PRACTICAL APPLICATION

Transporting Comparative Effectiveness Evidence Between Countries: Considerations for Health Technology Assessments

Alex J. Turner[1] · Cormac Sammon[1] · Nick Latimer[2,3] · Blythe Adamson[4] · Brennan Beal[4] · Vivek Subbiah[5] · Keith R. Abrams[6,7] · Joshua Ray[8]

Abstract

Internal validity is often the primary concern for health technology assessment agencies when assessing comparative effectiveness evidence. However, the increasing use of real-world data from countries other than a health technology assessment agency's target population in effectiveness research has increased concerns over the external validity, or "transportability", of this evidence, and has led to a preference for local data. Methods have been developed to enable a lack of transportability to be addressed, for example by accounting for cross-country differences in disease characteristics, but their consideration in health technology assessments is limited. This may be because of limited knowledge of the methods and/or uncertainties in how best to utilise them within existing health technology assessment frameworks. This article aims to provide an introduction to transportability, including a summary of its assumptions and the methods available for identifying and adjusting for a lack of transportability, before discussing important considerations relating to their use in health technology assessment settings, including guidance on the identification of effect modifiers, guidance on the choice of target population, estimand, study sample and methods, and how evaluations of transportability can be integrated into health technology assessment submission and decision processes.

1 Introduction

Health technology assessment (HTA) agencies commonly express a preference for evidence from randomised controlled trials (RCTs) to guide decision making, citing that randomisation eliminates non-chance confounding from unmeasured patient characteristics [1–3]. Conditional on no loss to follow-up and no measurement error, blinded randomisation in large samples ensures RCTs produce unbiased estimates of comparative efficacy within the trial sample, i.e. effects that are internally valid [4].

However, evidence-based decision making also requires that evidence is externally valid, i.e. that treatment effects from a study represent unbiased estimates of efficacy in a specific target population of interest. When study populations are not randomly selected from a target population, external validity is more uncertain and it is possible that distributions of effect modifiers (characteristics that predict variation in treatment effects) differ between the trial sample and target population [5]. Trial samples rarely represent random samples of the target population, often having strict inclusion/exclusion criteria, targeting patients with the highest risk of an outcome, and containing a selected sample of patients agreeing to participate [6].

Studies using real-world data (RWD) are increasingly common sources of comparative-effectiveness evidence in both regulatory and HTA submissions, particularly in rare cancers and other diseases, where conducting well-powered RCTs is difficult [7–9]. Despite the absence of randomisation reducing the likelihood of internal validity,

✉ Joshua Ray
joshua.ray@roche.com

1 Putnam PHMR, London, UK

2 School of Health and Related Research, University of Sheffield, Sheffield, UK

3 Delta Hat, Nottingham, UK

4 Flatiron Health, New York, NY, USA

5 Sarah Cannon Research Institute, Nashville, TN, USA

6 Department of Statistics, University of Warwick, Coventry, UK

7 Centre for Health Economics, University of York, York, UK

8 Global Access, F. Hoffmann-La Roche Ltd, Grenzacherstrasse 124, 4070 Basel, Switzerland

Key Points

The transportability of comparative effectiveness evidence generated using data from other countries has become a key concern of health technology assessment agencies.

Methods have been developed to measure and adjust for a lack of transportability but their consideration in a health technology assessment is limited.

This article provides an introduction to transportability and related methods, and discusses important considerations regarding their use in health technology assessment settings.

RWD studies are often conducted in samples of patients being treated in clinical practice, which are likely more representative of a decision-maker's target population.

However, external validity does not just refer to generalisability but also to transportability [6]. Whereas *generalisability* relates to whether inferences from a study can be extended to a target population from which the study dataset was sampled, *transportability* relates to whether inferences can be extended to a separate (external) population from which the study sample was not derived [10]. The availability and quality of RWD vary substantially by country, meaning that submissions to some HTA agencies often include evidence from patients residing in other countries. Similarly, in many countries, new therapies will not be available to patients prior to reimbursement, meaning local RWD sources will not collect data on all treatments of interest, and manufacturers must use RWD from countries where the therapy is already adopted [11]. The transportability of evidence from other countries, estimated either from RCTs or RWD, has therefore become a key concern of HTA agencies [12], driven for example by cross-country differences in potential effect modifiers such as disease characteristics, comparator therapies and treatment settings, and has led to a preference for local data [1, 13].

When evidence from other countries is used to inform comparative effectiveness (or efficacy), consideration of transportability in study design and analysis should be a key step. Methods have been developed to correct for a lack of transportability [14]. However, these methods have received little attention in an HTA setting. This may be because of limited knowledge of the methods and/or uncertainties in how best to utilise them within existing HTA frameworks.

Existing studies have provided detailed technical information on these methods [14, 15]. This article aims to provide a concise summary of methods available for identifying and correcting for a lack of transportability, with a specific focus on transporting evidence from either RCTs or RWD and transporting evidence across countries. We also extend existing studies by discussing important considerations regarding their use in HTA settings. Although we focus specifically on transporting results across countries, we note that HTA agencies may also be interested in transporting inferences across patient subgroups and treatment settings. Similar considerations will also be relevant when transporting results across regions within the same country. We also focus on transporting estimates of relative treatment effects, although RWD studies in particular may be used for a variety of other purposes in HTAs, for example, for deriving model inputs such as utilities relating to disease states, meaning transporting absolute rather than comparative effects is a concern.

2 Methods for Transportability

2.1 Key Assumptions

Typically, decision makers are interested in the target population average treatment effect (PATE): the average effect of treatment if all individuals in the target population were assigned the treatment. However, researchers commonly have access only to a sample and must estimate the study sample average treatment effect (SATE).

Westreich et al. define the ability of the SATE to approximate the PATE as "target validity", and deviations from the PATE as "target bias" [6]. These deviations can occur because of threats to internal validity, such as confounding and threats to external validity (Fig. 1).

Transportability is a specific form of external validity, where the study sample is not a subset of the target population (Fig. 2), for example, when extending inferences from a sample of patients from one country to patients in a different country. This differs from generalisability, where the study sample is a subset of the target population, for example, if extending inferences from a sample of patients with a given condition to all patients with that condition in the *same* country.

