Detecting Anomalies Among Practice Sites Within Multicenter Trials

An Application of Transportability Methods to the TOPCAT and ACCORD BP Trials

Seth A. Berkowitz, MD, MPH
Kara E. Rudolph, PhD
Sanjay Basu, MD, PhD

BACKGROUND: Recent multisite trials reveal striking heterogeneities in results between trial sites. These may be because of population differences indicating different treatment benefits among different types of participants or site anomalies, such as failures to adhere to study protocols that could negatively affect study validity. We sought to determine whether a new data analysis strategy—transportability methods—could suggest site anomalies not readily identified through standard methods.

METHODS AND RESULTS: We applied transportability methods to 2 large, multicenter cardiovascular disease treatment trials: the TOPCAT trial (Treatment of Preserved Cardiac Function Heart Failure With an Aldosterone Antagonist; n=3445) comparing spironolactone to placebo for heart failure (for which site anomalies were suspected) and the ACCORD BP trial (Action to Control Cardiovascular Risk in Diabetes–Blood Pressure; n=4733) comparing intensive-to-standard blood pressure treatment (for which site anomalies were not suspected). The transportability methods give expected results by standardizing from one site to another using data on participant covariates. The difference between the expected and observed results was assessed using calibration tests to identify whether treatment-effect differences between sites could be explained by participant population characteristics. Standard regression methods did not detect heterogeneities in TOPCAT between Russia/Georgia study sites suspected of study protocol violations and sites in the Americas (P=0.12 for difference in primary cardiovascular outcome; P=0.20 for difference in total mortality). The transportability methods, however, detected the difference between Russia/Georgia sites and sites in the Americas (P<0.001) and found that measured participant characteristics did not explain the between-site discrepancies. The transport methods found no such discrepancies between sites in ACCORD BP, suggesting participant characteristics explained between-site differences.

CONCLUSIONS: Transportability methods may be superior to standard approaches for detecting anomalies within multicenter randomized trials and assist data monitoring boards to determine whether important treatment-effect heterogeneities can be attributed to participant differences or potentially to site performance differences requiring further investigation.

Key Words: blood pressure ■ cardiovascular system ■ heart failure ■ methodology ■ reproducibility

Multisite trials that include international populations are an important method to rigorously evaluate clinical treatment strategies. However, it can be difficult for central data monitoring boards to ensure the quality of large, multicenter trials,[1] and the trials can be subject to heterogeneity across sites. This heterogeneity may be because of differences in how different populations benefit from the study intervention (an important scientific finding) or may be because of differences in how sites execute the trial protocol.[2–6] These execution differences may be legitimate variation in interpretation or implementation of the protocol or, of concern, represent fraudulent data generation and improper intervention delivery to participants.[3–5]

At present, the most common approach for monitoring heterogeneity in site-specific results is to statistically estimate site-by-treatment group interaction terms in a regression model, to determine whether the average treatment effect estimated at a given site differs from the average treatment effect estimated from the trial as a whole. But such an approach may have insufficient power to detect meaningful differences and does not clarify whether differences in characteristics of the participants from that site can explain any between-site heterogeneity that is detected. Recent methodological innovations known as transportability methods may offer improvements over this standard approach.[7–10]

To help understand the need for transportability methods, we first note that the average treatment effect in one setting, estimated by a randomized clinical trial, is the difference (or ratio, depending on effect type) in outcomes comparing treated to untreated participants. Thus, it is specific to the particular distribution of potential effect modifiers and prognostic factors (eg, age, sex, race/ethnicity, or geographic location), of those included in the trial. However, the sample participating in a trial may be distinct from the population we are trying to understand. This can raise questions about the generalizability or transportability of the findings. Transportability methods utilize understanding of the differences between populations and the conditions that license the extrapolation of causal effects to the population we are trying to understand. For example, we can use transportability methods to model the relationship between relevant covariates and a given treatment-outcome relationship in a particular setting (the source site) and then use those models to project what the outcome would be were the treatment applied to a setting with a different distribution of covariates (the target site). Hence, the methods seek to transport estimates of treatment effects from one setting to another. Such estimates can be useful if we wish to apply results from a clinical trial to a population with a different distribution of covariates. In addition, the transported estimates may allow one to determine whether results from certain sites within a multisite trial differ more than would be expected from results at other sites within the trial.

In this study, we used transportability methods to identify whether they could detect potential site anomalies in the TOPCAT trial (Treatment of Preserved Cardiac Function Heart Failure With an Aldosterone Antagonist). TOPCAT was an NIH-sponsored study of spironolactone therapy in individuals with heart failure with preserved ejection fraction.[2] Although the study did not find that spironolactone reduced the occurrence of the primary outcome, subsequent investigations suggested there were abnormalities when comparing sites in Russia/Georgia and sites in the Americas. We additionally sought to determine whether the transportability methods would be overly sensitive (and trigger false-positive warnings) about site variability by applying transportability methods to a trial for which site differences were not anticipated to explain effect size heterogeneity (ACCORD BP trial [Action to Control Cardiovascular Risk in Diabetes–Blood Pressure] of intensive versus standard blood pressure treatment).[11]

