Electronic Supplementary Material

Title: Transporting Comparative Effectiveness Evidence Between Countries: Considerations for Health Technology Assessment

Journal: PharmacoEconomics

Authors and Affiliations

Alex J Turner[a], Cormac Sammon[a], Nick Latimer[b,c], Blythe Adamson[d], Brennan Beal[d], Vivek Subbiah[e], Keith R Abrams[f,g], Joshua Ray[h*]

[a] *Putnam PHMR, London, UK*

[b] *School of Health and Related Research, University of Sheffield, Sheffield, UK*

[c] *Delta Hat, Nottingham, UK*

[d] *Flatiron Health, New York, NY, USA*

[e] *Sarah Cannon Research Institute, Nashville, TN, USA*

[f] *Department of Statistics, University of Warwick, Coventry, UK*

[g] *Centre for Health Economics, University of York, York, UK*

[h] *Global Access, F. Hoffmann-La Roche Ltd, Grenzacherstrasse, Basel, Switzerland*

** Corresponding author available at: joshua.ray@roche.com*

Supplementary Appendix: Overview of methods for transportability

Methods for transportability cover three key areas: (1) the identification and measurement of transportability; (2) methods to correct for a lack of transportability; and (3) sensitivity analysis to assess residual biases. Methods in each category have been detailed elsewhere [1, 2] and so are summarised below.

We focus on cases where IPD is available for the study sample and where either IPD or summary data is available from a target sample. Methods applicable where only summary data is available for both the study and target samples, such as meta-analytic techniques, are highlighted elsewhere [2].

Identification and measurement of transportability

The available approach to identification and measurement is dependent on whether outcome data is available in the target sample.

Outcome data not available in the target sample

Where outcome data is not available in the target sample, the requirements for the validity of the *conditional mean difference exchangeability of study selection* assumption are tested sequentially 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 are effect modifiers (i.e. explain heterogeneity in treatment effects) [3].

Where only summary data is available, simple methods for assessing similarity between characteristics of study and target samples must be used. These include qualitative comparisons of average characteristics [4, 5], and a generalisation score which takes an average of characteristic-specific ratios of average values [6]. However, they do not account for differences in dispersion and for the joint distribution of covariates. If variances are also reported in summary data, dispersion can in part be accounted for by calculating standardised mean differences (SMDs), with values above 0.1 indicative of extrapolation and reliance on correct model specification to correct for a lack of transportability [7]. With variance data, univariate tests of average differences can be conducted, with appropriate adjustment for multiple hypothesis testing [8].

With IPD from the target sample, multiple hypothesis testing can be avoided whilst also accounting for the joint distribution of covariates, by comparing distributions of propensity scores, generated as predicted probabilities of study sample membership from a logistic regression of study sample membership on baseline characteristics. SMDs of propensity scores greater than 0.1 can again be used to identify a lack of transportability [9].

To identify potential violations of the *positivity of selection* assumption (i.e. a lack of overlap), the proportion of target sample individuals with scores outside of the 5th and 95th percentiles can be calculated [7]. Comparisons of the distributions and cumulative densities of propensity scores can also be conducted using distance metrics such as Q-Q plots and the Levy distance and also the generalizability index, with an index score based proportion of study and target sample individuals in propensity score strata of <0.5 (on a scale of 0 to 1) being indicative of poor transportability [10].

Specifically for RCTs, the proportion of the target sample who meet the inclusion and exclusion criteria of the trial can also be examined [11-13], and bias scores can be created based on whether high risk individuals are included/excluded [14].

Multiple methods are available to test for effect modification. The most common test examines the statistical significance of interaction-terms between treatment assignment and potential effect modifiers. This requires assumptions regarding the functional form of effect modification e.g., whether effect modification is linear and additive [15, 16]. However, in small samples and when multiple effect modifiers are being tested concurrently, these tests may have limited power, although power issues may be less problematic if an RWD source is used as the study sample since sample sizes are often larger [17]. Solutions to power issues include the use of sequential tests for treatment-covariate interactions [18], and/or use of thresholds for clinical significance rather than statistical significance [19]. Where potential effect modifiers and the functional form of effect modification are unknown, machine learning approaches such as Bayesian additive regression trees (BART) and other classification and regression tree (CART) variants, as well as flexible parametric models (e.g. spline-based approaches), may be useful [20-26].

It must be noted that this approach will only be able to examine whether a lack of transportability is being driven by observed characteristics, and the absence of differences in characteristics between target and study samples and/or identifying no predictors of treatment effect heterogeneity will not rule out unobserved effect modification.