In addition to data on the study sample, methods to identify and adjust for a lack of transportability require a sample of data on the target population (the "target sample"). The key assumptions required to identify the PATE under transportability are outlined in Box 1 [14, 16, 17]. Standard assumptions for internal validity are required, including: (1) no differences in unmeasured outcome predictors

Fig. 1 Drivers of the differences between the sample average treatment effect (SATE) and the population average treatment effect (PATE). Adapted from Degtiar and Rose [14]. External validity bias and internal validity bias can be positive or negative and may bias the SATE in different directions. Sampling variability may also cause deviation from the PATE
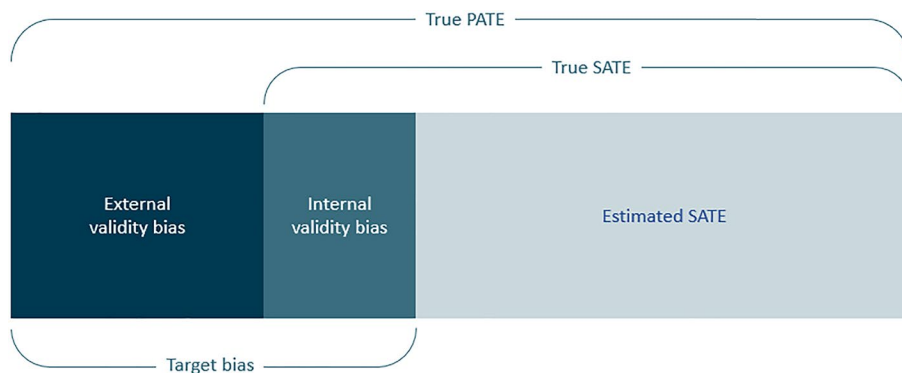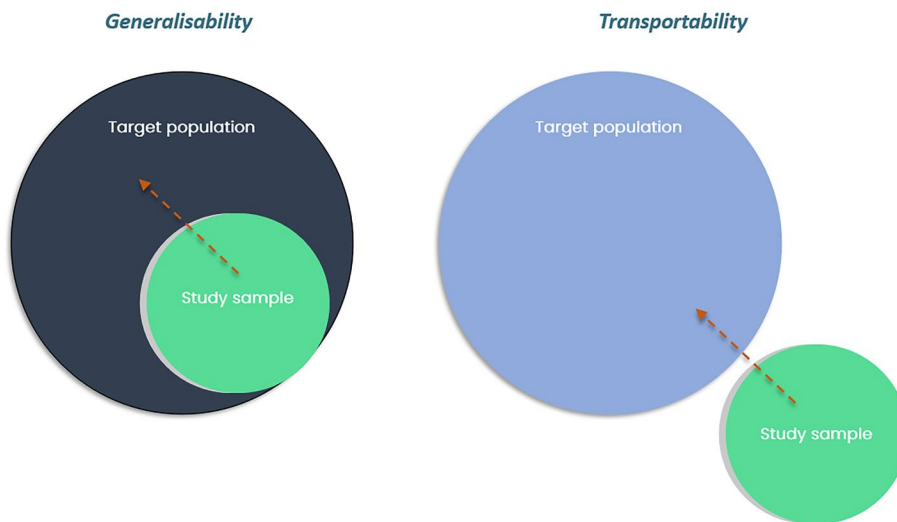

Fig. 2 Relevant study and target populations in the case of generalisability and transportability

(confounders) between treatment arms (*conditional treatment exchangeability*); (2) overlap in the characteristics of measured outcome predictors between treatment arms (*positivity of treatment assignment*); and (3) that the outcomes of individuals do not depend on the treatments received by others (*stable unit treatment value assumption*). Analogues of these assumptions for external validity are then required, and relate to differences between the study and target samples (rather than between treatment arms).

The *conditional mean difference exchangeability of study selection* assumption will often be most crucial, and will be the focus of the analysis. This includes identification and adjustment for observed effect modifiers, and arguments over the plausibility of no unmeasured effect modification. Where treatment effects are estimated from RWD, the plausibility of no unobserved confounding is also key. Whilst transporting absolute effects is not the focus of this paper, no unobserved confounding is also required for external validity in this case.

To explain these concepts, a hypothetical example is provided in Box 2.

2.2 Overview of Key Methods

Methods for identifying and correcting for barriers to transportability are summarised in Fig. 3, based on information in existing published reviews [14, 15]. Detailed information on these methods is provided in the Appendix. As in HTA submissions, the burden of evidence generation falls on the manufacturers (who will commonly have access to the individual patient-level data [IPD] from the study sample), we focus on methods where IPD is available for the study sample and where either IPD or summary data are available from a target sample. However, we note that alternative methods such as meta-analytic techniques can be used when only summary data are available on both samples [14].

Identifying transportability involves testing its assumptions, and methods depend on the availability of outcome data in the target sample. Where outcome data are unavailable, requirements underlying the *conditional mean difference exchangeability of study selection* assumption are tested in two steps by: (1) assessing differences in the distributions of characteristics between study and target populations and (2) identifying whether characteristics driving these differences

Box 1 Assumptions for identifying the PATE under transportability

---

Assumptions for Internal Validity

- **Conditional treatment exchangeability:** requires no unmeasured confounding when estimating the treatment effects in the study sample.

- **Positivity of treatment assignment:** requires each subject in the study sample to have a positive probability of receiving each treatment.

- **Stable unit treatment value assumption:** requires no interference between subjects, i.e. an individual's outcome depends only on their assigned treatment not on treatments of others.

Assumptions for External Validity (Transportability)

- **Conditional mean difference exchangeability for study selection:** mean difference exchangeability requires that all characteristics explaining treatment effect heterogeneity across individuals (i.e. all observed and unobserved effect modifiers) have equivalent average values in the study sample and target population. This implies the SATE can equal the PATE without adjustment for effect modifiers. When this assumption fails, *conditional* mean exchangeability requires that mean differences in outcomes between treatments at each covariate value is identical in the study sample and target population. This requires that all effect modifiers that differ between study and target samples are measured. Failure of this assumption means adjustment for observed effect modifiers will not identify the PATE.