METHODS

Data Source and Description

Data for this study came from the public release individual participant data files for the TOPCAT and ACCORD BP trials, available from the Biologic Specimen and Data Repository Information Coordinating Center of the National Heart, Lung, and Blood Institute. The TOPCAT study was a randomized, multisite, clinical trial comparing the use of spironolactone versus placebo in individuals with a history of heart failure with preserved ejection fraction. Individuals >50 years of age were eligible if they reported at least 1 sign and 1 symptom of heart failure, had a left ventricular ejection fraction >45%, controlled systolic blood pressure, and serum potassium <5.0 mmol/L. Further, eligible patients were required to have had a hospitalization for heart failure within the last 12 months or an elevated BNP (brain natriuretic peptide) or N-terminal pro-BNP level (or both).[2] Exclusion criteria were limited life expectancy, severe renal dysfunction, and other comorbidities.[2] The median study follow-up time for the primary outcome was 3.0 years.[2] A full study protocol and primary and secondary results have been published. Sites within TOPCAT were located in the United States, Canada, Argentina, and Brazil (the Americas) and in Russia and the Republic of Georgia. The primary result of the TOPCAT trial was that spironolactone was not superior to placebo, but subsequent analyses found that spironolactone metabolite was more often absent from individuals who reported taking spironolactone enrolled at Russian sites, compared with study sites in the Americas (≈30% versus ≈3%).[4] Further, spironolactone did not produce a significant reduction in the risk of the primary outcome among patients in the Russia/Georgia sites but did produce a significant reduction elsewhere (hazard ratio, 1.10; 95% CI, 0.79–1.51 in Russia/Georgia, versus hazard ratio, 0.82; 95% CI, 0.69–0.98 elsewhere).[3] Despite this, the site-by-treatment interaction tested was not significant.[3]

ACCORD BP was a randomized, multicenter, open-label trial of intensive (target systolic blood pressure <120 mm Hg) versus standard blood pressure treatment (target systolic blood pressure <140 mm Hg) among adults with type 2 diabetes mellitus, conducted at 77 clinical sites in the United States and Canada between January 2003 and June 2009, with a mean follow-up of 4.7 years.[11] Inclusion criteria for the ACCORD BP trial included age at least 40 years with cardiovascular disease or at least 55 years with anatomic evidence of substantial atherosclerosis, albuminuria, left ventricular hypertrophy, or at least 2 additional cardiovascular disease risk factors (dyslipidemia, hypertension, smoking, or obesity); systolic blood pressure of 130 to 180 mm Hg taking ≤3 blood pressure agents and having a 24-hour protein excretion rate <1 g; and type 2 diabetes mellitus with a hemoglobin A1c level of at least 7.5%. Exclusion criteria included having a body mass index >45 kg/m^2, serum creatinine >1.5 mg/dL, or other serious illness.[11]

The institutional review board at the University of North Carolina at Chapel Hill decided that approval was not required for this secondary analysis of deidentified data.

Outcomes

The primary outcome for the TOPCAT analysis in this study was the same as in the original TOPCAT trial—a composite outcome of death from cardiovascular causes, aborted cardiac arrest, or hospitalization for the management of heart failure.[2] The secondary outcome in this study was total (all-cause) mortality.[2]

For the ACCORD BP analyses, we used the ACCORD BP primary outcome (a composite of nonfatal myocardial infarction, nonfatal stroke, and death from cardiovascular causes).[11]

Covariates

We considered an extensive set of covariates that may differ between study sites and thus may explain differences in the observed treatment-effect estimates. All data were taken from baseline data in the TOPCAT public data release, and details of their assessment and measures are provided in the study documentation. For our main TOPCAT analysis, we considered a set of variables that, based on prior literature regarding clinical outcomes in individuals with heart failure with preserved ejection fraction, we hypothesized could be related to differences in observed outcomes across sites.[12–14] These variables were age (years), sex, race/ethnicity (non-Hispanic white, non-Hispanic black, Hispanic, or Asian/multi/other), history of congestive heart failure hospitalization, history of implantable cardioverter defibrillator placement, use of angiotensin-converting enzyme inhibitor or angiotensin receptor blocker, functional status as indicated by New York Heart Association heart failure class (I or II versus III or IV), systolic blood pressure, estimated glomerular filtration rate (using the modification of diet in renal disease equation), serum potassium level, study eligibility via hospitalization, and study eligibility via BNP value.

As robustness checks, we also considered an extended set of variables, described in the Appendix in the Data Supplement.

Variables used for modeling the outcome in the ACCORD BP analyses were study arm, age (years), sex, race/ethnicity (non-Hispanic white, non-Hispanic black, Hispanic, or Asian/multi/other), educational attainment (categorized as less than high school diploma, high school diploma, some college, or college degree and higher), history of cardiovascular disease, years with diabetes mellitus, smoking status, body mass index, hemoglobin A1c level, systolic blood pressure, diastolic blood pressure, fasting plasma glucose, estimated glomerular filtration rate, urine albumin-to-creatinine ratio, total cholesterol, triglycerides, high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, health insurance status, history of myocardial infarction, history of stroke, history of congestive heart failure, and study glycemia treatment arm (because ACCORD was a factorial design in which participants were also randomized to intensive or standard glycemic control for type 2 diabetes mellitus).

