Journal of Clinical Epidemiology 162 (2023) 160−168

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Transportability of two heart failure trials to a disease registry using individual patient data

Lili Wei[a,*], David M. Phillippo[b], Anoop Shah[c], John G.F. Cleland[d], Jim Lewsey[a], David A. McAllister[a]

[a]School of Health and Wellbeing, University of Glasgow, Glasgow, UK
[b]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK
[c]Department of Noncommunicable Disease, London School of Hygiene & Tropical Medicine, London, UK
[d]School of Cardiovascular & Metabolic Health, University of Glasgow, Glasgow, UK

Accepted 27 August 2023; Published online 1 September 2023

Abstract

Objectives: Randomized controlled trials are the gold-standard for determining therapeutic efficacy, but are often unrepresentative of real-world settings. Statistical transportation methods (hereafter transportation) can partially account for these differences, improving trial applicability without breaking randomization. We transported treatment effects from two heart failure (HF) trials to a HF registry.

Study Design and Setting: Individual-patient-level data from two trials (Carvedilol or Metoprolol European Trial (COMET), comparing carvedilol and metoprolol, and digitalis investigation group trial (DIG), comparing digoxin and placebo) and a Scottish HF registry were obtained. The primary end point for both trials was all-cause mortality; composite outcomes were all-cause mortality or hospitalization for COMET and HF-related death or hospitalization for DIG. We performed transportation using regression-based and inverse odds of sampling weights (IOSW) approaches.

Results: Registry patients were older, had poorer renal function and received higher-doses of loop-diuretics than trial participants. For each trial, point estimates were similar for the original and IOSW (e.g., DIG composite outcome: OR 0.75 (0.69, 0.82) vs. 0.73 (0.64, 0.83)). Treatment effect estimates were also similar when examining high-risk (0.64 (0.46, 0.89)) and low-risk registry patients (0.73 (0.61, 0.86)). Similar results were obtained using regression-based transportation.

Conclusion: Regression-based or IOSW approaches can be used to transport trial effect estimates to patients administrative/registry data, with only moderate reductions in precision.

Keywords: Heart failure; Randomized controlled trial; Transportability; Individual patient data; Inverse odds of sampling weights; Calibration

1. Introduction

Randomized controlled trials (RCTs) are the gold-standard for determining the efficacy and safety of treatments [1,2]. However, participants in heart failure (HF) RCTs are generally younger, more likely to be men, and have fewer comorbidities such as chronic respiratory or kidney disease than those encountered in clinical practice [3,4]. If the patient characteristics that are underrepresented are also associated with differences in treatment efficacy (e.g., if efficacy is lower in older people), the applicability of trial findings to clinical practice is attenuated. Partly for this reason, RCTs report baseline characteristics (such as age, sex, and disease severity) as well as treatment effects stratified by subgroups. However, individual patients may have many coexisting characteristics (for instance anemia and renal dysfunction) which are not represented in trial analysis with one-variable-at-a-time subgroup reporting [5].

Statistical trial transportation, also called calibration or population adjustment in other contexts [6−8], addresses these difficulties by weighting trial results to reflect the characteristics of target populations more closely and, importantly, without *breaking* randomization. Briefly, transportation apportions greater weight to randomized

Data availability: The data that has been used is confidential.

Funding: This work was supported by a grant from the Wellcome Trust (201492/Z/16/Z).

* Corresponding author. School of Health and Wellbeing, University of Glasgow, Clarice Pears Building, 90 Byres Road, G12 8TB Glasgow, United Kingdom. Tel.: +44 (0) 0141-330-5009; fax: +44 (0) 141-330-3299.

E-mail address: Lili.Wei@glasgow.ac.uk (L. Wei).

https://doi.org/10.1016/j.jclinepi.2023.08.019

What is new?

Key findings

- Regression-based and inverse odds of sampling weights approach can be performed in heart failure trials to improve trial applicability without breaking randomization, albert with a moderate loss of precision.

What this adds to what was known?

- While trials may have limited applicability when applied to the real-world population, our statistical transportation methods have demonstrated they can be enhanced.

What is the implication and what should change now?

- Accessing individual participant data (IPD) for trials and registries is challenging. Changes in trial reporting (providing variance-covariance matrix for a treatment effect model including treatment-covariate interactions) could improve the situation.
- Similarly, if summary data on the joint distribution of patient characteristics was provided from the registry, trialists with IPD could calibrate trials accordingly.

participants that were underrepresented in the trial compared to the target population and less weight to participants who were overrepresented in the trial, compared to the target population [6,9]. Calibration has been used in other conditions such as HIV [10] and lung cancer [11] and in dual antiplatelet therapy (DAPT) study [12] but we are not aware of any previous attempt to transport HF trials to a clinical practice registry for patients with HF to estimate effects in clinical practice. RCTs require considerable resources, in terms of research staff, finances and patient commitment [13]; it is important to maximize their utility for clinical practice.

Accordingly, we examined the effect of transporting two landmark HF trials to patients from a Scottish clinical practice HF registry using two different methods with differing assumptions—inverse odds of sampling weights (IOSW) and regression modeling.