- **Positivity of selection:** states that the probability of study participation, conditional on covariates, lies between zero and one. This requires that all members of the target population are represented by individuals in the study, i.e. there is overlap in the distribution of effect modifiers, such that there exists individuals from both study and target samples in every stratum of effect modifiers. Where propensity score models are used to adjust for differences between the study sample and target population, this assumption instead requires that the distribution of propensity scores in the study sample and target population have a sufficient overlap or common support. This assumption enables adjustment for effect modifiers without extrapolation.

- **SUTVA for study selection:** requires no interference between subjects in the study sample and target sample, and no differences in how outcomes are measured and in the distribution of versions of treatment across the study sample and the target population.

Adapted from Degtiar and Rose [14]. Assumptions of internal validity (such as no unmeasured confounding) will hold automatically in randomised controlled trials, but will need to be carefully considered where non-randomised studies using real-world data are used to estimate comparative effectiveness. Other estimands may be of interest to decision makers, such as the target population average treatment effects among the treated. These estimands are not the focus of this article, but similar assumptions to the above will apply. *PATE* population average treatment effect, *SATE* sample average treatment effect

---

Box 2 Assumptions for identifying the PATE under transportability

---

Background

Consider a case where a manufacturer is submitting evidence to an HTA agency in a hypothetical developing country, Aldova, to support reimbursement of a hypothetical oncology treatment, immunotuzumab. The Aldovan HTA agency prefers local data, but because of the lack of ability to recruit Aldovan patients for an RCT, trial evidence is not available to provide estimates of comparative effectiveness. In addition, although Aldova has a robust oncology RWD source, because immunotuzumab has not yet launched in Aldova, the RWD source does not provide data on the treatment of interest so cannot be used for the submission. However, immunotuzumab has previously received approval in a developed country, Belgravia, enabling the manufacturer to estimate comparative effectiveness using Belgravian RWD for the submission. Evidence suggests income is an important determinant of the effectiveness of immunotuzamab, i.e. is an effect modifier.

Target population and estimands

In this case, the target population is the population of Aldova and the PATE refers to the ATE in this population. The SATE is the ATE in the Belgravian study sample. Transportability here refers to the degree to which the ATE estimated using Belgravian RWD reflects the ATE that would be estimated in Aldova if suitable Aldovan data were available.

Assumptions for transportability

For the Belgravian ATE to equal the (unobserved) Aldovan ATE, mean difference exchangeability requires that all effect modifiers have equivalent average values in the Belgravian and Aldovan populations. This is unlikely to be plausible in this example because income is an effect modifier and average incomes are likely to differ between the less developed Aldova and the more developed Belgravia. Adjustment for the lack of transportability (using the Aldovan RWD source as the target sample) is therefore needed. As a result, the conditional mean difference exchangeability assumption requires that data on income (and other effect modifiers that differ across the two countries) be available in both the Belgravian study sample and the Aldovan target sample. The validity of the positivity of selection assumption may also be a concern and would need to be examined, as given the likely large differences in the distribution of income across developed and developing countries, it is possible that no individual in the Belgravian sample has an income as low as individuals with the lowest incomes in Aldova. Adjustment methods which allow for extrapolation (see Appendix) beyond the common support are likely more appropriate in these cases.

Given comparative effectiveness is estimated from a RWD source, the manufacturer must also consider threats to internal validity (as well as transportability).

ATE population average treatment effect, *HTA* health technology assessment, *PATE* population average treatment effect, *RCT* randomised controlled trial, *RWD* real-world data, *SATE* sample average treatment effect
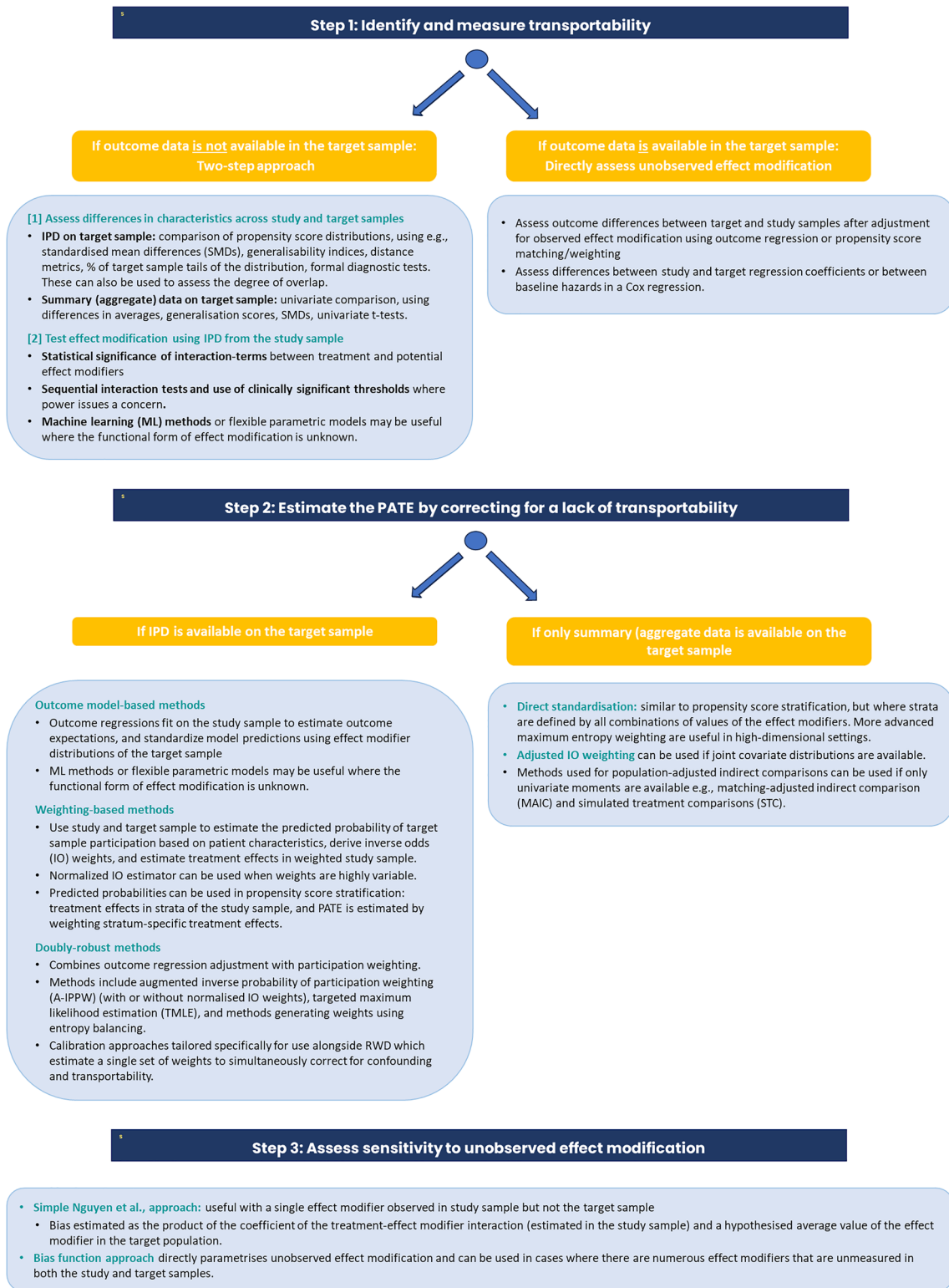
Step 1: Identify and measure transportability