Transportability Approach

We used a doubly robust, semiparametric targeted maximum-likelihood estimation (TMLE) transport estimator to predict the intent-to-treat effect in a target site, using data from both the target-site participants and a source-site treatment-effect estimate.[7,8,15] The approach models how participant covariates relate to the outcome in the source site and how covariates that may affect the outcome differ between the source and the target site. These models are then used to predict what results would be expected in the target site by standardizing the results of the source site over the covariate distribution in the target site (see conceptual illustration in Figure I in the Data Supplement). The mathematical details of the transported intent-to-treat estimated have been published previously.[7] We inferred study site from the participant's country of residence. Because transport formulae for time-to-event data have not yet been developed, we conducted our analyses using a dichotomous outcome (whether or not a person had the outcome by 36 months, with 36 months approximately being the median follow-up time) using logistic regression. Individuals who were censored before 36 months had their last outcome observation carried forward to 36 months and were retained in the analysis. Before using this approach for estimating the transport equations, we checked whether this was a reasonable approximation. The logistic regression analysis in TOPCAT data, with the primary outcome dichotomized at 36 months, produced an odds ratio of 0.86 (95% CI, 0.71–1.03), which is similar to the estimate using Cox proportional hazards regression (hazard ratio, 0.89; 95% CI, 0.77–1.04) reported in the main TOPCAT analysis.[2]

For the ACCORD BP analyses, we similarly dichotomized the primary outcome (whether or not a person had the outcome by 60 months, with 60 months being again approximately the median follow-up time) and compared results from a logistic regression model to those of the Cox proportional hazards model used as the primary analysis for the ACCORD BP study. The logistic regression model yielded an odds ratio of 0.87 (95% CI, 0.71–1.07), which was similar to the estimate from the Cox regression (hazard ratio, 0.88; 95% CI, 0.73–1.07).

Assumptions of Transportability Methods

To estimate the expected intention-to-treat average treatment effect in a target site using data from a source site, 3 assumptions need to be met. The first assumption is that of a

common outcome model. This can be expressed as E_0 (Y|S=0, W, A)=E_0 (Y|S=1, W, A). See Figure II in the Data Supplement for a graphical depiction of this. In words, this means that the mean value of the outcome (Y) in the source site (S=0) with its distribution of covariates (W) and treatment (A) would be the same as the mean value of the outcome at the target site (S=1) given the same covariates and treatment. The second assumption is that there are no unobserved confounders. This means that assignment of treatment (A) is independent of potential outcomes given observed covariates at the source site (S=0). Because both of the datasets analyzed in this study were randomized clinical trials, randomization provides this independence. Finally, the third assumption is positivity. This means that there is a nonzero probability of selection into a particular site and level of treatment given the observed covariates. In this study, the data are from randomized clinical trials, which helps ensure that, at least theoretically, all combinations of covariates included in the study have a nonzero chance at being assigned to a given treatment condition. In practice, however, rare combinations of covariates may not occur in all treatment levels.

Statistical Analysis

We first used data from the TOPCAT sites in the Americas and the TMLE transport estimator to generate an intent-to-treat estimate of expected study outcomes in the Russia/Georgia sites. These could then be compared against the observed outcomes at the Russia/Georgia sites. If the expected values were close to the observed outcomes, differences in observed patient characteristics could be responsible for the differences noted in Russia/Georgia treatment effects. If the expected values were different from the observed values, however, unobserved factors, such as site protocol adherence or unmeasured participant characteristics, could explain the differences in treatment effects.

Our analysis proceeded in 3 steps. First, we analyzed the data in the TOPCAT trial, stratified by site to generate the observed treatment-effect estimates by site. We expressed these values as cumulative incidence of the primary and of the secondary outcome at 36 months and the difference between study arms as risk differences at 36 months. We also replicated the Cox proportional hazard regression analysis conducted in the main trial analysis, adding an interaction term to determine whether site-level differences in treatment effect would have been detected using this approach. Finally, we used logistic regression analyses, adjusted for the same covariates as in the transport analyses, to test a site-by-treatment interaction term to determine whether this approach could detect site-level differences.

Next, we applied the TMLE transport estimator to generate estimates of the expected results, based on the distribution of covariates in study participants in the Russia/Georgia sites of TOPCAT. In addition to conducting robustness checks using additional variables in the TMLE transport equations, we also conducted a robustness check adjusting for follow-up time because our approach required the use of logistic regression rather than a time-to-event analysis.

Finally, we assessed the differences between observed and expected values for the outcome using calibration metrics (analogous to the comparison of observed and expected values for risk prediction models). We specifically applied the Hosmer-Lemeshow test,[16] which tests whether the observed event rates match the expected event rates by deciles of expected rates. We additionally applied the calibration belt approach proposed by Nattino et al[17] (the Italian Group for the Evaluation of the Interventions in Intensive Care Units [GiViTI] calibration test), which derives confidence bands around a polynomial fit between the expected and observed outcome rates.[18] For both the Hosmer-Lemeshow and GiViTI calibration test, the null hypothesis is that the model is well calibrated, so lower *P* values suggest a more significant difference between expected and observed values (more indication of site variations, rather than observed population variations, as explanations for treatment-effect heterogeneity between sites). In addition to testing the overall goodness of fit, we explored the stratified goodness of fit in the placebo groups and spironolactone groups individual. This would allow us to determine whether any discrepancies occurred in either or both arms of the trial.

In this study, we used the standard $P<0.05$ threshold to indicate statistical significance (and thus miscalibration between observed and expected event rates), but we believe it is worth noting that other thresholds may be useful. Particularly as our approach may be used to flag trial sites that need further investigation, rather than to definitely establish a discrepancy, using a higher threshold would increase sensitivity for detecting aberrations at the expense of decreasing specificity, which some data monitoring boards may desire. To that end, we estimated 80% CIs along with 95% CIs on calibration plots.

Falsification Testing

To test whether the methodology may be overly sensitive, identifying even slight and inconsequential variations across sites as being important for investigation, we repeated our analyses using a trial where no clinically meaningful differences across sites were expected based on prior monitoring: the ACCORD BP trial.