2. Methods

2.1. Data sources

The rationale, design, methods, and principal results of COMET, digitalis investigation group trial (DIG) have been described by the original investigators [14−17], but each are briefly described below.

2.2. Carvedilol or Metoprolol European Trial (COMET)

COMET was a multicenter, randomized, double-blind, parallel-group comparison of carvedilol, and metoprolol in participants with a left ventricular ejection fraction (LVEF) of 35% or less. Conducted in 15 European countries, 1,511 participants were randomly assigned to carvedilol and 1,518 to metoprolol tartrate. The mean trial duration was 58 months. The primary end points were all-cause mortality, and a composite of all-cause mortality or all-cause hospitalization [14,15].

2.3. The Digitalis Investigation Group Trial

DIG was a randomized, double-blind trial of the effect of digoxin compared to placebo among people with chronic HF. Studied in 302 centers, 7,788 participants were involved. The main trial was conducted for 6,800 participants with an LVEF of 45% or less. The average follow-up time was 37 months. The primary outcome was all-cause mortality. Worsening HF culminating in death or hospitalization was a combined secondary outcome [16,17].

2.4. Heart failure registry

A clinical practice HF registry of 8,012 individuals predominantly with reduced ejection fraction (HFrEF) was obtained from the largest regional health authority in Scotland (National Health Service Greater Glasgow & Clyde, NHSGGC), which covers 1.3 million people (almost a quarter of the Scottish population) [18]. People with HF in the region who were assessed by community HF nurses or HF clinics were included in the registry. Each patient's clinical features (diabetes, ischemic heart disease, etc.); therapy; vital signs including heart rate, systolic blood pressure (SBP), etc; results of blood tests (serum sodium, potassium, creatinine, etc.) were routinely recorded in an electronic health record to support clinical care. Missing values in the HF registry were imputed by a predictive mean matching algorithm with one imputed dataset being generated for simplicity [19]. The imputation includes diabetes (12%), SBP (14%), heart rate (12%), serum sodium (24%), estimated glomerular filtration rate (eGFR, 24%) and loop diuretics (17%) with dose expressed in furosemide equivalents (e.g., 1 mg of bumetanide = 40 mg of furosemide).

2.5. Statistical methods

Each trial was analyzed separately. For the primary end point (all-cause mortality in both trials) and the composite end point (all-cause mortality or all-cause hospitalization in COMET and worsening HF culminating in death or hospitalization in DIG), each trial was calibrated to the 8,012 patients in the HF registry, first using a regression-based

method (Supplementary Figure S2) and then using IOSW (Figure S3). Each method is described below, with detailed steps provided in the Supplementary.

2.6. Regression-based transportation

A model based on the trial data was first constructed. Variables were included as model covariates based on their availability in both the trial and registry and their potential to modify treatment effectiveness (as Table 1).

Detailed description and implementation of the regression-based transportation are described in the Supplementary Appendix. Briefly, 2 parametric survival models were built separately for the trial data and the registry with the same distribution of the best model fit and with model diagnostics. The model conducted for the trial included main effects of all covariates (age, SBP, etc.) and 2-way interactions with the treatment variable (treatment by age interaction, etc.). The natural history model for registry did not contain treatment main effects and interactions compared with the trial model as the treatment effects are estimated solely using the trial data. Coefficients for treatment main effect and interactions from the trial model and coefficients from the registry model of the main effects of other variables (age, SBP, etc.) were extracted to form an integrated set of coefficients. Assuming that patients in the registry receive the same intervention and comparator as in the trial and then applying the integrated set of coefficients to the registry, enables the risk of each outcome and the effect of intervention to be estimated. Uncertainty in the coefficients was propagated to the final model via simulation, with 100,000 samples being generated. The outcome predictions and treatment effects were calculated for each sample and summarized via the mean (geometric mean for relative measures) with the uncertainty expressed via the 2.5th and 97.5th percentiles (95% confidence interval (CI)).

In additional analyses, we used the natural history model fitted to the registry data to estimate the risk of the covariates (e.g., age, male, SBP) and outcomes (all-cause death et al. corresponding with trial outcomes) to estimate the predicted risk for each individual in the register, ranked these, then selected the top 10 percentile highest and top 10 percentile lowest into the highest and lowest risk subgroups respectively. We then repeated the final step of the above analysis (applying the regression coefficients to individuals with this set of covariates) for these high and low risk subgroups.

2.7. Inverse odds of sampling weights (IOSW) transportation

Briefly, using the same covariates as regression-based method, the trial and registry datasets were aggregated to obtain counts of individuals with each combination of characteristics (Supplemental Table S5). The probability that an individual from the HF registry is included in the trial sample, conditional on covariates, divided by the probability of not being in the trial sample—the inclusion odds—was then estimated by comparing these counts.

We then estimated the treatment effects as standard by comparing the odds of the outcome in each treatment arm; except that instead of all participants having the same weight in the analysis, different participants were weighted differently according to their inclusion odds. As an example, if there were 500 individuals with a given set of characteristics in the registry, and 5 participants with that set of characteristics in the trial, the odds of being sampled would be 1% (5/500). This would translate to a raw weighting of 100 (1/odds) for those 5 participants. Final weights for all participants would then be calculated by dividing each participant's weight by the sum of weights for all participants. See Supplementary Appendix for details on the methods used to calculate the inclusion odds and weightings, and to estimate the treatment effects using the weightings.

