



Transportability of the comparative effect of finerenone for the treatment of symptomatic chronic heart failure with left ventricular ejection fraction of $\geq 40\%$: insights from the FINEARTS-HF trial

Journal of **Comparative Effectiveness Research**

Alex J Turner¹ , Claire Leboucher^{*,2,3} , Cécile Remuzat², Yik Ming Fung⁴, Elena Pessina⁵ & Kerstin Folkerts⁶

¹ Health Organisation, Policy & Economics (HOPE), School of Health Sciences, University of Manchester, Manchester, M13 9PL, UK

² Putnam, 207 rue de Bercy, Paris, France

³ CERESS-Health Services Research & Quality of Life Center, Aix-Marseille University, Marseille, France

⁴ Bayer AG, Real-World Data Science & AI, Berlin, Germany

⁵ Bayer AG, Real World Evidence Center of Excellence, Milan, Italy

⁶ Bayer AG, HEOR & HTA, Wuppertal, Germany

*Author for correspondence: claire.leboucher@putassoc.com

Aim: Global randomized controlled trials (RCTs) are used to inform reimbursement decisions in multiple markets, meaning the transportability of findings from these RCTs to multiple country-specific populations is critical for evidence-based decision making. This study evaluated the transportability of the FINEARTS-HF trial, which assessed finerenone in patients with heart failure (HF) and left ventricular ejection fraction $\geq 40\%$, to a real-world US population. **Materials & methods:** A three-phase transportability assessment was conducted. First, potential effect modifiers were identified through systematic literature review and confirmed through interaction analyses using data from FINEARTS-HF. Second, representativeness was assessed by comparing the distribution of effect modifiers in the trial and a population derived from a US electronic health record dataset. Third, direct proxy tests explored heterogeneity of treatment effects in FINEARTS-HF between US and non-US patients. **Results:** Effect modifier analysis identified that treatment effects of finerenone were homogeneous across subgroups, with limited evidence of effect modification. Comparisons with the US target population indicated overall good alignment across key characteristics, with only modest imbalances, suggesting that the trial results may underestimate positive treatment effects in the US target population (i.e., effect on the primary outcome over 0.84 [0.74; 0.95]). Direct proxy tests found no statistically significant regional heterogeneity in treatment effects. **Conclusion:** This study provides a structured assessment of finerenone trial transportability. Findings support the robustness of FINEARTS-HF results for US clinical practice and are not impacted by the prevalence of SGLT2-is use. The risk of transportability bias is likely to be low.

Plain language summary: Can results from a global heart failure trial be used in different countries?

What is this article about? Clinical trials are often run in many countries. However, patients from trials may not always look the same as patients treated in everyday clinical practice in a specific country. This raises an important question: can results from a trial be reliably used to inform decisions at a country level? This article examines whether the results of the FINEARTS-HF trial, evaluating finerenone in patients with chronic heart failure and a left ventricular ejection fraction $\geq 40\%$, are applicable to patients from the US.

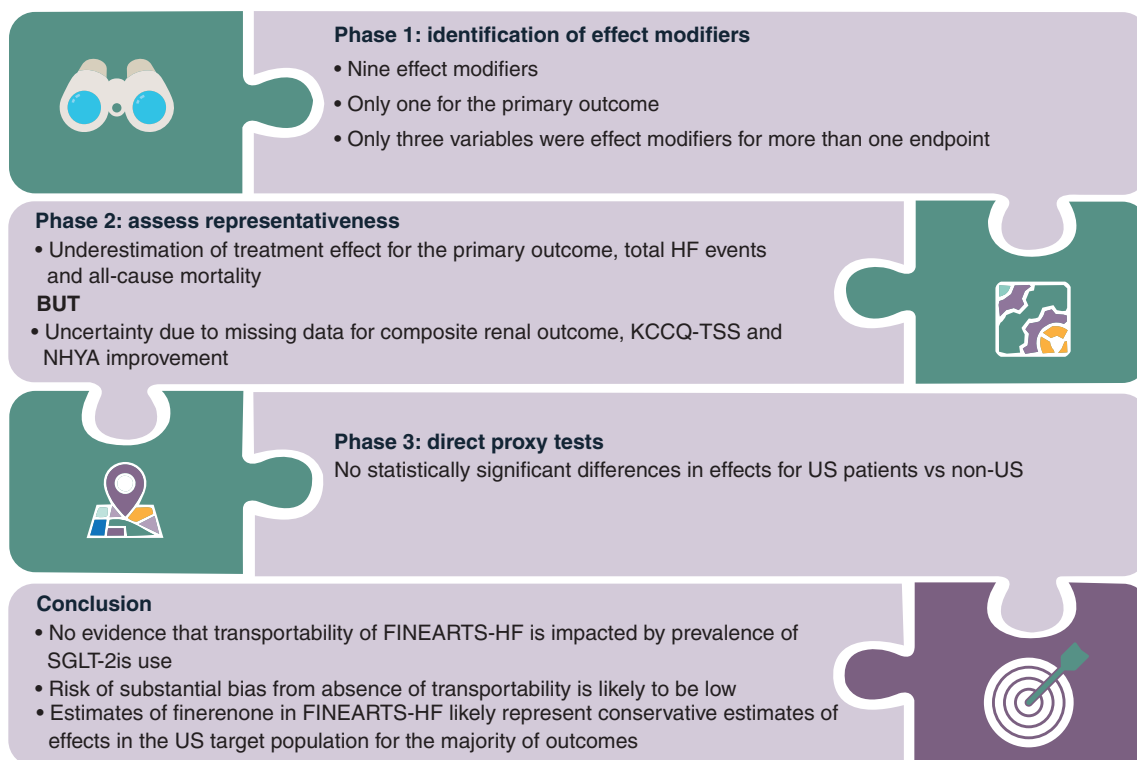
How was this assessed? The researchers used a step-by-step approach. First, they identified patient characteristics that could influence how well the treatment works. Next, they compared these characteristics between people in the trial and patients treated in routine care in the US. Finally, they checked whether the treatment worked differently in patients enrolled in the US compared with those enrolled in other countries.

What were the results? The people included in the trial were similar to patients treated in routine practice. There was no evidence that finerenone worked differently in one country compared with another. The small differences observed suggest that the trial results may slightly underestimate how much benefit patients could experience.



What do the results mean? Using the US as an example, the study shows that results from the FINEARTS-HF trial can be applied with confidence across different healthcare settings. This approach can help decision-makers judge whether results from global clinical trials are relevant in their own countries.