If outcome data is not available in the target sample:
Two-step approach

[1] Assess differences in characteristics across study and target samples
- **IPD on target sample:** comparison of propensity score distributions, using e.g., standardised mean differences (SMDs), generalisability indices, distance metrics, % of target sample tails of the distribution, formal diagnostic tests. These can also be used to assess the degree of overlap.
- **Summary (aggregate) data on target sample:** univariate comparison, using differences in averages, generalisation scores, SMDs, univariate t-tests.

[2] Test effect modification using IPD from the study sample
- **Statistical significance of interaction-terms** between treatment and potential effect modifiers
- **Sequential interaction tests and use of clinically significant thresholds** where power issues a concern.
- **Machine learning (ML) methods** or flexible parametric models may be useful where the functional form of effect modification is unknown.

If outcome data is available in the target sample:
Directly assess unobserved effect modification

- Assess outcome differences between target and study samples after adjustment for observed effect modification using outcome regression or propensity score matching/weighting
- Assess differences between study and target regression coefficients or between baseline hazards in a Cox regression.

Step 2: Estimate the PATE by correcting for a lack of transportability

If IPD is available on the target sample

Outcome model-based methods
- Outcome regressions fit on the study sample to estimate outcome expectations, and standardize model predictions using effect modifier distributions of the target sample
- ML methods or flexible parametric models may be useful where the functional form of effect modification is unknown.

Weighting-based methods
- Use study and target sample to estimate the predicted probability of target sample participation based on patient characteristics, derive inverse odds (IO) weights, and estimate treatment effects in weighted study sample.
- Normalized IO estimator can be used when weights are highly variable.
- Predicted probabilities can be used in propensity score stratification: treatment effects in strata of the study sample, and PATE is estimated by weighting stratum-specific treatment effects.

Doubly-robust methods
- Combines outcome regression adjustment with participation weighting.
- Methods include augmented inverse probability of participation weighting (A-IPPW) (with or without normalised IO weights), targeted maximum likelihood estimation (TMLE), and methods generating weights using entropy balancing.
- Calibration approaches tailored specifically for use alongside RWD which estimate a single set of weights to simultaneously correct for confounding and transportability.

If only summary (aggregate data is available on the target sample

- **Direct standardisation:** similar to propensity score stratification, but where strata are defined by all combinations of values of the effect modifiers. More advanced maximum entropy weighting are useful in high-dimensional settings.
- **Adjusted IO weighting** can be used if joint covariate distributions are available.
- Methods used for population-adjusted indirect comparisons can be used if only univariate moments are available e.g., matching-adjusted indirect comparison (MAIC) and simulated treatment comparisons (STC).

Step 3: Assess sensitivity to unobserved effect modification

- **Simple Nguyen et al., approach:** useful with a single effect modifier observed in study sample but not the target sample
  - Bias estimated as the product of the coefficient of the treatment-effect modifier interaction (estimated in the study sample) and a hypothesised average value of the effect modifier in the target population.
- **Bias function approach** directly parametrises unobserved effect modification and can be used in cases where there are numerous effect modifiers that are unmeasured in both the study and target samples.

Fig. 3 Methods of identifying and adjusting for barriers to transportability. *IO* inverse odds, *IPD* individual patient data, *ML* machine learning, *SMD* standardised mean difference

are effect modifiers (i.e. explain treatment effect heterogeneity). Differences in distributions can also be used to test for potential violations of the *positivity of selection*. Where outcome data are available, the degree of transportability can be measured directly by testing for an unobserved effect modification using outcome or treatment effect differences between study and target samples after adjustment for observed effect modifiers.

Accounting for a lack of transportability involves adjustment to ensure similarity in observed effect modifiers between study and target samples. Methods depend on whether IPD are available on characteristics of the target sample. Where IPD are available, analogues of methods used to adjust for observed confounding can be employed, including outcome regression-based methods, matching, stratification, inverse odds of participation weighting and doubly robust methods combining matching/weighting with regression adjustment. Where only aggregate data are available, methods including those employed for population-adjusted indirect treatment comparisons, such as matching-adjusted indirect comparisons, can be used.

A key assumption of all methods is no unmeasured effect modification, and a sensitivity analysis is available to assess the extent to which results are sensitive to deviations from this assumption [18, 19]. Appropriate methods depend on whether unmeasured effect modifiers are missing from the target sample only or from both study and target samples.

Finally, many of these methods have been developed for transporting treatment effects from RCTs, but can also be applied to comparative-effectiveness studies using RWD. This can involve pre-processing study sample data using weighting/matching to balance observed confounders prior to applying transportability methods [19], or using bespoke methods that generate weights to simultaneously correct for confounding and transportability [20].

3 Considerations for the Application of Transportability Methods for HTA

In the following, we discuss several considerations for the application of transportability methods in an HTA context (Table 1), building on good practices outlined in the epidemiological literature [14, 15].