In additional analyses we also refitted the IOSW models based on the same high and low risk subgroups from the HF registry identified using the regression-based approach.

All analyses were conducted in R (R 3.4.0 for trial and R 3.5 for the registry). The parametric survival models were fitted using the ''flexsurv'' package [20] and the weighted logistic regression models were fitted using the ''survey'' package [21].

3. Results

3.1. Baseline characteristics

The clinical characteristics of patients in the three datasets are shown in Table 1. Patients in the registry were older than trial participants and had lower eGFR and higher doses of loop diuretics. The proportion of patients who were men was higher in both trials compared to the registry.

3.2. Effect of baseline characteristics on outcomes

The parametric survival model with a generalized gamma distribution had the best fit and was used for the HF registry and each trial. The coefficient for the covariates for the HF registry and trials are shown in Fig. 1 and Supplementary Table S6. These coefficients are mutually adjusted. In both trials, male sex, older age, and history of diabetes predicted a worse prognosis and higher eGFR predicted longer survival. In COMET, use of higher dose loop diuretics also predicted a worse outcome, higher serum sodium concentration, and higher SBP predicted better prognosis.

3.3. Effect of baseline characteristics on treatment efficacy

The estimates for the treatment effects (at the mean of all the covariate levels) and the treatment-covariate

Table 1. Baseline characteristics in each dataset included in calibrations

Variables	HF registry (N = 8,012)	COMET (N = 3,029)	DIG (N = 6,800)
Age (yr)	73 (12)	62 (11)	64 (11)
Men, *n* (%)	4,906 (61%)	2,412 (80%)	5,281 (78%)
History of diabetes, *n* (%)	1,863 (23%)	728 (24%)	1,933 (28%)
Heart rate (beats per minute)	73 (13)	81 (13)	79 (13)
Systolic blood pressure (mm Hg)	120 (21)	126 (19)	126 (20)
Serum sodium (mmol/l)	138 (4)	140 (3)	–
eGFR (mL/min/1.73 m^2)	59 (23)	67 (21)	62 (21)
Loop diuretics (mg/day)	62 (31)	20 (46)	[a]

Categorical variables are shown as counts (%s) and continuous variables as means (standard deviations).
– Not available.
Abbreviations: COMET, Carvedilol or Metoprolol European Trial; DIG, Digitalis Investigation Group Trial; HF, heart failure; eGFR, estimated glomerular filtration rate.
[a] In DIG loop diuretics was recorded as a categorical variable (whether participants had taken it or not or unknown) and the dosage information was not available.

interactions are shown in Fig. 2 and Supplementary Table S7. The treatment-covariate interaction estimates were wide, and for some variables the magnitude and direction of the point estimates varied between trials. For both COMET and DIG, treatment efficacy appeared to be lower for patients with diabetes (accelerated failure time (AFT) ratio: 0.95 (0.66, 1.37) and 0.93 (0.67, 1.29) for COMET all-cause death and composite outcome; 0.90 (0.71, 1.15) and 0.81 (0.59, 1.10) for DIG all-cause death and composite outcome) and greater for heartrate (AFT ratio: 1.06 (0.90, 1.25) and 1.13 (0.98, 1.31) for COMET and 1.07 (0.96, 1.20) and 1.12 (0.97, 1.30) for DIG), but the CIs almost all included the null.

3.4. Effect of transportation on treatment effects

Fig. 3 and Supplementary Table S10 show the calibrated treatment effects. For either primary or composite outcome in DIG over a period of 3 years, the uncalibrated and calibrated effect estimates (odds ratios, ORs) were similar (OR: 0.99 (0.91, 1.07) vs. 1.06 (0.92, 1.21) vs. 1.05 (0.86, 1.28) for uncalibrated analysis, IOSW and regression-based

