(13):1485-505.

[24] Lähteenmaa J, Ilmonen P, Hännikäinen EN, Laaksolilja J, Ståhls A, Santti K, Karihtala P, Mattson J. Moving Beyond Crude Trial–Real-World Comparisons Through Population Alignment and Decomposition; 2026. GitHub repository. `https://github.com/juholahteenmaahus/trial-rw-population-alignment`.

[25] Gao HF, Li W, Wu Z, Dong J, Cao Y, Zhao Y, et al. De-escalated neoadjuvant taxane plus trastuzumab and pertuzumab with or without carboplatin in HER2-positive early breast cancer (neoCARHP): A multicentre, open-label, randomised, phase 3 trial. Journal of Clinical Oncology. 2025;43(17 Suppl):LBA500.

[26] U S Food and Drug Administration. Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products. Silver Spring, MD: U.S. Food and Drug Administration; 2019. Guidance for Industry; Accessed: 2026-02-11.

[27] U S Food and Drug Administration. Pathological Complete Response in Neoadjuvant Treatment of High-Risk Early-Stage Breast Cancer: Use as an Endpoint to Support Accelerated Approval. Silver Spring, MD: U.S. Food and Drug Administration; 2020. FDA Guidance for Industry. Accessed: 2026-02-11.

[28] Hutchins LF, Unger JM, Crowley JJ, Coltman CAJ, Albain KS. Underrepresentation of patients 65 years of age or older in cancer-treatment trials. The New England Journal of Medicine. 1999 Dec;341:2061-7.

[29] Rothwell PM. External validity of randomised controlled trials: to whom do the results of this trial apply? The Lancet. 2005;365:82-93.

[30] Unger JM, Hershman DL, Fleury ME, Vaidya R. Association of Patient Comorbid Conditions With Cancer Clinical Trial Participation. JAMA Oncology. 2019 Mar;5(3):326-33.

Table 1: Baseline characteristics of the HUS cohort.

Characteristic	**N**	**Summary**
Median age (n = 268)	—	52 (42–63)
Tumor size, median (mm)	221	26.0 (20–37)
Tumor size $\leq$ 20 mm	58	22% [17–27%]
Tumor size 20–50 mm	143	53% [47–60%]
Tumor size $>$ 50 mm	19	7% [4–11%]
Tumor size: unknown	48	18%
Nodal status: positive	155	58% [52–64%]
Nodal status: negative	93	35% [29–41%]
Nodal status: unknown	20	8% [5–11%]
Bilateral breast cancer	8	3% [1–6%]
Unilateral breast cancer	260	97% [94–99%]
Bilaterality: unknown	0	—
Histology: ductal	211	79% [73–84%]
Histology: lobular	13	5% [3–8%]
Histology: other	44	16% [12–21%]
Histology: unknown	0	—
Tumor grade 1	$<$ 3	$<$ 1% [0–2%]
Tumor grade 2	64	24% [19–29%]
Tumor grade 3	152	57% [51–63%]
Tumor grade: unknown	51	19% [15–24%]
ER and/or PR positive	88	33% [27–39%]
ER and PR negative	180	67% [61–73%]
ER or PR unknown	0	—
HER2 IHC/ISH: IHC 2+ and ISH+	71	27% [21–32%]
HER2 IHC/ISH: IHC 3+	183	68% [62–74%]
HER2 IHC/ISH: other	14	5% [3–9%]
HER2 IHC/ISH: unknown	0	—
Ki-67, median (n = 263)	—	40 (30–60)
Ki-67 $\leq$ 30%	91	34% [28–40%]
Ki-67 $>$ 30%	172	64% [58–70%]
Ki-67: unknown	5	2% [0–4%]

Notes: Continuous variables are reported as median (interquartile range). Categorical variables are reported as n (%) with 95% binomial confidence intervals. Counts between 1 and 2 are masked as "$<$ 3" to preserve patient confidentiality.

Abbreviations: ER, estrogen receptor; HER2, human epidermal growth factor receptor 2; IHC, immunohistochemistry; ISH, in situ hybridization; Ki-67, proliferation index; PR, progesterone receptor.