We selected ACCORD BP as a comparator trial because it had an approximately similar, but slightly larger number of participants, than TOPCAT (4733 in ACCORD BP versus 3445 in TOPCAT), such that ACCORD BP transport results should be more sensitive to minor deviations, and we would be modeling cardiovascular outcomes where the risk factors for the trial outcomes are thought to be well understood. We randomly selected 3 of the 7 clinical networks used in ACCORD BP to serve as the transport sites, with the remaining 4 serving as the source sites. The 3 selected sites represented a similar proportion of participants in ACCORD BP as the Russia/Georgia sites did in TOPCAT.

In the main analyses, 3434 of 3445 (99.7%) participants had complete data, so no imputation was used. In the first set of robustness checks, 3411 of 3445 (99.0%) had complete data, so we also did not use imputation. In the second set of robustness checks, 2827 of 3445 (82.1%) had complete data, owing to 18% missingness for the physical activity variable. Therefore, for the third set of robustness checks, we conducted both complete-case analyses and analyses using a nonparametric imputation method based on a random forest, called missForest.[19,20] For the ACCORD BP analysis, 4507 of 4733 (95.2%) observations had complete data, so no imputation was used.

Analyses were conducted in SAS, version 9.4 (SAS Institute, Cary, NC), and R, version 3.4.2 (R Foundation for Statistical Computing, Vienna, Austria). TOPCAT and ACCORD data are available from the National Heart, Lung, and Blood Institute Biologic Specimen and Data Repository Information Coordinating Center (https://biolincc.nhlbi.nih.gov/home/) under a data use agreement but cannot be shared by the authors. Statistical code for the transportability analyses was adapted from Rudolph and van der Laan[7] and will be avail-able, at time of publication, from the authors' website.[21]

RESULTS

Characteristics of the Participant Sample From TOPCAT

The demographic and clinical characteristics of participants in TOPCAT, both overall and stratified by site (Russia/Georgia versus all other sites) are presented in Table 1 and an extended set of demographics in Table I in the Data Supplement. Overall, there were many significant differences between the samples at the Russia/Georgia site versus the other site. Participants at the Russia/Georgia site were younger, more commonly of non-Hispanic white race, had higher prevalence of previous heart failure hospitalization, had higher baseline systolic blood pressure, and had better baseline renal function.

Primary and Secondary Outcomes Across Study Sites

Across all TOPCAT sites, 16.8% of individuals in the placebo group and 14.8% of individuals in the spironolactone group experienced the primary outcome by 36 months. At the Russia/Georgia sites, 6.4% of the placebo group and 6.5% of the spironolactone group experienced the primary outcome; at the other sites, 26.8% of the placebo group and 22.6% of the spironolactone group experienced the primary outcome. A similar discrepancy was present for the secondary outcome of total mortality (Table 2).

Table 1. Demographics Overall and by Study Site

	Overall	Russia/ Georgia Site	Other Countries	
	n=3445	n=1678	n=1767	
	Mean (SD) or n (%)	Mean (SD) or n (%)	Mean (SD) or n (%)	*P* Value
Age at study entry, y	68.56 (9.59)	65.44 (8.42)	71.52 (9.69)	<0.001
Women	1775 (51.5)	893 (53.2)	882 (49.9)	0.057
Race/ethnicity				<0.001
Non-Hispanic white	2824 (82.0)	1675 (99.8)	1149 (65.0)	
Non-Hispanic black	269 (7.8)	0 (0.0)	269 (15.2)	
Hispanic	321 (9.3)	3 (0.2)	318 (18.0)	
Asian/multi/other	31 (0.9)	0 (0.0)	31 (1.8)	
Country				NA
United States	1151 (33.4)	0 (0.0)	1151 (65.1)	
Canada	326 (9.5)	0 (0.0)	326 (18.4)	
Russia	1066 (30.9)	1066 (63.5)	0 (0.0)	
Republic of Georgia	612 (17.8)	612 (36.5)	0 (0.0)	
Brazil	167 (4.8)	0 (0.0)	167 (9.5)	
Argentina	123 (3.6)	0 (0.0)	123 (7.0)	
Assigned to spironolactone	1722 (50.0)	836 (49.8)	886 (50.1)	0.878
History of CHF hospitalization	2489 (72.3)	1449 (86.4)	1040 (58.9)	<0.001
History of ICD placement	44 (1.3)	2 (0.1)	42 (2.4)	<0.001
Systolic blood pressure	129.22 (13.97)	131.00 (11.37)	127.52 (15.87)	<0.001
NYHA class III or IV heart failure at baseline	1136 (33.0)	516 (30.8)	620 (35.2)	0.007
Use of ACE inhibitor/ARB	2900 (84.3)	1505 (89.7)	1395 (79.0)	<0.001
Estimated glomerular filtration rate, mL/min	67.67 (20.15)	71.03 (18.03)	64.47 (21.50)	<0.001
Serum potassium, mmol/L	4.25 (0.45)	4.32 (0.45)	4.19 (0.43)	<0.001
Met hospitalization inclusion criterion	2464 (71.5)	1488 (88.7)	976 (55.2)	<0.001
Met BNP inclusion criterion	1444 (41.9)	306 (18.2)	1138 (64.4)	<0.001

ACE indicates angiotensin-converting enzyme; ARB, angiotensin receptor blocker; BNP, brain natriuretic peptide; CHF, congestive heart failure; ICD, implantable cardioverter defibrillator; and NYHA, New York Heart Association.