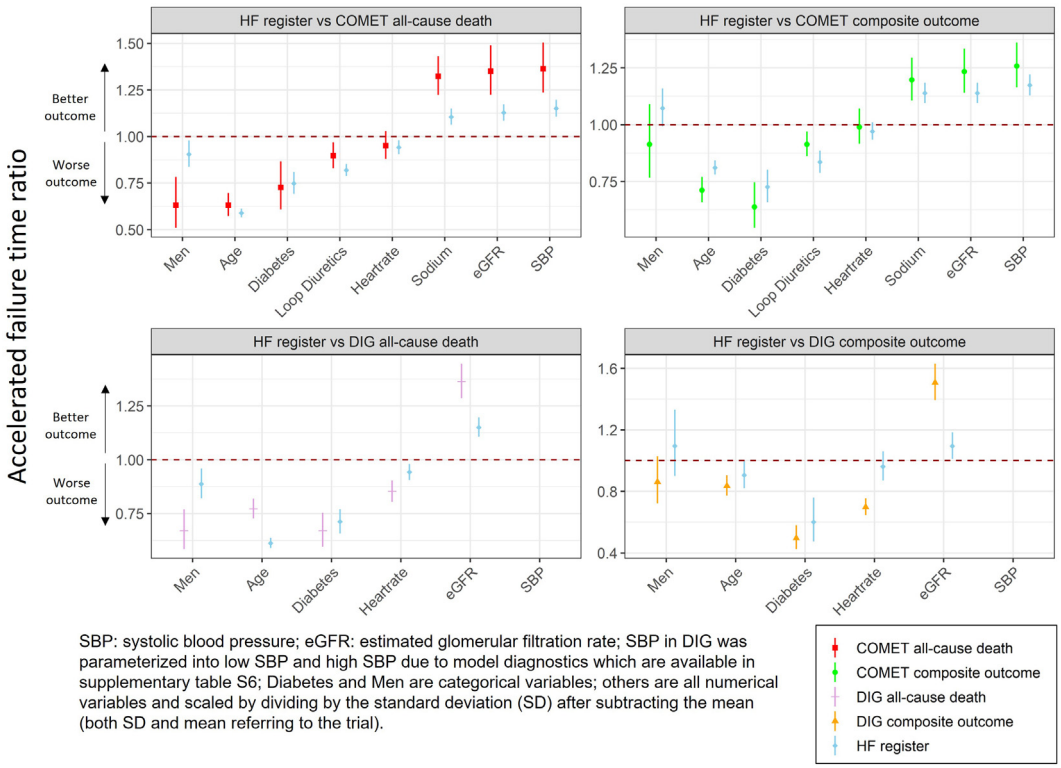

Fig. 1. Main effects in HF registry and two trials. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

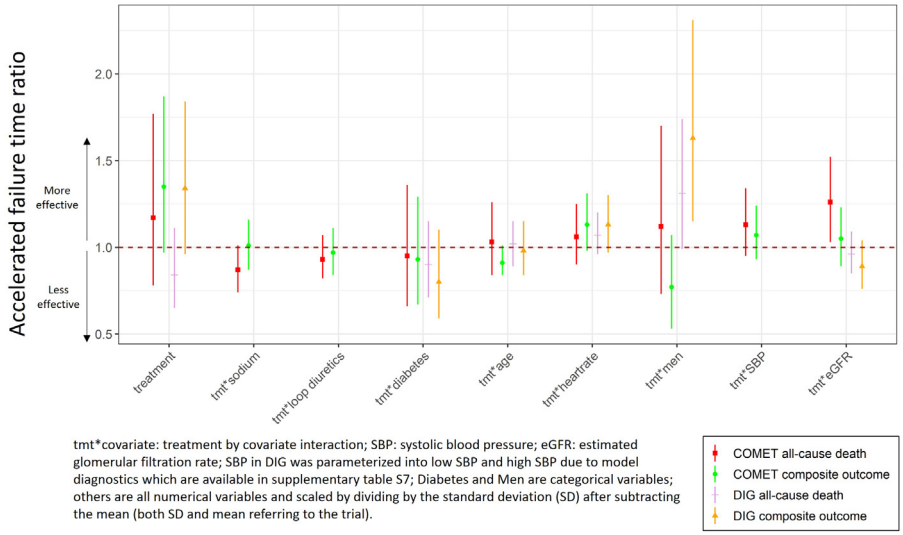

Fig. 2. Treatment and treatment-covariate interactions in two trials. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

transportation for all-cause death and 0.75 (0.69, 0.82) vs. 0.73 (0.64, 0.83) vs. 0.84 (0.78, 0.91) for the composite outcome), indicating similar efficacy in the trial and HF registry. For COMET the efficacy was higher for IOSW (OR: 0.62 (0.39, 0.99) and 0.87 (0.59, 1.30) for all cause death and composite outcome over a period of 4 years) but lower for the regression-based transportation (0.97 (0.72, 1.27) and 1.08 (0.81, 1.39)) although the 95% CIs

