Paper RE05

External Comparator Studies in the Era of Real-World Data

Vishal Goriya, IQVIA, Thane, India
Anshul Sinha, IQVIA, Noida, India
Gerd Rippin, IQVIA, Frankfurt, Germany

ABSTRACT

External Comparator (EC) studies are increasingly recognized as an additional valuable study design type, particularly when randomization is infeasible or unethical. These studies typically leverage real-world data (RWD) from observational sources such as electronic health records, claims databases, and registries to construct an EC cohort that reflects routine clinical practice. Methodological rigor is essential to address potential biases of non-randomized data, like applying carefully suitable causal inference models (e.g., propensity score matching, inverse probability of treatment weighting (IPTW)) and missing data methods (e.g., multiple imputation). Other tasks besides causal inference and missing data handling include the assessment of data appropriateness (data being fit-for-purpose), baseline / endpoint data alignment, selection of marginal estimators, model checks, and performing many supplementary and sensitivity analyses (e.g., for unmeasured confounding).

INTRODUCTION

Randomized controlled trials (RCTs) are the gold standard for evaluating the efficacy and safety of medical interventions. However, in some contexts such as rare diseases, oncology, or urgent public health crises conducting an RCT may be infeasible, unethical, or impractical due to limited patient populations, ethical constraints, or logistical challenges. In these cases, EC studies may offer an alternative approach to contextualize trial findings and to perform formal statistical testing.

However, if an RCT is feasible, it should be performed, as EC studies come with additional methodological challenges. The absence of randomization introduces risks of confounding and biased estimations. For example, utilizing external data comes generally with potential data heterogeneity and baseline / endpoint misalignment. To address these issues, researchers have adopted rigorous frameworks, such as the ICH E9(R1) estimand guidance, the target trial emulation approach to improve the operational design process and interpretability of findings, and a wealth of other methodological advice (see references). Our aim is to summarize the literature to provide practical guidance for designing and analyzing EC studies.

METHODS

STUDY DESIGN FRAMEWORKS

Designing an observational EC study benefits from a structured approach that combines the ICH E9(R1) estimand framework with the target trial emulation framework (TTEF). The estimand framework specifies what treatment effect is being estimated by defining the 5 estimand attributes treatment condition, target population, endpoint, handling of intercurrent events, and the population-level summary measure. The TTEF, on the other hand, details how to emulate a hypothetical randomised trial using RWD, by outlining the 7 TTEF components eligibility criteria, treatment strategies, assignment procedures, follow-up period, outcomes, causal contrasts of interest, and the analysis plan.

There is considerable conceptual overlap between these frameworks (for example, the estimand attribute “population” aligns well with the TTEF’s “eligibility criteria”), but frameworks also complement each other well by bringing in distinct strengths. The estimand framework provides clarity on which treatment effect is estimated and excels in clarity how to handle post-baseline intercurrent events, whereas the TTEF has the strength of focusing on operationalizing trial design elements in an observational setting. A unifying approach of both frameworks consists according to recent research of nine core design elements encompassing both frameworks. These elements cover: (1) treatment conditions and strategies; (2) population; (3) endpoint specification and validation; (4) handling of intercurrent events; (5) population-level summary measure; (6) follow-up period; (7) baseline alignment (time zero definition); (8) assignment procedures; and (9) choice of the marginal estimator. By systematically addressing each of these elements, researchers can ensure that the EC study is conceptually aligned with a target trial and that the estimand of interest is clearly defined. This unified approach promotes transparency, reproducibility, and regulatory alignment, serving as a practical blueprint for designing EC studies that are both methodologically rigorous and fit for decision-making.

MARGINAL ESTIMATOR CHOICES

In EC studies, the Average Treatment Effect on the Treated (ATT) and the Average Treatment Effect (ATE) are often applied for defining a marginal estimator. The ATT represents the effect of the treatment in the population that actually received it (the trial's patient cohort), while the ATE represents a treatment effect as expected in the pooled sample.