Outcome data available in the target sample

Where outcome data on patients receiving the treatments examined in the study sample are available in the target sample, unobserved effect modification can be identified by observing outcome differences between study and target samples after propensity scores are used to weight samples to ensure similarity in observed effect modifiers [9]. Formal equivalence test for this are available in a generalizability setting [27]. Outcome regressions can also be used, with outcomes of study and target sample individuals receiving the same treatment compared after conditioning on measured effect modifiers [2]. Alternatively, differences between study and target regression coefficients or between baseline hazards in a Cox regression can be assessed [28].

We note that in many cases where transportability methods will be applied, outcomes for only the comparator treatment would be available in the target sample, meaning these outcome-based tests cannot confirm a lack of unobserved effect modification for the experimental treatment.

Methods to correct for transportability bias to estimate target population average treatment effects

To mitigate transportability bias, methods aim to correct for differences in the distribution of effect modifiers between study and target samples. Available approaches are dependent on the availability of individual patient data (IPD) or aggregate data on the target sample.

All approaches are based on the assumption on no unmeasured effect modification.

IPD available on the target sample

Where IPD is available on the target sample, methods typically fall into one of three classes: (1) outcome modelling and standardisation, (2) study sample participation modelling with weighting; and (3) doubly-robust (DR) methods which combine matching/weighting with outcome modelling.

Dahabreh et al. review methods to correct for transportability of the effects of time-fixed treatments from RCTs where IPD on outcomes and baseline covariates is available from the study sample, but only IPD on baseline covariates is available from the target sample [29].

Outcome-model-based methods

Outcome-model based methods use outcome regressions fit using data for individuals in the study sample to estimate expectations of the outcome (using G-estimation) and then standardize (average) the model predictions with respect to the distribution of characteristics of the target sample. Differences in conditional means for potential outcomes of the treatment and comparator groups are then used to estimate the PATE. To account for all possible interaction between the treatment and covariates, separate outcome models are typically estimated for each treatment group of the study sample. A range of parametric models can be used depending on the nature of the outcomes. Machine learning methods, for example tree-based methods such as BART, have also been used. Assuming the outcome model is correctly specified, these methods produce a consistent estimate of the PATE.

Weighting-based methods

Weighting-based approaches are commonly based on a model for the probability of study sample participation. After appending data from the study and target samples, a parametric model (typically logistic regression) is used to generate predicted probabilities of participation based on individual characteristics. For transportability, these predictions are used to derive inverse odds (IO) of participation weights, with treatment effects in the study sample re-estimated after applying the weights to estimate the PATE [30]. When the probability of participation model is correctly specified and there is sufficient overlap in study and target sample characteristics, the IO weighting estimator provides consistent estimates of the PATE.

A limitation of weighting approaches is that they produce unstable results and estimates with a large variance in the presence of extreme weights [31]. Plotting of the distribution of the estimated probabilities of study sample participation and subsequent IO weights should be conducted to visually inspect spread and identify extreme values. This can be combined with diagnostics for positivity violations [32], including a check that sample averages of IO weights multiplied by the study sample participation indicator, divided by the proportion of individuals of the combined sample being in the target sample should be approximately equal to one. Values different from one suggest positivity violations or model misspecification. Note that these checks are recommended even when using the outcome model-based estimators, because estimated probabilities of study sample participation near zero indicate possible positivity violations [1].

The choice of characteristics to include in participation models should be guided by a trade-off between adding covariates that may provide bias reduction but, if unnecessary, can decrease precision and increase the chance of extreme weights [33]. The former is typically of greater concern [1, 34]. Trimming of patients with extreme weights can be used, but that changes the population of interest. An alternative estimator generates IO weights which are normalised to sum to the number of individuals in the target population. This estimator ensures estimated outcomes lie within the support of the outcome variable, and has better finite performance when weights are highly variable [1].

With all weighting methods, balance checking should be conducted to ensure characteristics of study and target samples are similar post-weighting, with an SMD<0.1 conventionally used as a threshold [34].

Propensity score stratification

Propensity score stratification can also be used, where a combined study and target sample is created and split into strata based on the predicted probability of study sample participation. Treatment effects are then estimated in each strata using the study sample data, before the PATE is estimated by weighting stratum-specific treatment effects using the proportion of the target population in each stratum [35]. Stratification can be more stable that IO weighting estimators but involve a trade-off with the choice of the number of

strata: too few strata risks residual unobserved effect modification, whilst too many strata can lead to small numbers in each strata and therefore unstable estimates [7, 34].