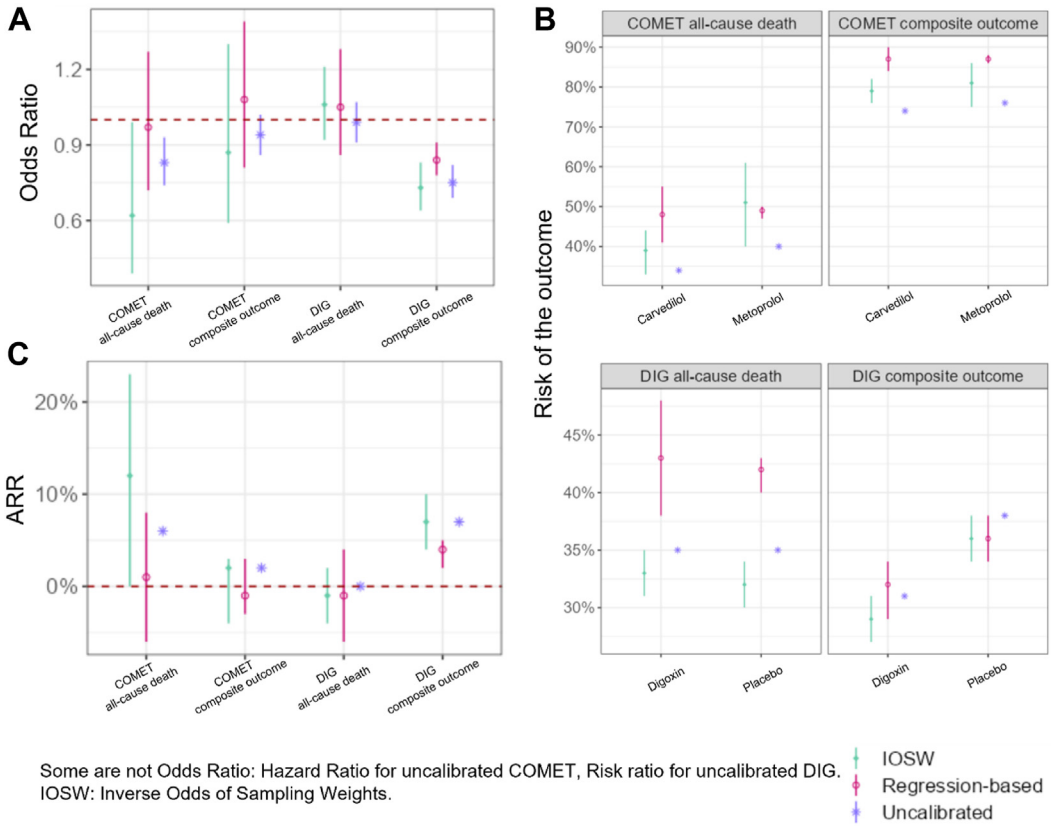

Fig. 3. Measure of effects in uncalibrated and calibrated analyses in two trials. (A) Odds ratio; (B) Risk of the outcome; (C) Absolute risk reduction (ARR). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

overlapped those of the uncalibrated estimates (0.83 (0.74, 0.93) and 0.94 (0.86, 1.02)). The impact of transportation was similar where the trials were calibrated to the high-risk and low-risk subgroups (Supplementary table S12 and Figure S6).

Where the calibrated and uncalibrated results differ, the contribution of each covariate to this difference can be estimated as the covariate-treatment interaction multiplied by the mean difference in the covariate between the registry and trial population. For COMET (death), eGFR and loop diuretics dose were the main drivers, for COMET (composite) age and heartrate were the main drivers and for DIG, male sex and heart rate were the main drivers (Supplementary Table S8).

Compared to the uncalibrated and IOSW models, the estimated risk of the outcome within each treatment arm (except for metoprolol arm in COMET all-cause death, placebo arm in DIG composite) was larger for the regression-based model (Fig. 3B), for example, the estimated mortality in the digoxin arm of DIG was 35%, 33% and 43% in the uncalibrated, IOSW and regression-based analysis respectively. However, these differences of risk in each treatment arm did not translate to large differences in the absolute risk reductions (see Fig. 3C).

3.5. Effect of transportation on precision of treatment efficacy

As expected, compared to the standard analysis, the standard errors (SEs) were generally larger for the calibrated effect estimates (Supplementary Table S13). For transportation to the overall target population, this ranged from no increase to 4.6-fold wider SEs (e.g., SEs are 0.06, 0.15, and 0.24 for uncalibrated analysis, regression-based and IOSW transportation for COMET all-cause mortality). Where the results were calibrated to the highest and lowest risk subgroups of the registry, which by design were more different from the trial populations based on baseline characteristics than was the overall population, the SEs ranged from 1.2-fold to 12.7-fold wider. After excluding the 1% patients with the lowest odds (highest weights) of trial inclusion (Supplementary Tables S12 and S13), the SE ranged from no increase to 3.1-fold wider for overall target population, and it ranged from 1.2-fold to 7.8-fold wider for highest and lowest risk subgroups.

4. Discussion

We calibrated two landmark HF trials to a Scottish ‘‘real-world’’ population using two approaches, regression-based and IOSW. Both were straightforward to perform, with only moderate loss of precision manifested as larger SEs. This suggests that trials can be calibrated to registry data, maximizing representativeness and applicability while preserving the benefits of randomization.

Previous studies have used calibration using IOSW or generalization via inverse probability of sampling weights (IPSW) [10−12]. In DAPT study, using IOSW no longer showed a significant effect of prolonged DAPT on reducing stent thrombosis, major adverse cardiac and cerebrovascular events, but the increase in bleeding persisted [12]. Cole and Stewart used IPSW to calibrate a major HIV trial, using counts of people with HIV in the US stratified by age, sex and cluster of differentiation 4 count to define the target population [10]. The GetReal project calibrated a trial of chemotherapy for nonsmall cell lung cancer to a cohort study using IPSW with 15 baseline characteristics and IPSW showed a similar hazard ratio for pemetrexed compared with gemcitabine with greater uncertainty (a wider CI) [11]. We add to this literature by showing that HF trials can be calibrated to the more complex populations encountered in clinical practice with only moderate loss in precision, yielding similar results for both IOSW and a regression-based approach. Furthermore, HF trials can be calibrated to different risk subgroups based on multiple characteristics. Unlike conventional subgroup analyses this approach simultaneously accounts for the impact of all measured characteristics which differ between the trial and real-world settings.