3.1 Choice of Target Population and Estimand

Discussions regarding the relevant target population(s) should be driven by the decision the study aims to inform. Where HTA agencies make decisions over national reimbursement, the relevant target population will be the whole population of individuals with a disease. However, where HTA agencies make decisions over regional reimbursement or reimbursement within certain subgroups (e.g. patients with high-risk genetics), national disease-wide populations will not be appropriate target populations.

The decision an analysis aims to inform should also guide the appropriate estimand. Decision makers will typically be interested in the PATE; however, if treatment discontinuation or switching is substantial or is not expected to reflect routine practice or outcomes, other estimands may be appropriate.

Following the choice of estimand, investigators must clearly outline the assumptions required to identify it. This will guide the choice of methods and allow HTA committees to judge the plausibility of the assumptions.

3.2 Identification of Effect Modifiers

Potential effect modifiers should be identified during study design [15]. Four classes of effect modifiers should generally be considered: patient/disease characteristics (e.g. biomarker prevalence), setting (e.g. location of and access to care), treatment (e.g. timing, dosage, comparator therapies, concomitant medications) and outcomes (e.g. follow-up or timing of measurements) [14, 21]. Beal et al. have derived a checklist for effect modifiers specifically relevant for the transportability of data from US electronic health records for HTA use in cases in oncology [22]. Although this checklist is designed for oncology, many of its components (such as differences in baseline demographics, disease and biomarker prevalence, disease assessment frequency, concomitant medications and access to supportive care) are relevant in the majority of disease areas. However, development of checklists for other specific disease areas is warranted.

Directed acyclic graphs represent a simple and transparent approach to identifying potential causal relationships between variables [23], and may be useful both in identifying effect modifiers and in communicating sources of transportability bias to HTA agencies. This should be supplemented with systematic literature searches and elicitation of expert opinion. As similar processes are recommended by HTA agencies to identify potential confounders [1, 24], we recommend that confounders and effect modifiers are identified concurrently when a RWD study is being used to estimate comparative effectiveness. Once study and target samples have been selected, the two-step approach used to identify transportability outlined above could be used to identify effect modifiers relevant to the particular study setting.

Table 1 Summary of the key considerations for the application of transportability methods in HTAs

Consideration	Summary
Defining the target population(s) and estimand of interest	Select the target population and estimand based on the decision a study aims to inform Clearly outline the assumptions required to identify the estimand
Identification of effect modifiers	Systematic and transparent approach to identifying effect modifiers during study design using directed acyclic graphs, systematic literature searches and expert opinion Assess effect modification using data from the study sample
Choice of study and target samples	Choose a study sample based on a trade-off between internal validity and external validity, guided by a tool used to assess data suitability (e.g. Data Suitability Assessment Tool [Data-SAT] or REQueST tools) Where HTA dossiers will be submitted to multiple markets, the study sample can be chosen to ensure representativeness of a single or set of similar key markets, but should also contain information to enable transporting of evidence to other markets Choose target sample based on availability of IPD, representativeness of the target population, availability of effect modifiers, and consistency in the measurement of key variables between target and study samples
Choice of methods to identify and adjust for a lack of transportability	To identify/measure transportability, global and univariate measures should be used to summarise differences between study and target samples, before testing effect modification using the study sample Available methods are dependent on access to IPD on the target sample Doubly robust methods have been suggested as the base-case method for adjusting for a lack of transportability when IPD on the target sample available Use of complex methods should consider the trade-off between potential benefits and drawbacks relating to reduced accessibility to investigators and HTA agencies
Integrating transportability into the HTA submission and decision process	Transportability methods should be pre-specified in a study protocol and statistical analysis plan Early engagement with HTA agencies and other key stakeholders crucial to ensuring alignment on appropriate approaches HTA agencies should provide clear guidelines for transportability analysis Need for demonstration studies and methods development relating to dynamic treatments Primary responsibility for implementing transportability methods often lies with the manufacturer. However, without a manufacturer-submitted analysis, HTA agencies should consider conducting their own simple transportability analysis

HTAs health technology assessments, *IPD* individual patient-level data

3.3 Choice of Study Sample

The choice of the study sample (e.g. between a RCT and a RWD source) should be guided by a trade-off between internal validity and external validity. Where a RWD source is considered, the choice of a specific RWD source could be informed by completion of tools such as EUnetHTA's REQueST tool [25] or the Data Suitability Assessment Tool (DataSAT) tool produced by the National Institute for Health and Care Excellence [26], which assess data quality, suitability of use for HTA purposes and appropriateness for answering a specific research question. This should be supplemented with a checklist similar to that produced by Beal et al. [22].

Where possible, a study sample should be chosen that both has characteristics ensuring internal validity and draws samples directly from the target population (country), avoiding problems with transportability (e.g. RCTs conducted within the target population or high-quality local RWD). Where prospective studies (either trial or RWD) are used, these should be designed to ensure generalisability of the study sample, such as the use of pragmatic RCTs or purposeful sampling [14], or where this is not possible designed so that relevant effect modifiers are collected.

However, where existing RWD sources are used, manufacturers may be forced to use study samples from countries outside of a decision-maker's population. This could be owing to local data sources either not existing or lacking clinical detail or because of an intervention not currently being available in the target population. In this case, it is important that the study sample is chosen to maximise transportability, by choosing a sample that is most similar in effect modifiers to the target population and/or collects data on all effect modifiers. Importantly, the absence of data at a patient level will generally prevent adjustment for effect modifiers relating to setting and treatment differences across countries, and thus ideally, the study sample should be located in a country with a similar healthcare design and with similar treatment guidelines for the disease of interest.

We must also recognise that manufacturers typically seek reimbursement in multiple markets, resulting in separate target populations for each HTA agency. As it is unlikely that separate study samples will be chosen to estimate comparative effectiveness in each market, manufacturers are likely to choose a single study sample. This could be chosen to be similar in effect modifiers to a population in a single key market, or chosen to be representative as possible to multiple key markets (with the caveat that this may mean the sample is not representative of any market). In each case, data should be available to apply methods to transport evidence to other markets.