Graphical abstract:



First draft submitted: 9 January 2026; Accepted for publication: 31 March 2026; Published online: 6 May 2026

Keywords: external validity • randomized controlled trial • real-world evidence • transportability

Background & objectives

As randomization eliminates nonchance confounding and increases the likelihood of internal validity, health technology assessment (HTA) agencies commonly express a preference for evidence from randomized controlled trials (RCTs) to guide decision making [1,2,3]. However, while RCTs are generally regarded as the gold standard for establishing internal validity, their external validity can be limited, i.e., that treatment effects from a study represent unbiased estimates of efficacy in a decision-maker's target population(s) of interest [4]. Strict inclusion/exclusion criteria, targeting patients with the highest risk of an outcome, and participation determined by nonrandom self-selection [4], can result in RCT samples differing in characteristics from patients who will receive a treatment in clinical practice. In addition, where RCTs are conducted in indications where treatment options are rapidly evolving, the time taken to recruit to and conduct RCTs can mean that concomitant medications taken by trial participants can deviate from current clinical practice [5]. If characteristics that distinguish trial samples from target populations act as effect modifiers, these differences will compromise the external validity.

This is particularly important when pivotal evidence for new medicinal products comes primarily from a single, global RCT, which is then used as the basis of HTA submission in multiple jurisdictions. In such situations, external validity with respect to multiple target populations is needed, requiring consistency in effect modifiers with multiple populations which may differ in patient characteristics and/or treatment patterns [6].

When assessing the external validity of RCTs in multiple HTA jurisdictions, external validity can be defined as transportability, where at least a portion of the study sample is drawn from data outside of the target population [7].

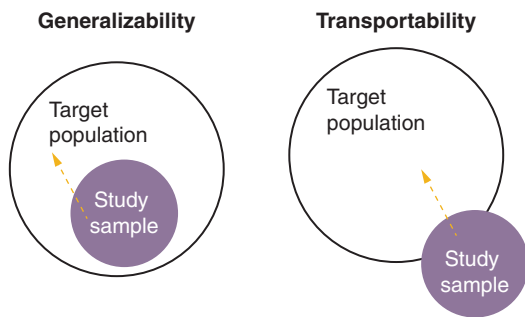


Figure 1. Generalizability and transportability concepts.

This is distinct from generalizability, where the study sample is a subsample of the target population [8] as shown in Figure 1.

Several studies have summarized methods to measure transportability and, where transportability is a concern, adjust/re-weight trial evidence to better reflect treatment effects that would be observed in the decision-maker's target population [7,8,9,10,11,12,13]. Other studies have discussed the relevance of these approaches in a HTA setting [6,14]. Recent reviews have described applications of these methods [15,16]; however, none have specifically focused on assessing the transportability of a global RCT to multiple countries.

FINEARTS-HF was conducted between September 2020 and June 2024. Its findings are expected to inform HTA submissions across North America, Europe, Asia and Latin America. Concerns have been raised regarding the representativeness of the usual therapy and the transportability of the trial results to patients with heart failure (HF) and left ventricular ejection fraction (LVEF) $\geq 40\%$. In particular, the regulatory approvals of Sodium-glucose co-transporter-2 inhibitors (SGLT-2is) and their subsequent incorporation and recommendation into treatment guidelines (refer to [Supplementary Table 1](#)) led to a concern that patients receiving SGLT-2is may be underrepresented in the study population compared with routine clinical practice. This underrepresentation could affect the transportability of the results if SGLT-2is use acts as an effect modifier [11]. To address this, heterogeneity in treatment effects was evaluated according to baseline SGLT-2is use, as well as other potential effect modifiers previously identified in studies of patients with HF.

The aim of this study was to assess the transportability of the FINEARTS-HF trial, a global RCT assessing the efficacy and safety of finerenone in addition to usual therapy in patients with HF and LVEF $\geq 40\%$ [17]. The US were used as a case study of transportability assessment due to the availability of robust data on the US target population.

Materials & methods

Data sources

Study sample: FINEARTS-HF

FINEARTS-HF was a multicenter, randomized, double-blind, placebo-controlled Phase III clinical trial (ClinicalTrials.gov number: NCT04435626) designed to evaluate the efficacy and safety of finerenone, a nonsteroidal mineralocorticoid receptor antagonist, in patients with HF and LVEF $\geq 40\%$. Patients were randomized in a 1:1 ratio to receive once-daily oral finerenone or placebo in addition to usual therapy [17].

Eligible participants included adults with New York Heart Association (NYHA) class II–IV HF, and LVEF $\geq 40\%$. The trial enrolled a geographically diverse population including patients treated at sites located in 37 countries primarily across North America, Europe, Asia and Latin America [17].

The primary efficacy outcome was a composite outcome of cardiovascular death and total (first and recurrent) HF events (HHF or urgent HF visit). The secondary efficacy outcomes included: the timing and occurrence of total (first and recurrent) HF events, the change from baseline in Kansas City Cardiomyopathy Questionnaire total symptom score (KCCQ-TSS), improvement in NYHA class, a composite renal endpoint (sustained decrease of eGFR $\geq 50\%$ relative to baseline over at least 4 weeks, or sustained eGFR decline to <15 ml/min/1.73 m², or initiation of dialysis or renal transplantation), and the time to all-cause mortality [17].

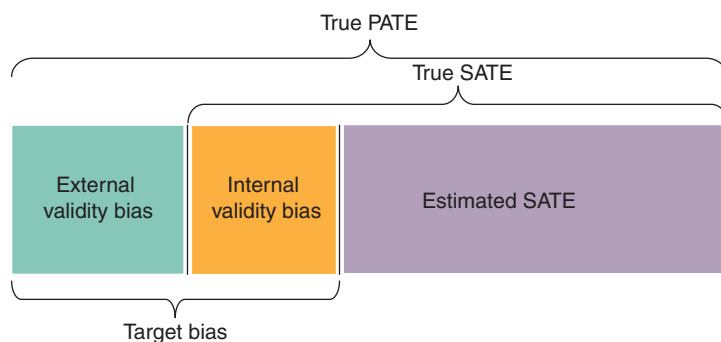


Figure 2. PATE, SATE, internal and external validity.

PATE: Target population average treatment effect; SATE: Sample average treatment effect.

US target sample: TULIP-US

TULIP-US was a retrospective cohort study using observational data derived from electronic health records (EHR) from the US, provided by Optum®. The study included data from 1 January 2013 to the most recent date of available data (2023) [18,19].

The base cohort was composed of newly diagnosed HF patients identified from 2014 to 2023 (end of available data) using International Classification of Diseases (ICD)-9 & ICD-10 diagnostic codes recorded in the inpatient and/or outpatient setting, as well as LVEF measurements, available within structured data and derived from unstructured fields in patients' charts by Optum, using natural language processing. Patients were included if they received an LVEF measurement within ± 90 days of the first observed ICD-9 or ICD-10 code for HF. The index date was defined as either the date of the first observed HF code or the date of the corresponding LVEF value at index, whichever came second. Patients were required to be at least 18 years of age at index, with at least 365 days of baseline observation data. Patients with a pre-existing HF diagnosis recorded during the 365-day pre-index period were excluded. Further cross-sectional analysis of both incident (newly diagnosed) and prevalent HF patients with at least one ICD-9/10 code for HF (within the specified year), at least one LVEF measurement $\geq 40\%$ (within the specified year) and at least 18 years of age on 1 January of the specified year was conducted [18,19].