Table 2. Observed and Expected Incidence of Primary Outcome and Total Mortality at 36 mo in TOPCAT

	Observed				Expected			
	Primary Outcome		Total Mortality		Primary Outcome		Total Mortality	
	n (%)	Risk Difference	n (%)	Risk Difference	n (%)	Risk Difference	n (%)	Risk Difference
All sites								
Placebo	290 (16.83)	…	192 (11.14)	…	…		…	
Spironolactone	254 (14.75)	−2.1%	164 (9.52)	−1.6%	…		…	
Russian/Georgian sites								
Placebo	54 (6.41)	…	46 (5.46)	…	225 (26.8)	…	123 (14.63)	…
Spironolactone	54 (6.46)	0.05%	42 (5.02)	−0.4%	134 (16.0)	−10.7%	77 (9.21)	−5.4%
All other sites								
Placebo	236 (26.79)	…	146 (16.57)	…	…		…	
Spironolactone	200 (22.57)	−4.2%	122 (13.77)	−2.8%	…		…	

TOPCAT indicates Treatment of Preserved Cardiac Function Heart Failure With an Aldosterone Antagonist.

Standard Analysis

To investigate whether a discrepancy across sites could have been found using a standard statistical approach, we fit a Cox proportional hazards regression (the analysis strategy specified in the trial protocol) with terms for site (Russia/Georgia versus the Americas), treatment (spironolactone versus placebo), and site-by-treatment interaction. The interaction term was not significant in the analysis of either the primary outcome (P=0.12) or total mortality (P=0.20), indicating that this approach did not detect differences across sites. In addition, we conducted a logistic regression analysis, adjusting for the same factors used in the transport analysis, to test a site-by-treatment interaction term. This interaction term was not significant when analyzing either the primary outcome (P=0.17) or total mortality (P=0.42) either, indicating that this approach also did not detect differences across sites.

Transport Analysis

TMLE transport analyses revealed that expected rates for the primary and secondary outcome, based on the demographic and clinical characteristics of the participants, were actually higher in the Russia/Georgia sites than the other sites (Table 2). For example, we would have expected 26.8% of individuals in the placebo group in the Russia/Georgia site to have experienced the primary outcome, based on their demographic and clinical characteristics, instead of the 6.4% who actually did, suggesting that observed participant characteristics included in our model were unlikely to explain the Russia/Georgia treatment-effect results. Goodness-of-fit testing showed that the expected values did not match those observed (Hosmer-Lemeshow test P <0.001 and GiViTI calibration test statistic <0.001 for both the primary outcome and total mortality). As Figures 1 and 2 show, expected outcome rates were significantly greater than observed outcome rates across all levels of cardiovascular event risk for both outcomes. In robustness checks adjusting for follow-up time, using additional covariates, or imputing missing values, the Hosmer-Lemeshow test and GiViTI calibration test P strongly indicated lack of fit in all cases, supporting our finding that the Russia/Georgia participant characteristics did not explain the heterogeneity in those sites' observations (and was instead potentially because of protocol violations) across the different specifications (Table II in the Data Supplement; Figures III through VI in the Data Supplement).

Finally, we explored whether the lack of fit between observed and expected values in the Russia/Georgia sites of TOPCAT occurred in the placebo group, spironolactone group, or both. We did this by conducting goodness-of-fit testing stratified by treatment group. We found that there was a lack of fit for both the placebo and spironolactone groups (Table III in the Data Supplement; Figures VII through X in the Data Supplement).

Falsification Testing

In falsification testing, there were differences across sites in ACCORD BP regarding a number of factors—notably race/ethnicity, education, history of cardiovascular disease, and hemoglobin A1c (Table IV in the Data Supplement). Despite this, however, we found that the observed outcomes matched the expected derived using the transport method (Hosmer-Lemeshow test, P=0.21; GiViTI calibration test, P=0.17; Figure XI in the Data Supplement), meaning that site differences could largely be explained by participant characteristic variations (Table 3).

DISCUSSION

In this study, we found that transport analyses detected anomalies in the Russia/Georgia sites of the TOPCAT

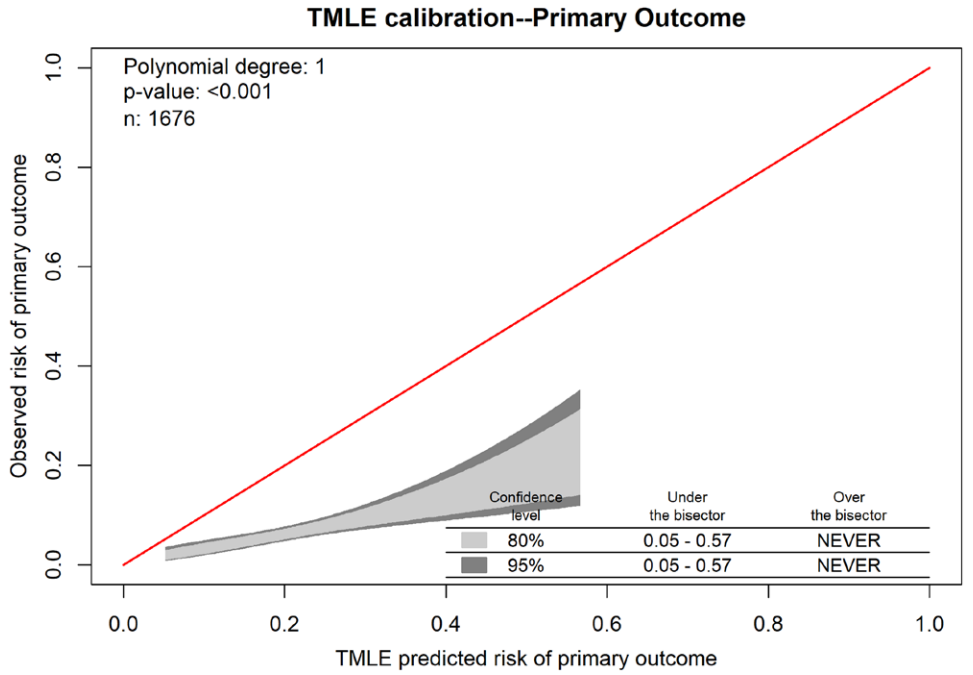

Figure 1. Calibration plot for primary outcome in the TOPCAT trial (Treatment of Preserved Cardiac Function Heart Failure With an Aldosterone Antagonist) Russia/Georgia sites.