Doubly-robust methods

Dahabreh et al. present three doubly-robust estimators that are easily implemented in statistical software [29]. Doubly-robust methods provide consistent estimates of the PATE if *either* the outcome-model or propensity score model is correctly specified. Augmented inverse probability of participation weighting (A-IPPW) involves estimating the PATE by combining predictions from both a model for conditional expectations of the outcome, and a model for the probability of study sample membership. A-IPPW using normalised IO weights described above can be used to ensure bounded outcome estimates. A final A-IPPW model involves fitting a model for the outcome conditional on covariates among trial participants using a weighted regression with weights derived from propensity score models, and then averaging the predicted values from this regression over the covariate distribution of the target population.

Existing reviews [2] discuss other doubly-robust methods, including targeted maximum likelihood estimation (TMLE), and augmented calibration weighting methods with sampling weights derived through other approaches that are not based on propensity scores [36, 37]. Josey et al use entropy balancing which is a convex optimization technique that derives weights that automatically satisfy a set of balance constraints [37]. Unlike propensity score methods, this means weighting ensures exact balance on characteristics between study and target samples.

For methods of all types, confidence intervals for transported treatment effects can be obtained using the “sandwich" approach or using bootstrap methods [1].

Performance of methods

Simulation studies comparing the performance of these methods have identified that outcome-based approaches are particularly useful when the outcome is common but when selection into the study is rare, and can yield better precision than weighting-based methods because they can adjust both for confounders, effect-modifiers, and factors only predictive of the outcome, thus decreasing variance in the estimate [2]. A further benefit of these methods is that predictions in the target sample can be made even in cases of non-overlap in the characteristics of study and target samples, although the predictions rely solely on extrapolation in these cases [38, 39] and often adjustments in the variance are not made to reflect this additional uncertainty [2].

Weighting approaches may be preferable to outcome-based approaches in the case of multiple outcomes, as one set of weights can be used to estimate treatment effects for all outcomes.

Doubly-robust estimators typically produce estimates that are more precise than those of weighting estimators and nearly as precise as those of the outcome model-based estimator while giving investigators two opportunities for valid inference. However, in some cases, when misspecification of the outcome model is combined with highly variable weights, doubly robust estimators can perform worse than other estimators that use the same misspecified outcome model [40, 41]. The diagnostic checks relating to weights described above should therefore also be conducted when using doubly-robust methods [1].

Use of transportability methods alongside RWD

Many of the methods described above have been developed for transporting treatment effects from RCTs, but many will be applicable to RWD studies. Prior to the application of methods to identify and correct for transportability, study sample data can be pre-processed using weighting or matching to ensure balance on observed confounders [42]. New methods are available that generate a single set of weights to simultaneously correct for confounding and sample selection. For example, Josey et al. [43] extend their previous entropy balancing approach [37] for this purpose. Each treatment group in the study sample could also be matched or weighted separately to the target sample [44], although this risks generating imbalance in characteristics between treatment arms [42].

Only aggregate data available on the target sample

Where only summary-level (aggregate) data is available on characteristics of the target population, a restricted set of methods are available. Direct standardisation is an identical approach to propensity score stratification, but where strata are instead defined by all combinations of values of the effect modifiers. This becomes infeasible with many effect modifiers and where effect modifiers are continuous, as empty strata become common. Continuous effect modifiers can be aggregated to reduce this issue but this increases the risk of unobserved effect modification. For higher dimensional settings or with continuous covariates, more flexible nonparametric approaches can be applied, such as maximum entropy weighting can be used [27].

If joint covariate distributions are available from the target population, then an adjusted version of the IO weighting method can be applied [33]. Where only univariate moments are available, methods typically used to conduct population-adjusted indirect comparisons of treatments that haven't been directly compared in a trial but have each been compared to a common comparator in other trials. The methods were designed for cases where IPD is available from one trial, but where only aggregate data is available from another. Matching adjusted indirect comparison (MAIC) and simulated treatment comparison (STC) are commonly-used methods. MAIC is derived from IO weighting but, given a lack of IPD in the target sample, weights are estimated using the method of moments [45]. STC is a modification of the outcome-based models above, but where mean covariate values are substituted into the outcome model fitted in the study sample [46]. Phillippo et al. describe these methods in more detail and discuss when each method is most appropriate [47].

Sensitivity analyses

A key assumption of all methods is no unmeasured effect modification. If unmeasured effect modification is suspected, sensitivity analysis can be performed to assess the extent to which it is likely to impact results [42, 48].