The study endpoints were patient clinical characteristics (demographics, comorbidities, laboratory values/clinical measurements), treatment characteristics (medication prescriptions, diagnostic measures, clinical procedures, HCRU) and patient outcomes (mortality and other heart/kidney outcomes) [19].

The cross-section for 2023 was chosen as the target sample for this study because both incident and prevalent patients were included (consistent with FINEARTS-HF) and treatment data in the most recent year in the sample most closely reflects current treatment patterns. TULIP-US represented a suitable target sample for the US target population since it is a single-country study capturing patients only treated in the US and was conducted in the target indication (i.e., contains patients and LVEF $\geq 40\%$). Moreover, the Optum EHR database includes both insured and uninsured patients across all age groups, thereby providing a broadly representative sample of the US population. It comprises longitudinal, patient-level EHRs for approximately 100 million individuals receiving care at over 700 hospitals and 7000 clinics across the US.

As the study period covered 2023, all included patients were treated after the date of SGLT-2is approval date in the US which reflects the currently available treatment options (Refer to [Supplementary Table 1](#)). Treatment patterns from this cross-sectional cohort therefore plausibly represented current clinical practice. In addition, many potential effect modifiers with data available in FINEARTS-HF could be observed in Optum ([Supplementary Table 2](#)).

Underlying assumptions to test for transportability

To assess transportability, analysis was conducted to test the key underlying assumptions for transportability. These assumptions determine the degree to which the treatment effect estimated in the trial (the sample average treatment effect; SATE) is likely to deviate from target population average treatment effect (PATE): the average effect of treatment if all individuals in the target population were assigned the treatment ([Figure 2](#)) [11].

Conditional on assumptions regarding internal validity holding, the transportability of an RCT depends on the following assumptions [11]:

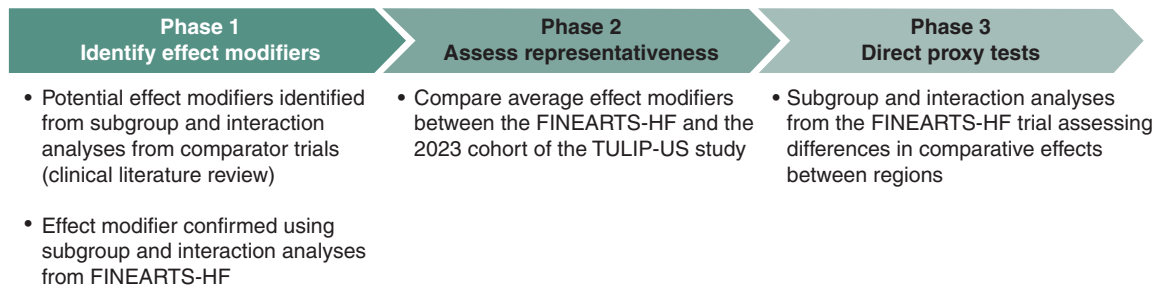


Figure 3. The 3-phase approach to assess transportability of FINEARTS-HF.

- Conditional mean difference exchangeability for study selection 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 are 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 exist 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.
- The Stable Unit Treatment Value Assumption for study selection requires no interference between subjects in the study sample and target sample, and no difference in how outcomes are measured and in the distribution of versions of treatment across the study sample and the target population.

The transportability assessment approach

The approach for assessing transportability in this study involved a three-phase approach aiming to test the plausibility of the conditional mean difference exchangeability assumption (Figure 3).

Phase 1 – identification of effect modifiers

To identify effect modifiers, a systematic literature review (SLR) focusing on HF comparator studies was conducted to identify potential effect modifiers. The Population, Intervention, Comparison, Outcomes and Study (PICOS) are presented in [Supplementary Table 3](#).

Information on p-values from interaction tests was not routinely reported in comparator studies and not formally extracted as part of the clinical SLR. However, where p-values were available these were used as the primary source of evidence for effect modification. Information on potential effect modifiers was also identified based on magnitudes of effect differences between subgroups based on demographic/clinical and treatment characteristics, using the following criteria:

- Comparative effects were statistically significant in one subgroup but not in others, and
- Differences in the magnitude of effects across subgroups were sufficiently large, defined as the ratio of the comparative effects across subgroups (e.g., hazard ratio (HR) HR_2/HR_1) being greater than 1.25.

These are likely to overstate effect modification since limited samples would mean that these differences may not have been statistically significant if results from formal interaction analysis were available. However, for the purpose of identifying potential effect modifiers, these criteria were considered appropriate.

Effect modifiers were confirmed using interaction analyses based on data from FINEARTS-HF which assessed heterogeneity in treatments across subgroups based on each potential effect modifiers and other baseline characteristics collected in the trial ([Supplementary Table 2](#)). These interaction analyses were conducted for each of

the six primary/secondary outcomes (Section: Study sample: FINEARTS-HF). Data from FINEARTS-HF provided the best evidence on effect modification relative to effect modifiers identified in comparator trials, because effect modifiers are typically indication-specific. Within indications, effect modifiers often differ by mechanism of action/treatment class, which has been demonstrated in patients with LVEF $\geq 40\%$ [20].

The statistical methods applied to estimate effect modification were tailored to endpoint type. For time-to-event outcomes, (two-sided) interaction p-values were based on stratified Cox proportional hazards model including identifiers for treatment (finerenone vs placebo), subgroup and subgroup by treatment interaction terms as covariates. For recurrent event outcomes, (two-sided) interaction p-values were based on stratified Anderson–Gill models including identifiers for treatment (finerenone vs placebo), subgroup, and subgroup by treatment interaction terms as covariates. For binary outcomes, (two-sided) interaction p-values were based on logistic regression including identifiers for stratification factors, treatment, subgroup and subgroup by treatment interaction terms as covariates. For change from baseline in KCCQ-TSS, (two-sided) interaction p-values were based on a mixed model for repeated measures including identifiers for baseline, study visit, baseline by visit interactions, stratification factors, treatment, subgroup and subgroup by treatment interaction terms as covariates.

A variable was considered as an effect modifier if the interaction p-value < 0.1 , recognizing restricted power to detect heterogeneity across subgroups at a 5% significance level. No adjustment for multiplicity was performed. Whether variables met the ‘magnitude’ criteria used to identify potential effect modifiers in the SLR was also identified. These criteria differed from those previously used to assess heterogeneity in the FINEARTS-HF results. In FINEARTS-HF, the results consistency across all pre-specified subgroups was carefully checked.