A special case is where RWD is used to derive an external control arm for a single-arm trial; a case increasingly common in HTA submissions [9]. If either the single-arm trial or the external control arm is sampled from a country outside of the target population, external validity (in addition to internal validity) requires transporting absolute effects, and therefore requires an assumption of no differences in unmeasured prognostic variables across two countries (in addition to no differences in unmeasured effect modifiers). Where arms are sourced from different countries both not located in the target population, adjustment would need to ensure a balance of all prognostic variables both between treatment arms and between each arm and the target population, and would require a stronger assumption of no differences in unmeasured prognostic variables across three countries. Given the limited plausibility of these assumptions in this setting, the use of methods to assess the sensitivity of results to unobserved confounding and effect modification will be crucial, although residual uncertainty surrounding results is likely to be unavoidable in these cases.

3.4 Choice of Target Sample

The target sample, if data are available, should be chosen with representativeness of the target population in mind, otherwise transporting results to the target sample may not ensure identification of the PATE. A RWD source will typically be used as the target sample. Unless a census forms the target sample (meaning the target sample equals the target population), representativeness is not guaranteed and will need to be assessed.

To enable identification and adjustment for a lack of transportability, the target sample should report data on all effect modifiers. For interventions where many effect modifiers are plausible, sources that include detailed information on patient/disease characteristics, such as registries or electronic health records, may be most appropriate. Where plausible effect modifiers are restricted to commonly recorded patient characteristics, less rich data sources such as claims databases may be sufficient. If the general population is the target population, data from censuses could serve as

the target sample, although analysis would be restricted to methods that only require aggregate data. All else equal, a target sample with IPD available should be preferred over aggregate data on the target population, as this allows a more robust adjustment for effect modification. Effect modifiers must also be defined similarly to in the study sample, and must be reported with sufficient completeness and accuracy.

3.5 Methods to Identify and Adjust for Barriers to Transportability

Determining the most appropriate transportability methods in practice is not trivial. Here, we provide some important considerations in an HTA setting.

3.5.1 Identifying/Measuring Transportability

As it is unlikely that outcomes of all treatments will be available in the target sample, the two-step approach that first examines differences in characteristics between the target and study samples before assessing treatment effect heterogeneity in the target sample will be most appropriate. Multiple methods should be used, including standardised mean differences of propensity scores, and examinations of propensity score distributions alongside formal diagnostic tests to identify the absence of an overlap [14]. Univariate standardised mean differences and tests can then be used to examine drivers of overall differences. With only summary target sample data available, analysis could include computing simple differences in averages, alongside standardised mean differences and univariate tests if standard deviations are reported.

Parametric models with treatment-covariate interactions can be used to detect effect modification. Where small study samples result in power issues or where unknown functional forms increase the risk of model misspecification, machine learning techniques such as Bayesian additive regression trees could be considered, and the use of directed acyclic graphs may be particularly crucial for selecting effect modifiers in this case.

Where tests indicate no differences in characteristics across target and study samples or that characteristics are not effect modifiers, comparative effectiveness can be estimated using the study sample without adjustment for transportability, noting that unobserved effect modification is always possible. If assessments of no overlap suggest methods to adjust for a lack of transportability that cannot be applied, investigators could consider changing the target population [27]. For example, whilst evidence may not be transportable to all patients with a condition, they may be transportable to specific patient subgroups, which may still be of interest to HTA agencies.

3.5.2 Adjusting for Barriers to Transportability

Where IPD is available for the target sample, existing reviews suggest that doubly robust methods should be used as the base case, as they are unbiased if only either the propensity model or outcome model is misspecified [15]. Normalised inverse odds weights are recommended in the event of variable weights. Machine learning could be explored to select functional forms in outcome/propensity models when sample sizes are small and/or where functional forms are unknown. Standardised mean differences between study and target samples should be compared before and after weighting to test for balance. Doubly robust or regression-based methods may be preferable in cases of a limited overlap in characteristics between study and target samples, as they allow for adjustment using extrapolation beyond the common support.

Where study samples are derived from RWD, pre-processing using matching/weighting to ensure balance on observed confounders across treatment arms should be conducted, with newer methods that simultaneously correct for confounding and transportability also considered [20]. Given existing simulation studies have been primarily conducted by creators of these methods, "neutral comparisons" could be useful in assessing the relative performance of methods that are free from potential investigator bias [28, 29]. Further targeted simulation studies that explore relative performance in scenarios similar to where methods would be used in practice could also be warranted.

Where only aggregate data are available on the target sample, the adapted inverse odds weighting approach should be considered if data are available on the joint distribution of covariates. If only univariate statistics are available, methods for population-adjusted indirect comparisons can be used, with existing guidelines informing method selection [30]. However, the potential for residual bias when using these methods must be noted.

Accessibility by HTA agencies could also be considered. Recognising existing concerns regarding the acceptability of RWD methods by HTA agencies [31], some argue that the potential benefits of complex approaches such as machine learning should be compared against the potential reduction in the ability to communicate findings [32]. However, a counter argument is that the optimal methods should be used with the onus on decision makers to ensure knowledge of these methods.

Unadjusted results should also be presented as a comparison to demonstrate the impact of adjustment. For effect modifiers relating to setting and treatment differences across countries, where data will generally be absent, their impact must primarily be argued qualitatively, minimised through a careful choice of the study sample and explored in a sensitivity analysis.

3.5.3 Sensitivity Analysis

Consistent with latest HTA guidance for estimating comparative effectiveness using RWD [1], sensitivity analyses should explore the impact of uncertain decisions made during an analysis, and explore the impact of violations of key assumptions. Results using alternative methods for adjusting for transportability could be presented alongside those using doubly robust methods. However, Bayesian model averaging approaches, which have been recommended for addressing model uncertainty more generally in health decision models, could be explored rather than presenting decision makers with numerous alternative treatment effects [33].