For additional insight and to test robustness, alternative marginal estimators such as the Average Treatment Effect in the Untreated (ATU) and the Average Treatment Effect in the Overlap Population (ATO) can be applied as supplementary analyses. The ATU represents the treatment effect as expected in the EC, which may be of primary interest for Health Technology Assessment (HTA) submission, but may also serve as a supplementary perspective for regulatory decision-making, as the trial cohort is often criticized to constitute a too narrow patient selection with many exclusion criteria, which will not always be met in the real world application after approval. The ATO focuses on internal validity (providing a more like-for-like comparison), such that it is highly informative, especially when internal validity is questioned for other marginal estimators.

Table 1. Marginal Treatment Effect Estimators and Relevance for External Comparator Studies

Estimand	Definition	Relevance to ECs
ATT	Effect of the treatment in the population that received the treatment (trial participants).	Commonly applied. Focuses on trial-eligible patients and avoids extrapolation beyond the trial cohort.
ATE	Effect of the treatment in the pooled population (both treated and untreated).	Commonly applied. The focus on the pooled sample changes the estimand by incorporating also the RW component, which may lead to a more realistic treatment effect estimate for the treatment application after approval.
ATU	Effect of the treatment in the population that did not receive the treatment (comparator population).	A natural marginal estimator to estimate treatment effects for HTA submissions, but also informative as a supplementary analysis for regulatory context
ATO	Effect of the treatment in a population being closest to the one in a hypothetical RCT	An important marginal estimator when aiming to present an analysis with highest internal validity. Often used as a supplementary analysis to support the robustness of primary findings.

The choice of the primary marginal estimator should be aligned with regulatory / HTA stakeholders but providing the whole set of marginal estimators is usually most informative to allow for multiple perspectives.

DATA SOURCES AND CURATION

The EC is derived from one or more RWD sources that capture the relevant patient population, baseline characteristics, treatments, and outcomes with sufficient data quality. Common data sources include disease registries or electronic health record databases, but sometimes administrative claims data (which often have insufficient clinical data granularity for confounder adjustment) can be considered as well. The choice of data source depends on its relevance and reliability, especially its ability to mirror the trial's inclusion and exclusion criteria, the follow-up duration, and the availability of outcomes with acceptable completeness and accuracy.

Trial eligibility criteria are applied to the RWD as much as possible to define a comparable EC. The index date is often selected as the treatment initiation date to ensure comparability with RWD, but other solutions are possible as well. Baseline covariates are captured within a pre-specified window prior to the index date (e.g. 30 days, 3 months or longer, depending on the indication), and this time window can be made dependent on the type of data (e.g., laboratory, genetic markers).

Follow-up for the EC begins at the index date (e.g., treatment initiation) and is typically truncated to match the trial's observation period or follow-up schedule. Overall survival (OS) can usually be obtained from death records or by recorded clinical documentation. For composite endpoints like progression-free survival (PFS), revised definitions may be employed in both cohorts, for example, by defining progression as the first occurrence of either a documented progression event in the records, the start of a new systemic therapy, or death. If outcome definitions in the external data differ from those in the trial, validation steps (such as blinded adjudications) should be implemented to ensure that the endpoints are as comparable as possible.

The derived dataset undergoes thorough curation by applying standardized coding systems (for diagnoses, procedures, medications), resolving data inconsistencies (e.g., dates out of order), and assessing occurrence and impact of any missing data. Key baseline variables that are missing for some patients are handled according to a predefined plan (for instance, by multiple imputation or a complete-baseline population subset). More than one missing data handling method should be applied to demonstrate robustness of results independent of the applied approach. All data processing and cleaning steps should be documented to allow for highest transparency and reproducibility.

CAUSAL INFERENCE METHODS

The lack of randomization in an EC study means that patients in the trial and those in the external cohort typically differ systematically, leading to potential confounding and selection bias. These biases can distort treatment effect estimates and may undermine the credibility of the results if not handled carefully. Therefore, proper causal inference is essential to ensure that EC studies can yield credible evidence for regulatory and HTA decisions. In this section, we outline the main bias-reduction techniques for EC studies by focusing on propensity score methods, diagnostics for covariate balance, supplementary and sensitivity analyses, and possible approaches for missing data and unmeasured confounding.