Phase 2 – assess representativeness

To assess the representativeness of FINEARTS-HF, the differences in average levels of effect modifiers between the FINEARTS-HF study sample and the TULIP-US target sample were estimated. Differences were calculated using standardized mean differences (SMDs). When mean and standard deviations were not available, proxies were applied (median was considered as mean and standard deviation as the interquartile range/1.35) [21]. Imbalance was concluded if SMDs exceeded 0.1 [8].

The likely impacts of imbalance in effect modifiers on differences in treatment effects for FINEARTS-HF versus the US target population for each outcome were also computed, considering the direction of effect modifier (i.e., whether presence or increase in a variable inflated or reduced on treatment effects) and the direction of differences in the average value of the effect modifier between the FINEARTS-HF trial and the US target sample.

Phase 3 – direct proxy tests

Transportability can only be evaluated using measured effect modifiers, acknowledging that unobserved effect modification may still be present. When assessing transportability of global RCTs to countries from which a subset of trial participants is sampled, assessing whether the country is an effect modifier in the RCT can serve as a direct proxy test of overall transportability since any differences in effects will be driven by country differences in both observed and unobserved effect modifiers. This assumes that the country-specific trial subsample represents appropriate proxies for the target samples in those countries. Interaction tests were therefore performed using data from FINEARTS-HF exploring heterogeneity in treatment effects between patients treated in the US versus elsewhere.

Missing data

Missing data were handled separately in the trial and target datasets according to the purpose of the analysis. In FINEARTS-HF and TULIP-US, no imputation of missing baseline covariate data was conducted. Patients were either excluded from the analysis, or the extent of missing data was reported. For the representativeness assessment, when missingness in the target sample was substantial for variables identified as effect modifiers, the direction of the potential impact on transportability was considered uncertain and interpreted cautiously rather than inferred from incomplete data.

Results

Phase 1 – effect modifiers identified

Twenty-six trials were identified in the SLR (selection flowchart is in [Supplementary Figure 1](#)). Across these studies, there was some suggestive evidence that comparative effects differed by the amount, type of concomitant

	Primary composite	Total HF events	All-cause mortality	Composite renal	KCCQ-TSS	P(NYHA improvement)
Effect modifiers (n)	1	2	3	3	2	2
List of effect modifiers	• eGFR	• eGFR • Age	• Serum potassium • CRM condition (n) • Beta-blocker use	• LVEF • eGFR • UACR	• NT-proBNP • Beta-blocker use	• NT-proBNP • ACEi or ARB use

Figure 4. Effect modifiers identified.

ACEi: Angiotensin-converting enzyme inhibitor; ARB: Angiotensin II receptor blocker; CRM: Cardio-renal-metabolic conditions; eGFR: Estimated glomerular filtration rate; LVEF: Left ventricular ejection fraction; NT-proBNP: N-terminal pro-B-type natriuretic peptide; UACR: Urinary albumin-to-creatinine ratio.

medication, prevalence of comorbidities or risk-factors relating to other diseases, and other clinical and demographic characteristics. Several potential effect modifiers were identified for at least one trial and are listed in [Supplementary Table 2](#).

Overall, results of the interaction tests and criteria relating to magnitude provided evidence of substantial homogeneity of comparative effects of finerenone versus placebo ([Figure 4](#) & [Supplementary Table 2](#), effect estimates are provided in [Supplementary Table 4](#)). This was particularly true for the primary composite endpoint, where only a single effect modifier was identified based on the interaction tests. Only three variables were deemed effect modifiers for more than one endpoint (eGFR, NT-proBNP and baseline use of beta-blocker). Only nine variables were identified as effect modifiers for at least one outcome.

With respect to concomitant therapies, there was no evidence that baseline SGLT-2is use acts as an effect modifier for any endpoint. This indicates that the comparative effects of finerenone were consistent whether or not concomitant therapies used at baseline includes SGLT-2is. Use of a beta-blocker at baseline modified the effect on all-cause mortality (reduction in mortality risk caused by finerenone was smaller for patients with baseline beta blocker use) and change in KCCQ-TSS (improvement in KCCQ-TSS as a result of finerenone was lower in patients with baseline beta blocker use). Use of ACEi or ARB at baseline modified the effect for NYHA improvement only, where the improvement was higher among patients with no use at baseline.

There was some evidence that kidney function modified effects for a subset of endpoints, where shorter time to a relevant renal event for finerenone versus placebo overall was driven by patients with lower eGFR and higher UACR, and lower eGFR was associated with less favorable effects of finerenone on the primary composite and total HF events (although no association with UACR is found for effects on these endpoints). However, there was little evidence that the majority of comorbidities were effect modifiers, with no differences in comparative effects by presence of atrial fibrillation, diabetes, obesity and high blood pressure. There was some evidence that the comparative effect of finerenone was more favorable in those with lower serum potassium (with the reduction in all-cause mortality only present for this group), although this was not found for other endpoints.

Impact of heart function/HF severity on comparative effects were mixed. There was no evidence of impact of NYHA functional class at baseline and HF duration on effects for any endpoint. LVEF at baseline was associated with effects on the composite renal endpoint (where strong favorable effects were present in the LVEF $< 60\%$ subgroup and strong deleterious effects were present in the LVEF $\geq 60\%$ subgroup, although the results should be interpreted with caution as only 27 events occurred in the $< 60\%$ subgroup). Baseline NT-proBNP above the median was associated with statistically more favorable effects of finerenone on the change from baseline in the KCCQ-TSS and the probability of NYHA improvement.

Finally, the number of CRM conditions was associated with a higher risk of all-cause mortality for finerenone versus placebo for patients with 2 or 3 conditions.

There was little evidence of effect modification by patient demographics, with no difference in comparative effects by sex for any endpoint, and differences in effects by age only observed for total HF events, with favorable reductions driven by patients at or below median age. Magnitudes of effects varied substantially by race, likely driven by very low sample sizes in Black and other ethnicities and reflected in the uncertainty around comparative effects in these subgroups. Large differences in magnitudes could therefore occur by chance.

Phase 2 – representativeness of FINEARTS-HF to the US population

Definitions of characteristics from FINEARTS-HF could be replicated for the majority of variables, as summarized in [Table 1](#), which compares clinical and demographic characteristics between the FINEARTS-HF and TULIP-US populations. Definitions of characteristics from FINEARTS-HF could be identified for the majority of variables. However, an absence of data from Optum, meant that information on frailty and number of CRM conditions could not be derived.

As only a limited number of effect modifiers were identified, their impact on treatment effects is expected to be minimal. [Table 2](#) summarizes the implications of these differences across outcomes. As NT-proBNP, eGFR and betablocker use were effect modifiers of three endpoints and had different distribution between FINEARTS-HF and TULIP-US ([Table 1](#)), they were considered as the most impactful effect modifiers.