In our analyses the calibration was performed to improve transportability rather than generalizability. When reweighting for generalizability, the technique is identical, except that the inverse of the probability of trial inclusion is used rather than the inverse odds.

Both IOSW and regression make assumptions (Supplementary Table S1). It is essential that the main effects and interactions are correctly modeled in the regression-based approach, and that all variables that predict both heterogeneity in participation and the outcome have been included in the trial inclusion odds model for the IOSW-approach. For both approaches, we also assume that there are no treatment-covariate interactions for unmeasured variables; although it is worth noting that the current standard approach of applying the relative treatment effect from trials to target populations makes the more extreme assumption that there are no covariate-treatment interactions of any kind (measured or unmeasured). Importantly, athough transportation helps address underrepresentation, caution is needed when extrapolating trial results to patients who could not have been included in the trial (thus violating the positivity assumption), in this case, for example, children or people living in Africa. From a purely technical point of view, there are differences between the different approaches when extrapolating the trial findings to patients with combinations of characteristics beyond the range of the trial data. Using the IOSW approach, it is technically impossible to reweight the estimates for levels of characteristics beyond the range of the trial data (e.g. if no trial participants were aged over 65 years one cannot estimate relative effects in a population over the age of 65). In contrast, in the regression-based approach,

so where covariates are modeled as continuous variables (e.g. linear terms, polynomials, etc.) it is technically straightforward to extrapolate beyond the data. Nonetheless, whether applying regression or IOSW it is important to consider whether the applicability of the predicted effect estimates, on the required scale, are genuinely transportable to the desired target population. In other words, whether the relevant assumptions are met. Furthermore, participant/patient characteristics are only one way in which the circumstances of the trial may differ from the target population, for example, there may be differences in clinical settings or time periods of enrollment. Differences in diagnosis, treatment delivery and monitoring may lead to differential efficacy (e.g. due to improved adherence, better tailoring of dosages, etc.) [22].These also need to be carefully considered when assessing the transportability of effect estimates, and are generally less amenable to the kind of adjustments described in our manuscript.

Differences in the assumptions of IOSW and regression approaches alone, provide justification for performing both. However, they also provide different information. For example, the IOSW approach involves calculating the trial inclusion odds, and this then provides an overall single summary measure for all trial participants and registry patients. This allows comparisons within and between these populations, to determine, for example, whether the trial and registry populations are sufficiently similar to undertake calibration. This is analogous to an advantage of propensity score weighting in pharmacoepidemiologic analyses (e.g., control for measured confounding, identify barriers for treatment such as age) [23]. In contrast, an advantage of the regression approach is that we can explore which differences between trial participants and registry patients are driving any observed discrepancies between calibrated and uncalibrated treatment effect estimates. This can be done by examining the magnitude of covariate-treatment interactions and comparing levels of these covariates between trial participants and registry patients.

The regression-based calibration approach builds on the standard evidence synthesis modeling process for producing absolute treatment effects in a target population, recommended in NICE technical support document 5 [24], wherein:- i) a standard care model for absolute outcomes is fitted to data representative of the target population, ii) a relative treatment effect model is fitted to trial data and iii) the two models are combined (usually using Monte Carlo methods or bootstrapping) to estimate absolute treatment effects. Our model differs in two ways. First, we do not assume homogeneity of relative treatment effects but allow these to differ according to individual participant characteristics. Secondly, rather than having a single estimate for the natural history model, or having two or more estimates stratified by some important characteristics (e.g. disease severity), we allow the rates to differ according to individual patient characteristics. Importantly, this approach works on the assumption that relative treatment effects are transportable between trial and target populations conditional on the covariates included in the relative effects model in the trial data, with the standard care model fitted in the target population providing the baseline absolute rates to which the transported relative effects are applied. This is in contrast to alternative standardization/g-computation approaches (e.g., as described by Dahabreh et al. [25]) which solely use the trial data-derived model to produce absolute predictions in the target population (i.e., the standard care model is estimated within the trial), and thus are based on the assumption that absolute effects are transportable between trial and target populations. This is a much more stringent assumption to meet because differences in all prognostic factors and effect modifiers between trial and target population must be accounted for instead of just the effect modifiers, and is generally considered far less plausible. When noncollapsible relative effects measures are used (e.g., odds ratios or hazard ratios), we must additionally take care to ensure that the parameters from the standard care model are compatible with the parameters from the relative treatment effects model; that is, that they are conditioned in the same manner. This is not necessarily true in our analysis as some of the individuals in the register were taking digoxin and/or carvedilol. However, in many applications where a standard care population can be readily defined (e.g., because a new treatment is being considered), this condition is likely to be true; because the standard care population is restricted to (i.e., conditioned on) a common standard treatment and so the parameters of the standard care model have the same interpretation as their counterparts in the relative treatment effects model. This condition is trivially met by alternative standardization/g-computation approaches that use only the trial data to produce absolute predictions from one single model, although as noted above these approaches make much stronger assumptions to transport absolute effects. Since they exhibit different assumptions, some researchers may wish to explore the use of both approaches as a triangulation exercise.