Propensity Score Methods: Propensity score (PS) methods are commonly applied in observational studies, including EC studies. The PS represents the probability of receiving the treatment (i.e., being in the trial cohort) given a patient's observed baseline covariates. By summarizing many covariates into a single score, PS methods help achieving balance in measured covariates between the two cohorts efficiently. There may be slightly better options than PS methods, but PS methods have been traditionally applied for EC study submissions, such that this overview concentrates on PS methods as the selected causal inference approach.

Calculation and Covariate Selection: PS are typically estimated using logistic regression, which models treatment occurrence (trial vs. EC) as a function of baseline covariates. Choosing which covariates to include is critical and should be pre-specified in the protocol and the Statistical Analysis Plan. The model includes usually both true confounders (variables associated with both treatment selection and outcome) but also more generally outcome predictors. Including variables that influence treatment but not the outcome can needlessly increase variance without reducing bias. While logistic regression is the most common approach, more flexible machine learning methods (such as gradient boosting or random forests) can also be used, especially to capture non-linear relationships, provided they are carefully tuned to avoid overfitting.

Applications for Propensity Scores: Several approaches use propensity scores to adjust comparisons between cohorts:

- **Matching:** Propensity score matching (PSM) pairs each treated patient with one or more EC patients who have similar PS values. Common algorithms include nearest-neighbor matching (often within a specified caliper distance) but also the more sophisticated optimal matching approach. Matching is particularly useful when the external data pool is large. It typically targets the ATT, but estimating the ATE by full matching or estimating the ATU by reversing the process is also possible. Even estimating the ATO is thinkable by first performing a matching step and adding a weighting step afterwards. However, matching needs to be applied carefully, as it can lead to an overly reduced sample size, leading to unnecessary or even harmful low power, such that at the end false conclusions may be drawn.

- **Weighting:** Inverse Probability of Treatment Weighting (IPTW) uses PS to assign weights to each individual, creating a weighted population ("pseudo-population") in which the distribution of observed covariates is very similar in both cohorts. This approach can estimate the ATT, ATE, ATO and ATU. To improve stability, weights may be truncated or stabilized to limit the influence of individuals with very extreme PS values, but in EC studies this is typically not necessary because both cohorts are usually sufficiently harmonized due to applying very similar eligibility criteria across cohorts. Weighting generally retains all data, applying the scientific principle to use all relevant data when appropriate.

- **Stratification:** Patients can be stratified into subgroups based on their PS (e.g. quintiles). Outcomes are then compared within each stratum, and an overall effect is obtained by combining the stratum-specific estimates. Stratification uses all the data and is straightforward to implement. It can reveal whether treatment effects differ across levels of the PS (which might indicate heterogeneity of the treatment effect or residual confounding), providing important additional information. However, as a downside, not all bias is removed by PS stratification, because the PS distribution within each stratum is still expected to be different.

- **Covariate Adjustment:** Another approach is to directly include the PS (or the individual covariates) in a multivariable outcome regression model. For instance, one could regress the outcome on treatment and the PS (often with a flexible functional form such as spline functions). This method is easy to implement but relies on the correct specification of the outcome model. It is generally considered less robust than the above methods and also estimates just a conditional (regression-type) treatment effect but not a true marginal (population-related) treatment effect, such that other methods may be preferred.

Diagnostics and Balance Assessment: After applying any PS-based adjustment, it is crucial to verify that sufficient covariate balance has been achieved between the trial and EC cohort. The main diagnostic tool is displaying the Standardized Mean Differences (SMDs) for each baseline covariate and comparing them before and after adjustment. As a rule of thumb, an absolute SMD below |0.1| (10%) indicates good balance. One should also provide plots that display SMDs for all variables ("love plots"), as well as plots which overlay the propensity score and potentially also single covariate distributions of the two cohorts. For continuous covariates, comparing variance ratios (which should lie between 0.5 and 2) and using quantile-quantile plots is informative. If significant imbalances remain after adjustment, a "plan B" as specified by the Statistical Analysis Plan should be followed, e.g., that the ATO becomes the primary analysis, that entropy balancing is used instead of propensity score weighting or that another causal inference method like weighting or g-computation is performed after the matching step. Other more ad-hoc approaches like refining the PS model, for example, by adding interaction terms or non-linear terms without a clear pre-documented rule can lead to criticisms of lacking transparency.