For two of six outcomes (primary composite outcome, and total HF events), patient groups with less favorable effects were over-represented in FINEARTS-HF, meaning the treatment effects observed in FINEARTS-HF were likely underestimating the favorable treatment effects if the trial was run in the US target population ([Table 2](#)). For the primary composite outcome, patients with less favorable effects were over-represented in FINEARTS-HF (i.e., patients with lower eGFR). As lower eGFR was associated with less favorable comparative effects of finerenone, the effect of finerenone was likely underestimated in the trial versus in the US target population. The over-representation of patients with lower eGFR had the same direction of impact on total HF events. Consistency of average age across the two samples indicated age would have limited impact on transportability of effects on total age events even if age was an effect modifier.

For all-cause mortality, patient groups with less favorable effects were over-represented in FINEARTS-HF (more patients with beta-blocker use and a marginally higher value of serum potassium in the trial). As lower serum potassium was associated with more favorable comparative effects of finerenone and use of beta-blocker was associated with less favorable comparative effects, the treatment effects observed in FINEARTS-HF were likely underestimating treatment effects if the trial was run in the US target population.

Clinically, the over-representation in FINEARTS-HF of subgroups with less favorable comparative effects suggests that the trial results were likely to be conservative estimates of those likely to be observed in the US population. In other words, the effect of finerenone versus placebo in the US population was likely to be better than the ratio of 0.84 [0.74; 0.95] observed in the trial. All effect estimates are provided in [Supplementary Table 4](#).

As no comparison of TULIP-US and FINEARTS-HF patients on the number of CRM conditions was available, the impact was uncertain.

For the remaining three endpoints (composite renal outcome, KCCQ-TSS and NHYA improvement), due to an important number of missing values (on NT-proBNP and UACR) in the US target sample, assessment of the direction of transportability bias was limited. As these variables were identified as effect modifiers for these endpoints, incomplete ascertainment in the target sample reduced confidence in determining whether FINEARTS-HF over- or under-estimated the effects that would be expected in the US population. This uncertainty should be interpreted as incomplete information rather than evidence against transportability.

Phase 3 – direct proxy tests

Interaction p-values from the interaction tests indicate that there were no statistically significant differences in effects for patients treated in the US versus elsewhere for any endpoint. Based on magnitudes alone, there was no evidence of numeric differences in comparative effects between US and the non-US FINEARTS-HF patients for all outcomes, except for total HF events ([Table 3](#)).

Discussion

This study assessed the transportability of findings from the FINEARTS-HF trial, and to our knowledge, represents the first study to consider the transportability of results from a global RCT to the US population. The study did so by assessing cross-country heterogeneity in treatment effects in FINEARTS-HF and by assessing effect modification within the trial. For a case study country, US, levels of effect modifiers were compared between the trial and an external real-world population reflecting the US target population.

The potential under-use of SGLT-2is as a concomitant medication in FINEARTS-HF participants relative to current use in key markets was the primary risk to the transportability of results from the trial. However, for the US target sample, available data indicated that current use of SGLT-2is is in line with the levels observed in the trial. In addition, strong evidence of homogeneity in treatment effects by SGLT-2is use at baseline was identified,

Table 1. Description of effect modifiers and comparison between FINEARTS-HF and TULIP-US.

Characteristic	Effect Modifier?	FINEARTS-HF		TULIP-US		SMD	Conclusion
		n = 6001		n = 163,843			
Age [†] , years [mean ± SD]	Yes	72	9.6	72.3	12.42	0.02	↔ Consistent
Sex [†] : Women [n (%)]		2731	45.5	81,977	50.03	0.09	↔ Consistent
Race [n(%)]							≈ Equivalent overall
Asian		996	16.6	2791	1.7	0.53	
Black		88	1.5	24,889	15.2	0.51	
Other		165	2.8	9505	5.8	0.15	
White		4735	79.1	126,658	77.3	0.04	
Comorbidities/history [n (%)]							
COPD		770	12.8	47,548	29	0.41	↓ Lower in FINEARTS-HF
Type 2 diabetes [†]		2438	40.6	57,910	35.3	0.11	↑ Higher in FINEARTS-HF
Hypertension		5323	88.7	150,824	92.1	0.11	↓ Marginally lower in FINEARTS-HF
AMI		1539	25.6	38,337	23.4	0.05	↔ Consistent
Stroke		707	11.8	14,771	9	0.09	↔ Consistent
AF		2293	38.2	76,120	46.5	0.17	↓ Lower in FINEARTS-HF
BMI [†] , kg/m ²							≈ Marginal difference
N non missing				159,005	97		
BMI, kg/m ² [mean ± SD]		29.9	6.1	30.8	7.77	0.12	
<18.5 (underweight)		65	1.1				
18.5-<25 (normal weight)		1241	20.7				
Underweight / Normal BMI (BMI < 25)		1306	21.8	36,561	23	0.03	
25-<30 (overweight)		1990	33.2	44,969	28.3	0.11	
Obese (BMI ≥ 30)		2692	44.9	77,475	48.7	0.08	
30-<35 (class I obesity)		1546	25.8				
>35 (class II-III obesity)		1146	19.1				
LVEF [†] , %	Yes						↑ More patients with <60% in FINEARTS-HF ↓ Less patients with ≥60% in FINEARTS-HF
LVEF, % [mean ± SD]		52.6	7.8				
<50%		2172	36.2	36,012	22	0.32	
≥50% to <60%		2674	44.6	60,138	36.7	0.16	
≥60%		1147	19.1	67,693	41.3	0.5	
NT-proBNP [†] , pg/ml	Yes						↑ Higher value in FINEARTS-HF
N non missing				101,792	62.1		
NT-proBNP, pg/ml [median (IQR)]		1401	[449–1946]	986	[333–2798]	0.23	
Systolic blood pressure [†] , mmHg [mean ± SD]		129.4	15.3	129.61	13.99	0.02	↔ Consistent
eGFR [†] , ml/min/1.73 m ²	Yes						↑ More patients with low eGFR in FINEARTS-HF
eGFR, ml/min/1.73 m ² [mean ± SD]		62.1	19.7	62.3	25.3	0.01	
eGFR < 60 ml/min/1.73 m ² [n (%)]		2888	48.1	62,149	37.9	0.21	
Potassium, mmol/L [mean ± SD]	Yes	4.4	0.5	4.2	0.5	0.33	↑ Marginally higher in FINEARTS-HF
UACR category, mg/g [n (%)]	Yes						↓ Less patients with high UACR in FINEARTS-HF
N non missing				18,634	11.4		
UACR, mg/g [median (IQR)]		18	[7–67]	32	[12–144]	0.14	
<30		3511	60.6	8087	43.4	0.35	
30 to <300		1712	29.5	7003	37.6	0.17	
≥300		574	9.9	3544	19	0.26	

[†]Identified in the clinical SLR.