A comparison of transported predicted primary outcome rates at Russia/Georgia sites based on results from other sites, vs observed outcomes in the TOPCAT trial. The diagonal bisecting line represents where observed equals expected outcome rates at every risk level and thus where the participant characteristics would be expected to explain heterogeneity in treatment effects between sites rather than site-specific anomalies in study protocol or other unobserved factors influencing the results. Light gray bands indicate the 80% confidence region, and darker gray bands represent the region of 95% confidence. Areas below the bisecting line indicate that predicted risk was higher than observed and vice versa. Inset chart shows the specific levels of predicted risk and whether they were over or under the bisecting line. Probability levels not in chart (eg, <0.05 or >0.57) were not present in this study and thus are not plotted. TMLE indicates targeted maximum-likelihood estimation.

study—specifically that observed participant characteristics did not explain differences in the treatment effects observed between the Russia/Georgia sites and sites in the Americas. Standard site-by-treatment interaction testing did not detect these differences. The transportability method did not appear overly sensitive to site variations when additionally tested in the ACCORD BP trial, where the methods suggested that between-site variations were attributable to differences in participant characteristics.

The transportability method adds important insights to the existing literature on the conduct of large multicenter trials. First, it offers the opportunity to contextualize and evaluate whether heterogeneity in treatment effects may be because of important observed participant characteristics that may modify the impact of therapy and, therefore, would be important for practitioners generalizing the results of a trial to their patient populations. Second, it offers a warning flag for post hoc evaluation of a trial where study sites may substantially differ in effect size estimates for reasons other than observed participant characteristics. The availability of such methods, coupled with the increased sharing of clinical trial data, may assist trialists who have identified that management and monitoring of international multisite trials is particularly important but challenging.[1,3–5] Third, the method can quantify whether the observed treatment effects are larger or smaller than the expected effects adjusted for population characteristics, which may help identify the direction and magnitude of the problem.

Table 3. Observed and Expected Incidence of Primary Outcome and Total Mortality at 36 mo in TOPCAT

	Observed		Expected	
	Primary Outcome		Primary Outcome	
	n (%)	Risk Difference	n (%)	Risk Difference
All sites				
Standard control	209 (8.81)	…	…	
Intensive control	185 (7.83)	–0.98	…	
Transport sites				
Standard control	71 (7.40)	…	85 (9.00)	…
Intensive control	73 (7.73)	0.33	80 (8.31)	–0.66
Source sites				
Standard control	138 (9.77)	…	…	
Intensive control	112 (7.90)	–1.87	…	

Randomly selected transport sites were networks 1, 2, and 3 (further characterization is not provided in publicly available data to preserve participant anonymity). Source sites were networks 4 to 7. TOPCAT indicates Treatment of Preserved Cardiac Function Heart Failure With an Aldosterone Antagonist.

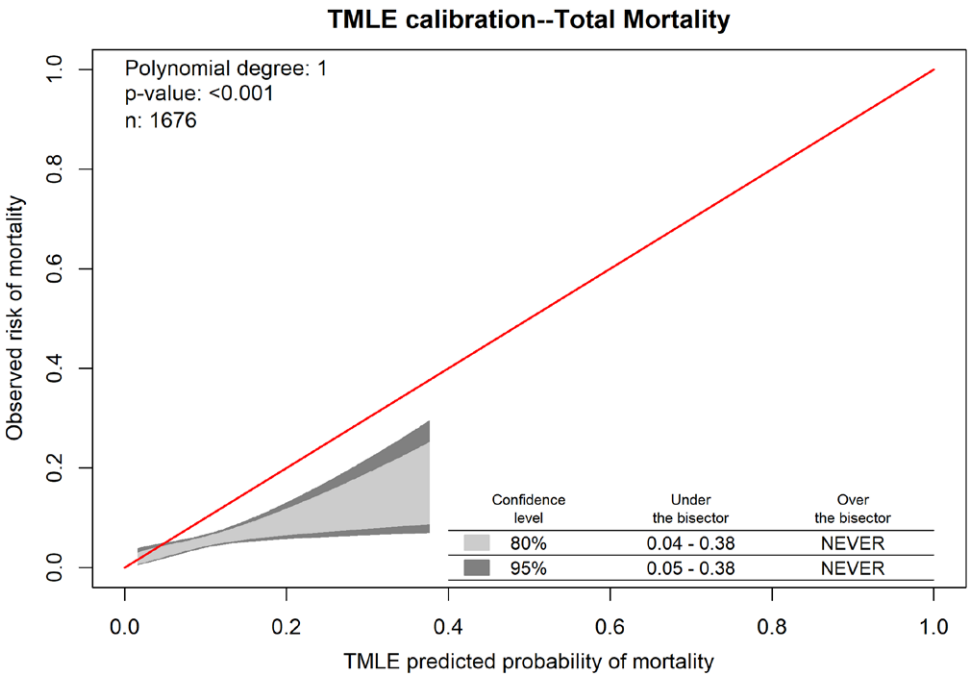

Figure 2. Calibration plot of total mortality in the TOPCAT trial (Treatment of Preserved Cardiac Function Heart Failure With an Aldosterone Antagonist) Russia/Georgia sites.