We focused on comparing two methods of calibration (regression-based and IOSW-based). However, it is also possible to combine both using what are termed doubly robust approaches where both regression and inverse-weighting are used together. For an example of this, see. Li et al. [26].

Butala et al. [12] suggested to trim the extremely large weights which may be caused by small sample size to ensure stable estimates. This can be achieved by truncating the top weights (such as 1%) or normalizing weights. In our additional analyses (Supplementary Table S12 and S13), the exclusion of individuals with largest 1% weights in the IOSW transportation slightly changed the point estimates, SEs and narrowed the CI. This 1% extreme large weights were characterized by older age, higher loop diuretics doses, and lower eGFR (Figure S5).

A challenge of calibration is the need to access individual participant data (IPD) for both the trial and target populations. This is complex (e.g., data sharing agreements and regulatory approvals) and requires considerable analyst time. However, as illustrated in Figure S2 in the supplementary, changes in trial reporting could improve this situation. Were trialists to provide the coefficients and the variance-covariance matrix for a treatment effect model including all nonnegligible treatment-covariate interactions, secondary researchers (with access to registry IPD) could produce calibrated treatment effect estimates for specific target settings. To enable such an approach would also require trialists to select the relevant covariates and to correctly specify the treatment covariate analysis. To be widely practiced, it would likely also require consensus among trialists and guidance from regulatory agencies. Similarly, it may also be possible in the future for estimates to be produced by trialists if those managing disease registries (such as NHSGGC) were able to provide adequate summary data to reconstruct the joint distribution of patient characteristics. We illustrate some of the information that would be need in Supplementary Table S14 for HF clinical trials (age, sex, SBP, etc.). As has previously been shown, joint covariate distributions may be reconstructed from routinely collected data given published marginal summary statistics (e.g. means, standard deviations) and correlation matrices if we are willing to make assumptions about the functional form of the marginal distributions and the correlation structure, for example, by using a multivariate normal a copula to capture the correlation structure [27]. Moreover, simulation studies have shown that the results are likely to be robust to the assumptions used to reconstruct the joint distribution [28]. However, for such an approach to be adopted, additional methodological work is first needed to i) reassure those holding routinely collected data that the risk of reidentifying individuals is sufficiently low and ii) reassure analysts that this parametric summary of the data is generally adequate for trial calibration. For the widespread adoption of transportation, the reporting of such summaries would need to be standardized [29]. Clinical trials are already highly standardized and sophisticated with mature ontologies and reporting standards [30]. These would need to be expanded to cover reporting of treatment-covariate interactions from multivariable models. Current proposals to standardize and harmonize HF registries would also need to incorporate reporting standards for population summaries. Considerable efforts by the HF research community would be required to implement such changes in both trial and registry settings. Our observation that calibration yielded more applicable estimates with only a moderate loss of precision suggests that this effort is worthwhile.

There are several important limitations in our analysis. We used routine data to define the target population because it was highly representative of patients encountered in clinical practice. However, some important variables were incompletely recorded, such as the New York Heart Association classification and LVEF and therefore could not be included in the transportation. Although a numerical value for LVEF was available for only 15% of patients, a semiquantitative measure of LVEF was available and indicated that patients in the registry are predominantly HFrEF.

5. Conclusion

Calibration of HF trials to HF registry data is feasible and may be used, without breaking randomization, to help address concerns about the representativeness of trials to patient population encountered in clinical practice. Consideration should be given to trial reporting standards and harmonization of HF registry data to facilitate transportation of clinical trials into clinical practice.

CRediT authorship contribution statement

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

Declaration of competing interest

J.G.F.C. is supported by a British Heart Foundation Centre of Research Excellence award RE/18/6/34217; he has received personal honoraria for advisory boards and lectures from Abbott, Amgen, Astra-Zeneca, Bayer, Bristol Myers Squibb, Johnson & Johnson, Novartis, Medtronic, Myokardia, NI Medical, Pharmacosmos, Idorsia, Respicardia, Servier, Torrent, Vifor, and Viscardia; nonfinancial support from Boehringer-Ingelheim and Boston Scientific; and research funding for his institution from Bayer, Bristol Myers Squibb, Medtronic, and Vifor. All other authors have nothing to disclose.

Acknowledgment

We acknowledge the SafeHaven support team at Robertson Center for Biostatistics in the University of Glasgow and NHSGGC for their prompt responsiveness to requests and assistance.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2023.08.019.

References

[1] Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 2000;342(25):1887−92.

[2] Kilcher G, Hummel N, Didden EM, Egger M, Reichenbach S, GetReal Work P. Rheumatoid arthritis patients treated in trial and real world settings: comparison of randomized trials with registries. Rheumatology 2018;57(2):354−69.

[3] Ezekowitz JA, Hu J, Delgado D, Hernandez AF, Kaul P, Leader R, et al. Acute heart failure perspectives from a randomized trial and a simultaneous registry. Circ Heart Fail 2012;5(6):735−41.