ACEi: Angiotensin-converting enzyme inhibitor; AF: Atrial fibrillation; AMI: Acute myocardial infarction; ARB: Angiotensin II receptor blocker; ARNI: Angiotensin receptor-neprilysin inhibitor; BMI: Body mass index; COPD: Chronic obstructive pulmonary disease; eGFR: Estimated glomerular filtration rate; HF: Heart failure; ICD: International classification of diseases; IQR: Interquartile range; KDIGO: Kidney disease: Improving global outcomes; LVEF: Left ventricular ejection fraction; MRA: Mineralocorticoid receptor antagonist; NT-proBNP: N-terminal pro-B-type natriuretic peptide; NYHA: New York Heart Association (functional class); SD: Standard deviation; SGLT-2is: Sodium-glucose co-transporter-2 inhibitor; UACR: Urinary albumin-to-creatinine ratio.

Table 1. Description of effect modifiers and comparison between FINEARTS-HF and TULIP-US (cont.).

Characteristic	Effect Modifier?	FINEARTS-HF		TULIP-US		SMD	Conclusion
		n = 6001		n = 163,843			
KDIGO risk, [n (%)]							
N non missing		5797		18,544			
Low risk		2022	34.9	3726	20.1	0.32	↑ More low-risk in FINEARTS-HF
Moderate risk		1688	29.1	4458	24	0.1	↔ Identical
High/very high risk		2087	36	10,360	55.9	0.43	↓ fewer high/very-high in FINEARTS-HF
Concomitant/background therapy [n (%)]							
Beta-blocker [†]	Yes	5096	84.9	107,133	65.4	0.46	↑ Higher in FINEARTS-HF
ACEi or ARB	Yes	4767	79.4	73,692	45	0.76	↑ Higher in FINEARTS-HF
ARNI		509	8.5	13,848	8.5	0	↔ Identical
SGLT-2is		816	13.6	23,312	14.2	0.02	↔ Identical
MRA [†]		313	5.22	29,268	17.9	0.4	↓ Lower in FINEARTS-HF
Loop diuretic [†]		5240	87.3	96,720	59	0.67	↑ Higher in FINEARTS-HF
Frailty [†]							
Cardio-renal metabolic condition, n [†]	Yes						
NYHA functional class at baseline [†] (II, III/IV)							

[†]Identified in the clinical SLR.

ACEi: Angiotensin-converting enzyme inhibitor; AF: Atrial fibrillation; AMI: Acute myocardial infarction; ARB: Angiotensin II receptor blocker; ARNI: Angiotensin receptor–neprilysin inhibitor; BMI: Body mass index; COPD: Chronic obstructive pulmonary disease; eGFR: Estimated glomerular filtration rate; HF: Heart failure; ICD: Internation classification of diseases; IQR: Interquartile range; KDIGO: Kidney disease: Improving global outcomes; LVEF: Left ventricular ejection fraction; MRA: Mineralocorticoid receptor antagonist; NT-proBNP: N-terminal pro-B-type natriuretic peptide; NYHA: New York Heart Association (functional class); SD: Standard deviation; SGLT-2is: Sodium-glucose co-transporter-2 inhibitor; UACR: Urinary albumin-to-creatinine ratio.

indicating that even large increases in the number of SGLT-2is users in FINEARTS-HF would not have impacted comparative effects. SGLT-2is use therefore did not pose a risk to transportability even in countries where current SGLT-2is use may be higher than in the trial. To reinforce this conclusion, further research could explore the impact of SGLT-2is use during follow-up and post-2023, when prescribing patterns may have evolved. Although baseline SGLT-2is use did not appear to modify the comparative effect of finerenone in FINEARTS-HF, HF treatment guidelines are evolving rapidly. Future transportability assessments incorporating updated guidelines may therefore be warranted.

Based on available data, there was evidence of substantial homogeneity in comparative effects for all endpoints for most potential effect modifiers. A subset of effect modifiers was identified, but these were few in numbers and differed by endpoint. Such homogeneity indicated that the transportability of FINEARTS-HF results was unlikely to be substantially compromised. This was supported by evidence of direct proxy tests of transportability, which identified the absence of statistically significant differences of treatment effects on any outcome across US and non-US regions. Differences in some effect modifiers between FINEARTS-HF and the US target sample were found, although they were marginal in some cases. In addition, differences in effect modifiers between the samples suggested that, for the majority of outcomes studied, treatment effects observed in FINEARTS-HF likely represented underestimates of the comparative effects of finerenone that would be observed if FINEARTS-HF had been conducted solely in the US target population.

Comparative effects from the global FINEARTS-HF trial are expected to inform reimbursement decisions either directly or indirectly as inputs into cost–effectiveness models. Findings from this study indicate that the external validity of FINEARTS-HF results was likely to be high, reducing concerns about their applicability to the US market. Our findings provide important insights for key stakeholders, including clinicians, payers and health policy makers. By assessing effect modifiers and comparing the trial population to a target sample, this analysis demonstrates how findings from the FINEARTS-HF study can be extrapolated to real-world populations. For payers in particular, such transportability assessments help clarify whether trial efficacy estimates are likely to be confirmed in practice and inform reimbursement decisions and reduce uncertainty around coverage.

A direct comparison between the US population of FINEARTS-HF sample and a US target sample (TULIP-US) has not been performed. Nonetheless, given that no significant differences in treatment effect were identified

Table 2. Impact of effect modifiers within FINEARTS-HF.

Outcome	Effect modifiers identified for the outcome	Heterogeneity in treatment effect by the effect modifier	Final conclusion
Primary composite	eGFR at baseline	Lower eGFR is associated with less favorable comparative effects of finerenone	Patients with less favorable effects are over-represented in FINEARTS-HF. Treatment effect is likely underestimated in the trial vs in the US target population.
Total HF events	eGFR at baseline	Lower eGFR was associated with less favorable comparative effects of finerenone	Patients with less favorable effects are over-represented in FINEARTS-HF. Treatment effect is likely underestimated in the trial vs in the US target population.
	Age	Lower age was associated with more favorable comparative effects on finerenone	
All-cause mortality	Baseline serum potassium	Lower serum potassium was associated with more favorable comparative effects of finerenone	Higher use of beta-blocker suggests a less favorable effect in FINEARTS-HF. Serum potassium level was marginally higher in the trial implying a less favorable effect in FINEARTS-HF. Treatment effect is likely underestimated in the trial vs in the US target population. But as the frequency of patients per number of CRM conditions was unknown in TULIP-US, the impact is uncertain.
	Number of CRM condition	Having 2 or 3 CRM conditions was associated with less favorable comparative effects of finerenone	
	Beta-blocker use at baseline	Use of beta-blocker was associated with less favorable comparative effects of finerenone	
Composite renal	LVEF	LVEF < 60%: Strong favorable comparative effects of finerenone	As patients with LVEF $\geq 60\%$ and with eGFR <60% are over-represented in the trial, treatment effect might be underestimated in the trial. The level of UACR suggests the contrary, but there are too many missing values to rely on. The impact is uncertain.
		LVEF $\geq 60\%$: Strong deleterious comparative effects of finerenone	
	eGFR at baseline	Shorter time to a relevant renal event for finerenone vs placebo overall was driven by patients with lower eGFR	
	Baseline UACR	Shorter time to a relevant renal event for finerenone vs placebo overall was driven by patients with higher UACR	
KCCQ-TSS (Δ from baseline)	Baseline NT-proBNP	Higher NT-proBNP was associated with statistically more favorable comparative effects of finerenone	Higher use of beta-blocker suggests a less favorable effect in trial. Higher value of NT-proBNP suggests the opposite. Due to the importance of NT-proBNP missing values, the impact is uncertain.
	Beta-blocker use at baseline	Use of beta-blocker was associated with less favorable comparative effects of finerenone	
P(NYHA improvement)	ACEi or ARB use at baseline	Use of RASi was associated with less favorable comparative effects of finerenone	Higher use of ACEi or ARB suggests a less favorable effect in trial. Higher value of NT-proBNP suggests the opposite. Due to the importance of NT-proBNP missing values, the impact is uncertain.
	Baseline NT-proBNP	Lower NT-proBNP was associated with less favorable comparative effects of finerenone	