Supplementary and Sensitivity Analyses: These additional analyses are performed to assess the robustness of the EC study results to key assumptions and analytical choices and to provide alternative perspectives. Typical analyses include:

- **Causal inference method:** Using an alternative causal inference method to see if the treatment effect estimate changes meaningfully. If the estimate remains consistent, it suggests that the result is not overly sensitive to the exact causal inference model used.

- **Handling missing data:** Applying different methods of dealing with missing baseline data. For instance, one could compare the primary analysis (say, multiple imputation) to an analysis that uses only complete baseline patient data or another different missing data approach. If the conclusions are similar, it increases confidence that the selected method to handle missing data is not decisively affecting the result.

- **Marginal estimators:** Applying more than one marginal estimator is recommended. It is most informative to present all 4 (ATT, ATE, ATO, ATU), at least for the primary endpoint, while one of them should be pre-selected to be the primary marginal estimator of interest.

- **Unmeasured confounding:** Although unmeasured confounders cannot be adjusted for directly, analysts can perform quantitative bias analysis to estimate how strong an unmeasured confounder would have to be to explain away the observed treatment effect. Calculating the E-value, for example, provides an indication of the minimum association that an unmeasured factor would need with both the treatment and the outcome to nullify the effect. If this threshold is implausibly high, the result can be considered robust.

All supplementary and sensitivity analyses should be pre-specified in the protocol or SAP and reported transparently. They are an integral part of demonstrating the credibility of an EC study, providing insights into how dependent the findings are on various methods or assumptions.

Advanced Methods Beyond Propensity Scores:

Causal Inference Methods: While propensity score methods handle bias from measured baseline confounders adequately, there are many alternative causal inference approaches which can be applied. Also, additional techniques are needed for issues like intercurrent events, informative censoring, time-dependent confounding or treatment switching.

Missing Data Handling Methods: Missing data on baseline covariates can introduce bias if not properly handled, especially if missingness is related to treatment or outcome. Multiple imputation is a common approach for dealing with missing baseline covariates: several complete datasets are created by imputing the missing values based on other observed data, and the analysis is replicated on each, and finally combined to account for uncertainty. It was shown that it is important to perform multiple imputation in the two cohorts separately, not by pooling the data. Sensitivity analyses should be conducted by applying different techniques or more generalized assumptions.

Addressing Unmeasured Confounding: The potential for unmeasured confounding remains a key limitation of any non-randomized study. As noted above, techniques like the E-value or quantitative bias analyses provide instruments to check robustness. Ultimately, acknowledging the possibility of unmeasured confounders and discussing their likelihood in conjunction with direction and magnitude of potential bias lends credibility to the analysis.

CONCLUSION

A well-designed EC study offers an approach to evidence generation when traditional RCTs are not feasible. However, according to the FDA, it should only be applied when the expected treatment effect is large (e.g., a hazard ratio of < 0.5 or potentially < 0.6) to minimize the impact of any residual bias. By leveraging RWD, EC studies can contextualize single-arm trial results especially for rare diseases where internal controls are too difficult to obtain. However, the validity of an EC study hinges on rigorous methodology: the application of the target trial emulation and estimand frameworks, meticulous data curation and cohort selection, similar eligibility criteria, the use of robust analytical techniques to adjust for differences between the trial and real-world patients, and applying many supplemental and sensitivity analyses is key.

Regulatory agencies are increasingly open to considering evidence from RW and EC studies, especially in situations of high unmet medical need, where an RCT is not possible. That said, RCTs will always remain the gold standard and should always be performed when adequate. Researchers planning an EC study should engage with agency stakeholders early to agree on the EC study operationalization and to implement highest levels of rigor and transparency. This requires pre-specifying all analyses, performing thorough diagnostics, and openly discuss limitations (such as the potential impact of unmeasured confounders). Such a careful methodological approach results in the EC study being held to highest standards.