A simple approach to calculating bias-adjusted treatment effects is available in cases when a single effect modifier is observed in the study sample but not in the target population [42]. Bias in the PATE is formalised as the product of the coefficient of the treatment-effect modifier interaction and the average value of the effect modifier in the target population (the "sensitivity parameter"). After estimating the treatment-effect modifier interaction coefficient using data in the study sample, the sensitivity parameter is varied to explore the impact on treatment effects. This approach is applicable with outcome-based models or doubly-robust methods and supports outcome models including linear models for additive effects, and models with log or logit link for multiplicative effects. A more general approach which uses bias functions to directly parametrise unobserved effect modification and can be used in cases where there are numerous effect modifiers that are unmeasured in both the study and target sample [49].

References

1. Dahabreh IJ, Robertson SE, Steingrimsson JA, Stuart EA, Hernán MA. Extending inferences from a randomized trial to a new target population. Stat Med. 2020 Jun 30;39(14):1999-2014.
2. Degtiar I, Rose S. A review of generalizability and transportability. Annual Review of Statistics and Its Application,. 2021;10.
3. Olsen RB, Orr LL, Bell SH, Stuart EA. External Validity in Policy Evaluations that Choose Sites Purposively. J Policy Anal Manage. 2013 Winter;32(1):107-21.
4. Beal B, Altomare I, Ray J, Bargo D, Adamson B. HTA3 Passport for Travel: Proposed Framework for Transportability of Oncology Real World Evidence. Value in Health. 2022;25(7):S504.
5. Ma X, Long L, Moon S, Adamson BJ, Baxi SS. Comparison of population characteristics in real-world clinical oncology databases in the US: Flatiron Health, SEER, and NPCR. Medrxiv. 2020.
6. Cahan A, Cahan S, Cimino JJ. Computer-aided assessment of the generalizability of clinical trial results. Int J Med Inform. 2017 Mar;99:60-6.
7. Tipton E, Hallberg K, Hedges LV, Chan W. Implications of Small Samples for Generalization: Adjustments and Rules of Thumb. Eval Rev. 2017 Oct;41(5):472-505.
8. Greenhouse JB, Kaizar EE, Kelleher K, Seltman H, Gardner W. Generalizing from clinical trial data: a case study. The risk of suicidality among pediatric antidepressant users. Stat Med. 2008 May 20;27(11):1801-13.
9. Stuart EA, Cole SR, Bradshaw CP, Leaf PJ. The use of propensity scores to assess the generalizability of results from randomized trials. J R Stat Soc Ser A Stat Soc. 2001 Apr 1;174(2):369-86.
10. Tipton E. How generalizable is your experiment? An index for comparing experimental samples and populations. Journal of Educational and Behavioral Statistics. 2014;39(6):478-501.
11. Sen A, Chakrabarti S, Goldstein A, Wang S, Ryan PB, Weng C. GIST 2.0: A scalable multi-trait metric for quantifying population representativeness of individual clinical studies. J Biomed Inform. 2016 Oct;63:325-36.
12. He Z, Ryan P, Hoxha J, Wang S, Carini S, Sim I, et al. Multivariate analysis of the population representativeness of related clinical studies. J Biomed Inform. 2016 Apr;60:66-76.
13. Weng C, Li Y, Ryan P, Zhang Y, Liu F, Gao J, et al. A distribution-based method for assessing the differences between clinical trial target populations and patient populations in electronic health records. Appl Clin Inform. 2014;5(2):463-79.
14. Weisberg HI, Hayden VC, Pontes VP. Selection criteria and generalizability within the counterfactual framework: explaining the paradox of antidepressant-induced suicidality? Clin Trials. 2009 Apr;6(2):109-18.
15. Fang A. 10 Things to Know about Heterogeneous Treatment Effects. 2017.
16. Gabler NB, Duan N, Liao D, Elmore JG, Ganiats TG, Kravitz RL. Dealing with heterogeneity of treatment effects: is the literature up to the challenge? Trials. 2009 Jun 19;10:43.
17. Hernán MA, Alonso A, Logan R, Grodstein F, Michels KB, Willett WC, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. Epidemiology. 2008 Nov;19(6):766-79.
18. Qian M, Chakraborty B, Maiti R, Cheung YK. A sequential significance test for treatment by covariate interactions. arXiv preprint arXiv:190108738. 2019.
19. Gail M, Simon R. Testing for qualitative interactions between treatment effects and patient subsets. Biometrics. 1985 Jun;41(2):361-72.
20. Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. Proc Natl Acad Sci U S A. 2016 Jul 5;113(27):7353-60.
21. Green DP, Kern HL. Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. Public opinion quarterly. 2012;76(3):491-511.