ACEi: Angiotensin-converting enzyme inhibitor; ARB: Angiotensin II receptor blocker; CRM: Cardio-renal-metabolic conditions; eGFR: Estimated glomerular filtration rate; HF: Heart failure; KCCQ-TSS: Kansas City Cardiomyopathy Questionnaire–Total Symptom Score; LVEF: Left ventricular ejection fraction; NT-proBNP: N-terminal pro-B-type natriuretic peptide; NYHA: New York Heart Association (functional class); P(NYHA improvement): Probability of New York Heart Association improvement; RASi: Renin–angiotensin system inhibitor (includes ACEi, ARB, ARNI); UACR: Urinary albumin-to-creatinine ratio.

Table 3. Direct proxy tests results from FINEARTS-HF.

	Full trial	US assessment		
		Non-US	US	Interaction test (p-value)
Primary composite (RR [95% CI])	0.84 [0.74; 0.95]	0.82 [0.72; 0.94]	1.01 [0.66; 1.56]	0.3788
Total HF events (RR [95% CI])	0.82 [0.71; 0.94]	0.80 [0.68; 0.93]	1.03 [0.66; 1.62]	0.2951
Time to all-cause mortality (HR [95% CI])	0.93 [0.83; 1.06]	0.94 [0.83; 1.07]	0.84 [0.53; 1.34]	0.6462
Time to composite renal endpoint (HR [95% CI])	1.33 [0.94; 1.89]	1.40 [0.98; 2.02]	0.59 [0.14; 2.49]	0.267
Change from baseline in KCCQ-TSS (MD [95% CI])	1.56 [0.79; 2.34]	1.62 [0.82; 2.41]	0.86 [-2.75; 4.47]	0.5967
Probability of NYHA improvement (OR [95% CI])	1.01 [0.88; 1.15]	1.01 [0.88; 1.15]	1.02 [0.57; 1.82]	0.9776

CI: Confidence interval; HF: Heart failure; HR: Hazard ratio; KCCQ-TSS: Kansas City Cardiomyopathy Questionnaire–Total Symptom Score; NYHA: New York Heart Association (functional class); OR: Odds ratio; RR: Relative risk.

between the US subgroup and the overall trial population, using the full trial sample as the basis for comparison does not appear to introduce additional bias.

The main strength of the study is that it relies on a robust approach to assessing transportability through identification of effect modifiers identification based on an SLR of existing studies combined with interaction tests using data from FINEARTS-HF. The US target sample is plausibly representative of current practice in US target population and provides robust data on the majority of effect modifiers. However, as Optum does not randomly

sample patients from the US population, its representativeness cannot be guaranteed. But, Optum comprises longitudinal, patient-level EHRs for approximately 100 million individuals receiving care at over 700 hospitals and clinics across the US.

One limitation is that the trial was not designed to detect effect modification for some endpoints and some potential effect modifiers. To partially address this a threshold of 0.1 (rather than 0.05) was used to determine statistical significance. However, this was counterbalanced by the increased risk of false positives arising from multiple hypothesis testing.

Another limitation is that transportability can only be assessed based on data on observed effect modifiers. As with any study of this type, there is a potential for unobserved effect modification that may hamper transportability. However, data on a large number of effect modifiers were examined, and data from only two possible effect modifiers were not available in the target sample to assess the representativeness of the FINEARTS-HF trial. Frailty and number of CRM conditions could not be derived in Optum EHR and therefore could not be compared between FINEARTS-HF and TULIP-US. However, these unavailable variables are unlikely to materially change the primary conclusions for three reasons: the overall number of effect modifiers was small, the primary endpoint showed limited evidence of effect modification, and direct proxy tests found no statistically significant heterogeneity between US and non-US trial participants. Together, these findings reduce the likelihood that imbalance in a small number of unobserved characteristics would substantially alter the overall transportability conclusion.

Moreover, although the definitions of effect modifiers in FINEARTS-HF could be replicated in TULIP-US, in some cases, translation of commonly applied trial definitions was not possible, given a lack of relevant variables in the Optum EHR setting (e.g., NYHA class). In addition, where definitions could be replicated, differences in reporting between trial and real-world settings could have caused differences in the accuracy of measurement across FINEARTS-HF and TULIP-US. Measurement error or misclassification resulting from such measurement differences could have resulted in observed homogeneity or heterogeneity across FINEARTS-HF and target samples identified partial reflecting measurement process rather than true clinical homogeneity/heterogeneity. Results of the assessment of the representativeness of FINEARTS-HF should therefore be interpreted in this context.

Effects of baseline eGFR should be interpreted with caution, as they may be confounded by the titration scheme (only patients with eGFR > 60 ml/min/1.73 m² group could receive the 40 mg dose).

Future work could involve expanding the transportability assessment to other countries of interest by performing de-novo real world studies (to generate representative target samples) and subsequently to verify representativeness of FINEARTS-HF in other jurisdictions.

Conclusion

The transportability assessment concluded that there is no evidence that transportability of FINEARTS-HF is impacted by prevalence of SGLT-2is use or other potential effect modifiers. The risk of substantial bias from absence of transportability is likely to be low. Considering the US, estimates of finerenone in FINEARTS-HF likely represent conservative estimates of effects in the US target population for the majority of outcomes.