Figures

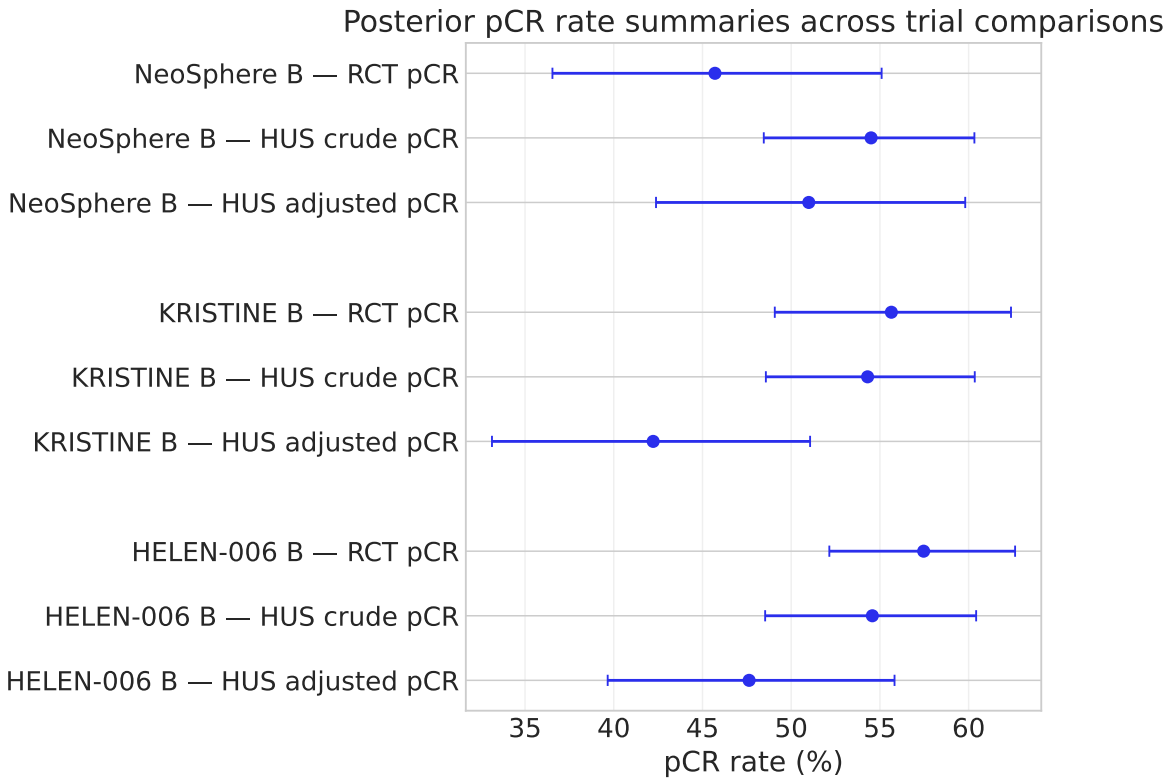


Figure 1: Posterior estimates of pathological complete response rates across trial comparisons. Points show posterior medians and horizontal intervals show 95% highest density intervals (HDIs) for the RCT arm, crude HUS cohort, and adjusted HUS cohort after baseline covariate alignment.

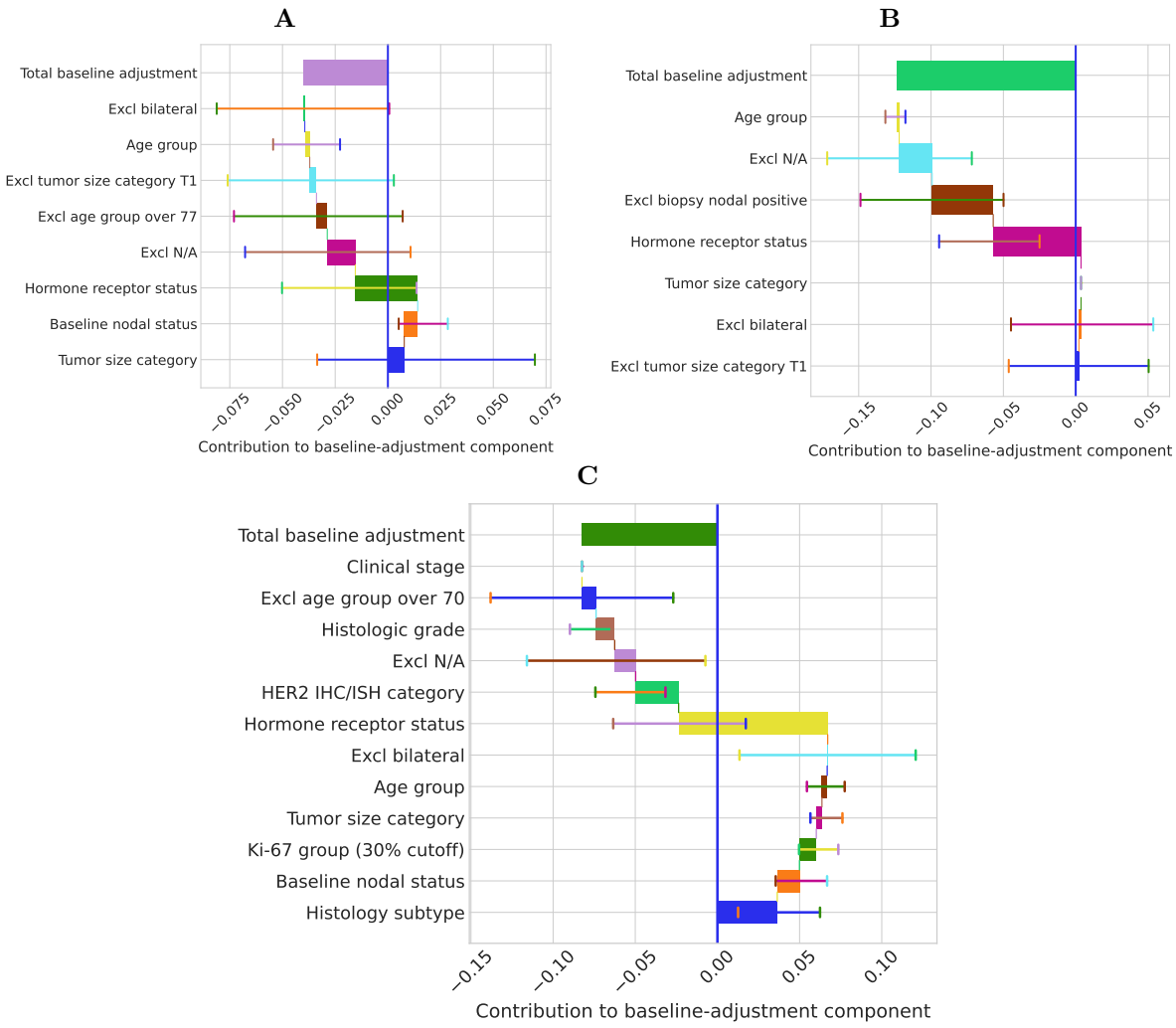


Figure 2: Shapley decomposition of adjustment differences across trial comparisons. Bars represent mean Shapley contributions and error bars show 95% HDIs. Mean contributions sum to the total adjustment difference for each comparison.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementaryinformationBMC.pdf