A sensitivity analysis assessing a potential unobserved effect modification should always be conducted, and approaches specifically designed for this purpose, such as the Nguyen et al. and bias function approaches, should be considered [18, 19]. Quantitative bias analysis methods addressing unobserved confounding, such as the E-value, are receiving increased attention in HTAs, and allow analysts to either (1) examine the size of bias required to change a result (e.g. treatment effect relative to some decision-making threshold) or (2) estimate the direction, magnitude and uncertainty of bias associated with treatment effects [34, 35]. However, to our knowledge, these have not yet been applied in this setting. Additional analyses may include comparing transported treatment-specific outcomes to outcomes in the target sample, if these outcome data are available in the target sample. A country-specific subgroup analysis may also be useful where a multi-country RCT forms the study sample, noting that reduced sample sizes will lead to a loss of power and will break randomisation unless the trial design ensured random treatment allocation within countries.

3.6 Integrating Transportability into the HTA Submission and Decision Process

Considerations regarding the choice of study and target samples and use of transportability methods should be pre-specified in a study protocol and statistical analysis plan. Where it is anticipated that transportability will form a key part of a submission, early engagement with HTA agencies and other key stakeholders will be crucial to ensure alignment on whether transportability is a concern and how transportability methods can help address that. To support this, HTA agencies should provide clear guidelines on transportability methods, including whether they should be used in

the base-case or as a sensitivity analysis in cost-effectiveness models. A review of RWD guidelines highlighted that only two HTA guidelines discuss transportability, and none describes methods to adjust for a lack of transportability [21]. Guidelines should also explicitly outline a hierarchy of study designs, recognising the trade-off between internal and external validity in the use of RCTs.

As with other evidence-generation activities for HTA, primary responsibility for carrying out transportability analysis will often lie with the manufacturer. However, where a manufacturer does not provide consideration of transportability in a submission, it may be prudent for HTA bodies to utilise simple methods to examine the likely impacts on results, noting any limitations. This could include the identification of effect modifiers from targeted reviews and/or expert elicitation, comparisons of summary statistics between the study sample and aggregate data on an easily accessible target sample, followed by a conservative sensitivity analysis. Knowledge that a default conservative approach would be applied in the absence of more considered approaches may incentivise manufacturers to conduct a thorough investigation addressing transportability.

As with evidence from non-randomised studies, acceptance of results transported from other countries by HTA agencies will depend on their confidence that transportability methods sufficiently address potential bias. We note that in many cases, providing HTA agencies with this confidence may be difficult. Unlike methods for correcting other biases, transportability methods require data from a second suitable data source (the target sample) to be identified. Often, the lack of appropriate data to conduct an analysis within the target population will be the reason such methods are needed. As a result, the applicability of these methods is likely higher when transporting data to a country with high-quality RWD in the disease of interest, but where the lack of adoption of the key treatment prevents the use of this data source for assessing comparative effectiveness. Even when an appropriate target sample is available, concerns over residual bias from unobserved effect modifiers will remain. As when addressing confounding, a sensitivity analysis will be crucial for highlighting and quantifying uncertainty in study findings to HTA agencies. However, the bar is arguably higher when adjusting for effect modification (across study and target samples) compared with when addressing confounding (between patients on different treatments), as the absence of patient-level data on important potential effect modifiers relating to treatment and setting characteristics will generally prevent adjustment (Sect. 3.3). Transportability methods will therefore be most suitable where differences in these effect modifiers can be minimised through a careful choice of the location of the study sample.

Given these threats to validity, building on applied studies in a generalisability setting [36] and examples transporting absolute effects [37], applied demonstration studies are needed to highlight that relative treatment effects can be transported across borders with minimal bias. To date, no such application exists. Replication studies, similar in style to RCT DUPLICATE and other initiatives that assess the ability of non-randomised studies to replicate RCT evidence [38], are warranted. Such studies would use published treatment effects as a benchmark and aim to replicate these effects by applying transportability methods to data from other countries. Continued methods development is also required, as methods are currently not available for transporting dynamic treatment effects.

4 Conclusions

A variety of methods are available to adjust comparative-effectiveness evidence to attempt to improve transportability, and are applicable using either patient-level or aggregate data. These methods require good data on the target population, which may not always be available, and the use of a sensitivity analysis to explore a likely unmeasured effect modification. Nevertheless, transportability methods can play an important role in ensuring that comparative-effectiveness evidence is appropriately considered in HTAs.

This article highlights several methodological and practical considerations that must be examined if these methods are to be successfully used in the HTA setting. These points include the choice of target population and study and target sample based on the relevant decision and decision maker(s) an analysis aims to inform, careful identification of effect modifiers and methods to adjust for them, and considerations on how these methods can be integrated into existing HTA evidence frameworks. Clear methodological guidelines on the application of transportability methods in the HTA setting, alongside further simulation studies to identify optimal methods and demonstration studies applying these methods, will facilitate their wider uptake.

References

1. National Institute for Health and Care Excellence. NICE real-world evidence framework. National Institute for Health and Care Excellence; 2022. London.

2. Institute for Quality and Efficiency in Health Care. General methods. Version 6.0. Institute for Quality and Efficiency in Health Care; 2020. Germany.

3. Makady A, Ham RT, de Boer A, Hillege H, Klungel O, Goettsch W. Policies for use of real-world data in health technology assessment (HTA): a comparative study of six HTA agencies. Value Health. 2017;20(4):520–32.

4. Hernán MA, Hernández-Díaz S. Beyond the intention-to-treat in comparative effectiveness research. Clin Trials. 2012;9(1):48–55.

5. Westreich D, Edwards JK, Lesko CR, Stuart E, Cole SR. Transportability of trial results using inverse odds of sampling weights. Am J Epidemiol. 2017;186(8):1010–4.

6. Westreich D, Edwards JK, Lesko CR, Cole SR, Stuart EA. Target validity and the hierarchy of study designs. Am J Epidemiol. 2019;188(2):438–43.

7. Bullement A, Podkonjak T, Robinson MJ, Benson E, Selby R, Hatswell AJ, et al. Real-world evidence use in assessments of cancer drugs by NICE. Int J Technol Assess Health Care. 2020;10:1–7.

8. Hatswell AJ, Baio G, Berlin JA, Irs A, Freemantle N. Regulatory approval of pharmaceuticals without a randomised controlled study: analysis of EMA and FDA approvals 1999–2014. BMJ Open. 2016;6(6): e011666.

9. Patel D, Grimson F, Mihaylova E, Wagner P, Warren J, van Engen A, et al. Use of external comparators for health technology assessment submissions based on single-arm trials. Value Health. 2021;24(8):1118–25.