A comparison of transported predicted primary outcome rates at Russia/Georgia sites based on results from other sites, vs observed outcomes in the TOPCAT trial. The diagonal bisecting line represents where observed equals expected outcome rates at every risk level and thus where the participant characteristics would be expected to explain heterogeneity in treatment effects between sites rather than site-specific anomalies in study protocol or other unobserved factors influencing the results. Light gray bands indicate the 80% confidence region, and darker gray bands represent the region of 95% confidence. Areas below the bisecting line indicate that predicted risk was higher than observed and vice versa. Inset chart shows the specific levels of predicted risk and whether they were over or under the bisecting line. Probability levels not in chart (eg, <0.04 or >0.38) were not present in this study and thus are not plotted. TMLE indicates targeted maximum-likelihood estimation.

Nevertheless, there are important limitations to the approach. Most importantly, transportability of a result from one site to other depends on the choice of participant covariates to standardize against and thus can only incorporate observed patient features. Clinical trials may not measure all relevant characteristics, and so important factors that influence the treatment effect may be unmeasured. In this case, site variations could lead to suspicions that sites did not follow protocols, when in fact unknown or unmeasured factors caused the site's population to be systematically different from the sample at other sites. Therefore, variations detected by transportability methods should serve only as a prompt for further investigation. Further, using the transportability methods in this way assumes that the relationship between baseline characteristics and the treatment on the outcome is similar enough across sites to be able to use the relationship at one site to predict the outcomes of another site. A multisite trial inherently makes a different, potentially stronger assumption—that there is a common overall treatment effect on the outcome (independent of covariates) across sites—or else one could not meaningfully pool the results to estimate a single average treatment effect of the intervention in the trial. If this assumption does not hold, then the differences in this relationship may explain any variation in outcomes observed. Second, our test here applied the method to only 2 trials. More subtle and smaller sample trials may need to be considered in the future to identify transportability method limitations. Third, the evidence presented here was to help post hoc trial analyses when heterogeneity in treatment effects across sites is observed or suspected. The results do not provide clear guidance for researchers performing interim analysis as part of the data monitoring team, which presents the difficulty of both false-positive findings and statistical power. A future analysis using simulations would be appropriate to decipher the power of traditional and novel transportability methods to determine the degree to which the method should be applied to interim monitoring exercises. Finally, because of the novelty of transportability methods, we used a dichotomous (as opposed to time-to-event analysis) outcome analysis strategy.

Given the ubiquity of multisite trials, the significant resources invested in them, and the tendency to prefer results from randomized trials over other study designs, confidence in trial results is of the utmost importance. The movement toward making individual patient data available, when coupled with innovative analytic techniques, such as transportability methods, may be an important way to increase our confidence in the results of

randomized trials and ultimately improve patient care by basing our treatments on the best available evidence.

ARTICLE INFORMATION

Received May 21, 2018; accepted January 16, 2019.

The Data Supplement is available at https://www.ahajournals.org/doi/suppl/10.1161/CIRCOUTCOMES.118.004907.

Correspondence

Seth A. Berkowitz, MD, MPH, Division of General Medicine and Clinical Epidemiology, Department of Medicine, University of North Carolina School of Medicine, 5034 Old Clinic Bldg, CB 7110, Chapel Hill, NC 27599. Email seth_berkowitz@med.unc.edu

Affiliations

Division of General Medicine and Clinical Epidemiology, Department of Medicine, University of North Carolina School of Medicine, Chapel Hill (S.A.B.). Cecil G. Sheps Center for Health Services Research, University of North Carolina at Chapel Hill (S.A.B.). Department of Emergency Medicine, School of Medicine, University of California, Davis, Sacramento (K.E.R.). Center for Primary Care and Outcomes Research (S.B.), Center for Population Health Sciences (S.B.), Department of Medicine (S.B.), and Department of Health Research and Policy (S.B.), Stanford University, CA. Center for Primary Care, Harvard Medical School, Boston, MA (S.B.).

Acknowledgments

This article was prepared using TOPCAT (Treatment of Preserved Cardiac Function Heart Failure With an Aldosterone Antagonist) and ACCORD (Action to Control Cardiovascular Risk in Diabetes) research materials obtained from the National Heart, Lung, and Blood Institute (NHLBI) Biologic Specimen and Data Repository Information Coordinating Center and does not necessarily reflect the opinion or views of the ACCORD or TOPCAT studies or the NHLBI. We also thank 2 anonymous reviewers for helpful comments incorporated into the article. Dr Berkowitz had full access to all of the data in the study and takes full responsibility for the work as a whole, including the study design, access to data, the integrity of the data, the accuracy of the data analysis, and the decision to submit and publish the manuscript. All authors had access to the data and agree to submission of the manuscript for publication. Dr Berkowitz affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that there are no discrepancies from the study as originally planned. Dr Basu conceived the study and revised the manuscript for critical intellectual content. Dr Berkowitz made significant contributions to the design of the study, conducted analysis of the data, and drafted the manuscript. Dr Rudolph made significant intellectual contributions to the design of the study and revised the manuscript for critical intellectual content.