As real-world data becomes more accessible and extensive, and as analytical methods continue to improve, EC studies are likely to play a growing role in clinical research and regulatory review, potentially bringing new therapies to patients faster.

REFERENCES

1. **Antunes L., Rippin G., Ralphs E., Arnold K., Luguzis A., Lee H. (2025)** Choosing an Index Date for Untreated Patients in External Comparator Studies. Drug Saf (2025). https://doi.org/10.1007/s40264-025-01613-x.
2. **Austin, P.C. (2011).** An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behavioral Research, 46(3), 399–424.
3. **Austin, P.C., & Stuart, E.A. (2015).** Moving towards best practice when using inverse probability of treatment weighting (IPTW) with the propensity score. *Statistics in Medicine, 34*(28), 3661–3679.
4. **Daniel, R.M., Cousens, S.N., De Stavola, B.L., Kenward, M.G., & Sterne, J.A. (2013).** Methods for dealing with time-dependent confounding. *Statistics in Medicine, 32*(9), 1584–1618.
5. **Hernán, M.A., & Robins, J.M. (2016).** Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology, 183*(8), 758–764.
6. **Hernán, M.A., & Robins, J.M. (2020).** *Causal Inference: What If.* Boca Raton, FL: Chapman & Hall/CRC.
7. **International Council for Harmonisation (ICH). (2019).** *ICH E9(R1) Addendum on Estimands and Sensitivity Analysis in Clinical Trials.* ICH Step 4 Guideline.
8. **Li, F., Morgan, K.L., & Zaslavsky, A.M. (2018).** Balancing covariates via propensity score weighting. *Journal of the American Statistical Association, 113*(521), 390–400.
9. **Mack, C., Christian, J., Brinkley, E., et al. (2020).** When context is hard to come by: External comparators and how to use them. *Therapeutic Innovation & Regulatory Science, 54*(4), 932–938.
10. **Rippin, G., Ballarini, N.M., Sanz, H., et al. (2022).** A review of causal inference methods for external comparator arm studies. *Drug Safety, 45*(8), 815–837.
11. **Rippin, G. (2024).** External comparators and estimands. *Frontiers in Drug Safety and Regulation, 3*, Article 1332040.
12. **Rippin, G., & Sanz, H. (2024).** External comparator studies and the joint application of the estimand and target trial emulation frameworks. *Frontiers in Drug Safety and Regulation, 4*, Article 1409102.
13. **Rippin, G., Sanz, H., Hoogendoorn, W.E., Ballarini, N.M., Largent, J.A., & Demas, E. (2024).** Examining the effect of missing data and unmeasured confounding on external comparator studies: case studies and simulations. *Drug Safety.* (Published online 2024).
14. **Rippin G., Largent J., Hoogendoorn W.E., Sanz H., Bosco J., Mack C. (2024).** External comparator cohort studies – clarification of terminology. Front Drug Saf Regul. 2024;3: 1321894. http://doi:10.3389/fdsfr.2023.1321894
15. **Rippin G., Sanz H., Hoogendoorn W.E., Largent J.** External Comparator Studies: Performance of 4 Missing Data Handling Approaches, Stratified by 4 Different Marginal Estimators, Drug Saf. 2025;48. https://doi.org/10.1007/s40264-025-01586-x
16. **Rosenbaum, P.R., & Rubin, D.B. (1983).** The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55.
17. **U.S. Food and Drug Administration (FDA) (2023).** Considerations for the Design and Conduct of Externally Controlled Trials for Drug and Biological Products. Draft Guidance for Industry, February 2023.

ACKNOWLEDGMENTS

The authors acknowledge the use of Microsoft Copilot as an assistive tool for refining the manuscript's language and structure; all content was reviewed and finalized by the authors in accordance with PHUSE guidelines on the use of AI in publications.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Vishal Goriya
Company: IQVIA RSD (India) Pvt Ltd
Email: vishal.goriya@iqvia.com