Summary points

- Transportability assessments are critical to determine whether treatment effects observed in randomized clinical trials (RCTs) can be generalized to real-world populations relevant for health technology assessment (HTA).
- A structured three-phase framework was applied to assess the transportability of results from the FINEARTS-HF trial to a real-world population in the US using Optum® electronic health records.
- The framework included identification of effect modifiers, assessment of the representativeness of the trial with respect to effect modifiers and direct proxy tests assessing cross-country heterogeneity in treatment effects in the trial.
- Only a limited number of variables were identified as effect modifiers across endpoints, suggesting relative homogeneity of treatment effects.
- Differences in effect modifiers between the trial and US populations were generally modest. However, patient groups with less favorable treatment effects were over-represented in the trial, suggesting treatment effects in the trial could underestimate the benefit of finerenone expected in the US population.
- For some endpoints, uncertainty remained due to missing data on key effect modifiers in the target population.
- No statistically significant heterogeneity in treatment effects was observed between US and non-US participants in the trial, supporting transportability.
- Overall, the findings support the transportability of FINEARTS-HF results to the US population.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at <https://becarispublishing.com/doi/epdf/10.57264/cer-2026-0003>

Author contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by AJ Turner and C Leboucher. TULIP-US analyses have been performed by YM Fung. The first draft of the manuscript was written by C Leboucher and all authors commented on all versions of the manuscript. All authors read and approved the final manuscript.

Financial disclosure

This study was funded by Bayer AG, Germany.

Competing interests disclosure

A Turner is a member of the University of Manchester. C Leboucher and C Remuzat are employees of Putnam. YM Fung and K Folkerts are employees of Bayer AG. E Pessina was an employee of Bayer AG at the time the study was conducted. The authors have no other competing interests or relevant affiliations with any organization or entity with the subject matter or materials discussed in the manuscript apart from those disclosed.

Writing disclosure

No funded writing assistance was utilized in the production of this manuscript.

Data sharing statement

The authors certify that this manuscript reports the secondary analysis of clinical trial data that have been shared with them, and that the use of this shared data is in accordance with the terms (if any) agreed upon their receipt. The source of this data is: FINEARTS-HF (ClinicalTrials.gov: NCT04435626).

Open access

This work is licensed under the Attribution-NonCommercial-NoDerivatives 4.0 Unported License. To view a copy of this license, visit <https://creativecommons.org/licenses/by-nc-nd/4.0/>

Reference

Papers of special note have been highlighted as: ● of interest; ●● of considerable interest

1. National Institute for Health and Care Excellence. NICE real-world evidence framework. (2022). <https://www.nice.org.uk/corporate/ecd9>
- **Provides guidance on the use of real-world evidence in health technology assessment (HTA) decision-making.**
2. Institute for Quality and Efficiency in Health Care. General methods. (2020). <https://www.iqwig.de/methoden/allgemeine-methoden-v8-0.pdf>
3. Makady A, Ham RT, de Boer A *et al.* Policies for use of real-world data in health technology assessment (HTA): a comparative study of six HTA agencies. *Value Health* 20(4), 520–532 (2017).
- **Describes how health technology assessment (HTA) agencies incorporate real-world data into decision-making frameworks.**
4. Westreich D, Edwards JK, Lesko CR *et al.* Transportability of trial results using inverse odds of sampling weights. *Am. J. Epidemiol.* 186(8), 1010–1014 (2017).
5. Rothwell PM. Factors that can affect the external validity of randomised controlled trials. *PLOS Clin. Trials* 1(1), e9 (2006).
6. Turner AJ, Sammon C, Latimer N *et al.* Transporting comparative effectiveness evidence between countries: considerations for health technology assessments. *Pharmacoeconomics* 42(2), 165–176 (2024).
- **Highlights practical challenges of transportability in HTA across jurisdictions.**
7. Dahabreh IJ, Robertson SE, Steingrimsson JA *et al.* Extending inferences from a randomized trial to a new target population. *Stat. Med.* 39(14), 1999–2014 (2020).
- **Provides a foundational statistical framework for extending trial results to target populations.**
8. Degtiar I, Rose S. A review of generalizability and transportability. *Annu. Rev. Stat. Appl.* 10(1), 501–524 (2023).
- **Comprehensive and up-to-date review of transportability and generalizability methods.**
9. Drummond M, Barbieri M, Cook J *et al.* Transferability of economic evaluations across jurisdictions: ISPOR good research practices task force report. *Value Health* 12(4), 409–418 (2009).

10. Goeree R, He J, O'Reilly D. Transferability of health technology assessments and economic evaluations: a systematic review of approaches for assessment and application. *Clin. Outcomes Res.* 3, 89–104 (2011).
11. Ling AY, Montez-Rath ME, Carita P. An overview of current methods for real-world applications to generalize or transport clinical trial findings to target populations of interest. *Epidemiology* 34(5), 627–636 (2023).
12. Pearl J, Bareinboim E. External validity: from do-calculus to transportability across populations. *Stat. Sci.* 29(4), 579–595 (2014).
- **Establishes the causal inference foundations of transportability theory.**
13. Stuart EA, Bradshaw CP, Leaf PJ. Assessing the generalizability of randomized trial results to target populations. *Prev. Sci.* 16(3), 475–485 (2015).
14. Jaksa A, Arena PJ, Chan KKW *et al.* Transferability of real-world data across borders for regulatory and health technology assessment decision-making. *Front. Med. (Lausanne)* 9, 1073678 (2022).
15. Levy NS, Arena PJ, Jemielita T *et al.* Use of transportability methods for real-world evidence generation: a review of current applications. *J. Comp. Eff. Res.* 13(11), e240064 (2024).
16. Vuong Q, Metcalfe RK, Ling A *et al.* Systematic review of applied transportability and generalizability analyses: a landscape analysis. *Ann. Epidemiol.* 104, 61–70 (2025).
17. Solomon SD, McMurray JJV, Vaduganathan M *et al.* Finerenone in heart failure with mildly reduced or preserved ejection fraction. *N. Engl. J. Med.* 391(16), 1475–1485 (2024).
- **Pivotal randomized trial forming the basis of the present transportability assessment.**
18. Lam CSP, Fonarow GS, Fung YM *et al.* Guideline-directed medical therapy treatment patterns in patients with newly diagnosed heart failure stratified by left ventricular ejection fraction in the United States. Presented at: *Heart Failure Society of America*. (26–29 September 2025).
19. Lam CSP, Fonarow GS, Fung YM *et al.* Clinical outcomes after onset of heart failure with left ventricular ejection fraction in the US. Presented at: *Heart Failure Society of America*. (26–29 September 2025).
20. Kittleson MM, Panjra GS, Amancherla K *et al.* 2023 ACC expert consensus decision pathway on management of heart failure with preserved ejection fraction. *J. Am. Coll. Cardiol.* 81(18), 1835–1878 (2023).
21. Cochrane. Cochrane Handbook for Systematic Reviews of Interventions. (2024). <https://www.cochrane.org/authors/handbooks-and-manuals/handbook>