[4] Sharma A, Ezekowitz JA. Similarities and differences in patient characteristics between heart failure registries versus clinical trials. Curr Heart Fail Rep 2013;10(4):373−9.

[5] Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. JAMA 2007;298:1209−12.

[6] Westreich D, Edwards JK, Lesko CR, Stuart E, Cole SR. Transportability of trial results using inverse odds of sampling weights. Am J Epidemiol 2017;186:1010−4.

[7] Frangakis C. The calibration of treatment effects from clinical trials to target populations. Clin Trials 2009;6:136−40.

[8] 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;38:200−11.

[9] Stuart EA, Ackerman B, Westreich D. Generalizability of randomized trial results to target populations: design and analysis possibilities. Res Soc Work Pract 2018;28(5):532−7.

[10] Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations. Am J Epidemiol 2010;172:107−15.

[11] 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.

[12] Butala NM, Faridi KF, Tamez H, Strom JB, Song Y, Shen CY, et al. Estimation of DAPT study treatment effects in contemporary clinical practice: findings from the EXTEND-DAPT study. Circulation 2022;145:97−106.

[13] Bentley C, Cressman S, van der Hoek K, Arts K, Dancey J, Peacock S. Conducting clinical trials-costs, impacts, and the value of clinical trials networks: a scoping review. Clin Trials 2019;16:183−93.

[14] Poole-Wilson PA, Swedberg K, Cleland JGF, Di Lenarda A, Hanrath P, Komajda M, et al. Comparison of carvedilol and metoprolol on clinical outcomes in patients with chronic heart failure in the Carvedilol or Metoprolol, European Trial (COMET): randomised controlled trial. Lancet 2003;362:7−13.

[15] Poole-Wilson PA, Cleland JGF, Di Lenarda A, Hanrath P, Komajda M, Metra M, et al. Rationale and design of the Carvedilol or Metoprolol European Trial in patients with chronic heart failure: COMET. Eur J Heart Fail 2002;4(3):321−9.

[16] Abernathy GT, Abrams J, Akhtar S, Albitar I, Amidi M, Anand IS, et al. Rationale, design, implementation, and baseline characteristics of patients in the DIG trial: a large, simple, long-term trial to evaluate the effect of digitalis on mortality in heart failure. Control Clin Trials 1996;17:77−97.

[17] Perry G, Brown E, Thornton R, Shiva T, Hubbard J, Reddy KR, et al. The effect of digoxin on mortality and morbidity in patients with heart failure. N Engl J Med 1997;336(8):525−33.

[18] NHS greater glasgow and clyde. Available at https://www.nhsggc.org.uk/about-us/who-we-are-what-we-do/. Accessed September 26, 2023.

[19] Package 'mice'. Available at https://cran.r-project.org/web/packages/mice/mice.pdf. Accessed September 26, 2023.

[20] Jackson C, Metcalfe P, Amdahl J, Warkentin MT, Kunzmann K. flexsurv: flexible parametric survival and multi-state models 2021. Available at https://cran.r-project.org/web/packages/flexsurv/index.html. Accessed September 26, 2023.

[21] Lumley T. Package 'survey' 2021. Available at https://cran.r-project.org/web/packages/survey/survey.pdf. Accessed September 26, 2023.

[22] Loudon K, Treweek S, Sullivan F, Donnan P, Thorpe KE, Zwarenstein M. The PRECIS-2 tool: designing trials that are fit for purpose. BMJ 2015;350:h2147.

[23] Stuermer T, Wyss R, Glynn RJ, Brookhart MA. Propensity scores for confounder adjustment when assessing the effects of medical interventions using nonexperimental study designs. J Intern Med 2014;275(6):570−80.

[24] Dias S, Welton NJ, Sutton AJ, Ades AE. Evidence synthesis for decision making 5: the baseline natural history model. Med Decis Making 2013;33:657−70.

[25] Dahabreh IJ, Robertson SE, Steingrimsson JA, Stuart EA, Hernan MA. Extending inferences from a randomized trial to a new target population. Stat Med 2020;39:1999−2014.

[26] Li X, Shen C. Doubly robust estimation of causal effect: upping the odds of getting the right answers. Circ Cardiovasc Qual Outcomes 2020;13(1):e006065.

[27] Phillippo DM, Dias S, Ades AE, Belger M, Brnabic A, Schacht A, et al. Multilevel network meta-regression for population-adjusted treatment comparisons. J R Stat Soc Ser A Stat Soc 2020;183(3):1189−210.

[28] Phillippo DM, Dias S, Ades AE, Welton NJ. Assessing the performance of population adjustment methods for anchored indirect comparisons: a simulation study. Stat Med 2020;39:4885−911.

[29] Aktaa S, Batra G, Cleland JGF, Coats A, Lund LH, McDonagh T, et al. Data standards for heart failure: the European unified registries for heart care evaluation and randomized trials (EuroHeart). Eur Heart J 2022;43:2185−95.

[30] CDISC standards in the clinical research process.. Available at https://www.cdisc.org/standards. Accessed September 26, 2023.