22. Lipkovich I, Dmitrienko A, Denne J, Enas G. Subgroup identification based on differential effect search--a recursive partitioning method for establishing response to treatment in patient subpopulations. Stat Med. 2011 Sep 20;30(21):2601-21.
23. Su X, Tsai C-L, Wang H, Nickerson DM, Li B. Subgroup analysis via recursive partitioning. Journal of Machine Learning Research. 2009;10(2).
24. Su X, Zhou T, Yan X, Fan J, Yang S. Interaction trees with censored survival data. Int J Biostat. 2008 Jan 28;4(1):Article 2.
25. Austin PC. A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. Stat Med. 2007 Jul 10;26(15):2937-57.
26. Lambert PC, Smith LK, Jones DR, Botha JL. Additive and multiplicative covariate regression models for relative survival incorporating fractional polynomials for time-dependent effects. Stat Med. 2005 Dec 30;24(24):3871-85.
27. Hartman E, Grieve R, Ramsahai R, Sekhon JS. From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2015;178(3):757-78.
28. Pan Q, Schaubel DE. Proportional hazards models based on biased samples and estimated selection probabilities. Canadian Journal of Statistics. 2008;36(1):111-27.
29. Dahabreh IJ, Robertson S, Stuart EA, Hernán MA. Extending inferences from randomized participants to all eligible individuals using trials nested within cohort studies. ADVANCES IN WATER RESOURCES, SOUTHHAMPTON, UK, COMPUTATIONAL MECHANICS PUBLICATIONS. 2017.
30. Westreich D, Edwards JK, Lesko CR, Stuart E, Cole SR. Transportability of Trial Results Using Inverse Odds of Sampling Weights. Am J Epidemiol. 2017 Oct 15;186(8):1010-4.
31. Tipton E. Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. Journal of Educational and Behavioral Statistics. 2013;38(3):239-66.
32. Petersen ML, Porter KE, Gruber S, Wang Y, van der Laan MJ. Diagnosing and responding to violations in the positivity assumption. Stat Methods Med Res. 2012 Feb;21(1):31-54.
33. Nie L, Zhang Z, Rubin D, Chu J. Likelihood reweighting methods to reduce potential bias in noninferiority trials which rely on historical data to make inference. The Annals of Applied Statistics. 2013;7(3):1796-813.
34. Stuart EA. Matching methods for causal inference: A review and a look forward. Stat Sci. 2010 Feb 1;25(1):1-21.
35. O'Muircheartaigh C, Hedges LV. Generalizing from unrepresentative experiments: a stratified propensity score approach. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2014;63(2):195-210.
36. Dong N, Stuart EA, Lenis D, Quynh Nguyen T. Using Propensity Score Analysis of Survey Data to Estimate Population Average Treatment Effects: A Case Study Comparing Different Methods. Eval Rev. 2020 Feb;44(1):84-108.
37. Josey KP, Berkowitz SA, Ghosh D, Raghavan S. Transporting experimental results with entropy balancing. Stat Med. 2021 Aug 30;40(19):4310-26.
38. Attanasio O, Meghir C, Szekely M. Using randomised experiments and structural models for'scaling up': evidence from the PROGRESA evaluation. 2003.
39. Kern HL, Stuart EA, Hill J, Green DP. Assessing methods for generalizing experimental impact estimates to target populations. J Res Educ Eff. 2016;9(1):103-27.
40. Kang JD, Schafer JL. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Statistical science. 2007;22(4):523-39.
41. Robins J, Sued M, Lei-Gomez Q, Rotnitzky A. Comment: Performance of double-robust estimators when" inverse probability" weights are highly variable. Statistical Science. 2007;22(4):544-59.
42. 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.
43. Josey KP, Yang F, Ghosh D, Raghavan S. A calibration approach to transportability and data-fusion with observational data. Stat Med. 2022 Jul 18.
44. Bennett M, Vielma JP, Zubizarreta JR. Building representative matched samples with multi-valued treatments in large observational studies. Journal of Computational and Graphical Statistics. 2020;29(4):744-57.
45. Signorovitch JE, Wu EQ, Yu AP, Gerrits CM, Kantor E, Bao Y, et al. Comparative effectiveness without head-to-head trials: a method for matching-adjusted indirect comparisons applied to psoriasis treatment with adalimumab or etanercept. Pharmacoeconomics. 2010;28(10):935-45.
46. Ishak KJ, Proskorovsky I, Benedict A. Simulation and matching-based approaches for indirect comparison of treatments. Pharmacoeconomics. 2015 Jun;33(6):537-49.
47. 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 Feb;38(2):200-11.
48. Dahabreh I, Robins J, Haneuse S, Saeed I, Robertson S, Stuart E, et al. Sensitivity analysis using bias functions for studies extending inferences from a randomized trial to a target population. arXiv preprint arXiv:190108738. 2019.
49. 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 Feb 27.