10. Dahabreh IJ, Hernán MA. Extending inferences from a randomized trial to a target population. Eur J Epidemiol. 2019;34(8):719–22.

11. Jonsson P, Mpofu P, Kent S, Copeland A, Groves B, Bargo D, et al. PCN170 can early US adoption of cancer drugs inform HTA decision-making? Value Health. 2021;24:S51.

12. Hogervorst MA, Pontén J, Vreman RA, Mantel-Teeuwisse AK, Goettsch WG. Real world data in health technology assessment of complex health technologies. Front Pharmacol. 2022;13: 837302.

13. Institute for Quality and Efficiency in Health Care. Concepts for the generation of routine practice data and their analysis for the benefit assessment of drugs according to §35a Social Code Book V (SGB V). Institute for Quality and Efficiency in Health Care; 2020. London.

14. Degtiar I, Rose S. A review of generalizability and transportability. Ann Rev Stat Appl. 2023;10(1):501–24.

15. Dahabreh IJ, Robertson SE, Steingrimsson JA, Stuart EA, Hernán MA. Extending inferences from a randomized trial to a new target population. Stat Med. 2020;39(14):1999–2014.

16. Josey KP, Berkowitz SA, Ghosh D, Raghavan S. Transporting experimental results with entropy balancing. Stat Med. 2021;40(19):4310–26.

17. Hernán MA, VanderWeele TJ. Compound treatments and transportability of causal inference. Epidemiology. 2011;22(3):368–77.

18. Dahabreh IJ, Robins JM, Haneuse SJA, Saeed I, Robertson SE, Stuart EA, et al. Sensitivity analysis using bias functions for studies extending inferences from a randomized trial to a target population. Stat Med. 2023;42:2029–43.

19. Nguyen TQ, Ackerman B, Schmid I, Cole SR, Stuart EA. Sensitivity analyses for effect modifiers not observed in the target population when generalizing treatment effects from a randomized controlled trial: assumptions, models, effect scales, data scenarios, and implementation details. PLoS One. 2018;13(12): e0208795.

20. Josey KP, Yang F, Ghosh D, Raghavan S. A calibration approach to transportability and data-fusion with observational data. Stat Med. 2022;41:4511–31.

21. Jaksa A, Arena PJ, Chan KKW, Ben-Joseph RH, Jónsson P, Campbell UB. Transferability of real-world data across borders for regulatory and health technology assessment decision-making. Front Med (Lausanne). 2022;9:1073678.

22. Beal B, Altomare I, Ray J, Bargo D, Adamson B. HTA3 passport for travel: proposed framework for transportability of oncology real world evidence. Value Health. 2022;25(7):S504.

23. Tennant PWG, Murray EJ, Arnold KF, Berrie L, Fox MP, Gadd SC, et al. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. Int J Epidemiol. 2021;50(2):620–32.

24. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. Am J Epidemiol. 2016;183(8):758–64.

25. EUnetHTA. Vision paper on the sustainable availability of the proposed Registry Evaluation and Quality Standards Tool (REQueST). 2019. https://www.eunethta.eu/wpcontent/uploads/2019/10/EUnetHTAJA3_Vision_paper-on-REQueST-tool.pdf. Accessed 19 Oct 2023.

26. National Institute for Health and Care Excellence. NICE real-world evidence framework. Appendix 1: Data Suitability Assessment Tool (DataSAT). National Institute for Health and Care Excellence; 2022.

27. Tipton E. Improving generalizations from experiments using propensity score subclassification: assumptions, properties, and contexts. J Educ Behav Stat. 2013;38(3):239–66.

28. Boulesteix AL, Lauer S, Eugster MJ. A plea for neutral comparison studies in computational sciences. PLoS One. 2013;8(4): e61562.

29. Heinze G, Boulesteix AL, Kammer M, Morris TP, White IR. Phases of methodological research in biostatistics: building the evidence base for new methods. Biom J. 2023;3: e2200222.

30. Phillippo DM, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. Methods for population-adjusted indirect comparisons in health technology appraisal. Med Decis Making. 2018;38(2):200–11.

31. The Economist Impact. Value of real-world evidence in health technology assessment: lost in translation? 2022. https://impact.economist.com/projects/rwe-in-hta/. Accessed 19 Oct 2023.

32. Padula WV, Kreif N, Vanness DJ, Adamson B, Rueda JD, Felizzi F, et al. Machine learning methods in health economics and outcomes research: the PALISADE checklist: a good practices report of an ISPOR Task Force. Value Health. 2022;25(7):1063–80.

33. Jackson CH, Thompson SG, Sharples LD. Accounting for uncertainty in health economic decision models by using model averaging. J R Stat Soc Ser A Stat Soc. 2009;172(2):383–404.

34. Leahy TP, Duffield S, Kent S, Sammon C, Tzelis D, Ray J, et al. Application of quantitative bias analysis for unmeasured confounding in cost-effectiveness modelling. J Comp Eff Res. 2022;11(12):861–70.

35. Leahy TP, Kent S, Sammon C, Groenwold RH, Grieve R, Ramagopalan S, et al. Unmeasured confounding in nonrandomized studies: quantitative bias analysis in health technology assessment. J Comp Eff Res. 2022;11(12):851–9.

36. Happich M, Brnabic A, Faries D, Abrams K, Winfree KB, Girvan A, et al. Reweighting randomized controlled trial evidence to better reflect real life: a case study of the Innovative Medicines Initiative. Clin Pharmacol Ther. 2020;108(4):817–25.

37. Ramagopalan SV, Popat S, Gupta A, Boyne DJ, Lockhart A, Hsu G, et al. Transportability of overall survival estimates from US to Canadian patients with advanced non-small cell lung cancer with implications for regulatory and health technology assessment. JAMA Netw Open. 2022;5(11): e2239874.

38. Franklin JM, Patorno E, Desai RJ, Glynn RJ, Martin D, Quinto K, et al. Emulating randomized clinical trials with nonrandomized real-world evidence studies: first results from the RCT DUPLICATE Initiative. Circulation. 2021;143(10):1002–13.