Sources of Funding

Research reported in this publication was supported by the National Institute for Diabetes and Digestive and Kidney Disease of the National Institutes of Health, the National Institute on Minority Health and Health Disparities, and the National Institute on Drug Abuse of the National Institutes of Health under award numbers DP2MD010478 (Dr Basu), U54MD010724 (Dr Basu), K23DK109200 (Dr Berkowitz), and R00DA042127 (Dr Rudolph). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the manuscript for publication.

Disclosures

None.

REFERENCES

1. Bristow MR, Enciso JS, Gersh BJ, Grady C, Rice MM, Singh S, Sopko G, Boineau R, Rosenberg Y, Greenberg BH. Detection and management of geographic disparities in the TOPCAT trial: lessons learned and derivative recommendations. *JACC Basic Transl Sci*. 2016;1:180–189. doi: 10.1016/j.jacbts.2016.03.001
2. Pitt B, Pfeffer MA, Assmann SF, Boineau R, Anand IS, Claggett B, Clausell N, Desai AS, Diaz R, Fleg JL, Gordeev I, Harty B, Heitner JF, Kenwood CT, Lewis EF, O'Meara E, Probstfield JL, Shaburishvili T, Shah SJ, Solomon SD, Sweitzer NK, Yang S, McKinlay SM; TOPCAT Investigators. Spironolactone for heart failure with preserved ejection fraction. *N Engl J Med*. 2014;370:1383–1392. doi: 10.1056/NEJMoa1313731
3. Pfeffer MA, Claggett B, Assmann SF, Boineau R, Anand IS, Clausell N, Desai AS, Diaz R, Fleg JL, Gordeev I, Heitner JF, Lewis EF, O'Meara E, Rouleau JL, Probstfield JL, Shaburishvili T, Shah SJ, Solomon SD, Sweitzer NK, McKinlay SM, Pitt B. Regional variation in patients and outcomes in the Treatment of Preserved Cardiac Function Heart Failure With an Aldosterone Antagonist (TOPCAT) trial. *Circulation*. 2015;131:34–42. doi: 10.1161/CIRCULATIONAHA.114.013255
4. de Denus S, O'Meara E, Desai AS, Claggett B, Lewis EF, Leclair G, Jutras M, Lavoie J, Solomon SD, Pitt B, Pfeffer MA, Rouleau JL. Spironolactone metabolites in TOPCAT - new insights into regional variation. *N Engl J Med*. 2017;376:1690–1692. doi: 10.1056/NEJMc1612601
5. George SL, Buyse M. Data fraud in clinical trials. *Clin Investig (Lond)*. 2015;5:161–173. doi: 10.4155/cli.14.116
6. The International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. General Principles for Planning and Design of Multi-Regional Clinical Trials [Internet]. 2017. [cited April 19, 2018]. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E17/E17EWG_Step4_2017_1116.pdf. Accessed April 19, 2018.
7. Rudolph KE, van der Laan MJ. Robust estimation of encouragement-design intervention effects transported across sites. *J R Stat Soc Series B Stat Methodol*. 2017;79:1509–1525. doi: 10.1111/rssb.12213
8. Rudolph KE, Schmidt NM, Glymour MM, Crowder R, Galin J, Ahern J, Osypuk TL. Composition or context: using transportability to understand drivers of site differences in a large-scale housing experiment. *Epidemiology*. 2018;29:199–206. doi: 10.1097/EDE.0000000000000774
9. Bareinboim E, Pearl J. Causal inference and the data-fusion problem. *Proc Natl Acad Sci U S A*. 2016;113:7345–7352. doi: 10.1073/pnas.1510507113
10. Westreich D, Edwards JK, Lesko CR, Stuart E, Cole SR. Transportability of trial results using inverse odds of sampling weights. *Am J Epidemiol*. 2017;186:1010–1014. doi: 10.1093/aje/kwx164
11. ACCORD Study Group, Cushman WC, Evans GW, Byington RP, Goff DC Jr, Grimm RH Jr, Cutler JA, Simons-Morton DG, Basile JN, Corson MA, Probstfield JL, Katz L, Peterson KA, Friedewald WT, Buse JB, Bigger JT, Gerstein HC, Ismail-Beigi F. Effects of intensive blood-pressure control in type 2 diabetes mellitus. *N Engl J Med*. 2010;362:1575–1585. doi: 10.1056/NEJMoa1001286
12. Udelson JE. Heart failure with preserved ejection fraction. *Circulation*. 2011;124:e540–e543. doi: 10.1161/CIRCULATIONAHA.111.071696
13. Redfield MM. Heart failure with preserved ejection fraction. *N Engl J Med*. 2016;375:1868–1877. doi: 10.1056/NEJMcp1511175
14. Zakeri R, Cowie MR. Heart failure with preserved ejection fraction: controversies, challenges and future directions. *Heart*. 2018;104:377–384. doi: 10.1136/heartjnl-2016-310790
15. Luedtke AR, Carone M, van der Laan MJ. An omnibus nonparametric test of equality in distribution for unknown functions. ArXiv.org. 2015. http://arxiv.org/abs/1510.04195. Accessed May 1, 2018.
16. Hosmer DW, Lemeshow S. Assessing the fit of the model. In: *Applied Logistic Regression*. Hoboken, NJ: John Wiley & Sons, Inc; 2005 [cited April 19, 2018]:143–202. Available from: http://doi.wiley.com/10.1002/0471722146.ch5.
17. Nattino G, Finazzi S, Bertolini G. A new calibration test and a reappraisal of the calibration belt for the assessment of prediction models based on dichotomous outcomes. *Stat Med*. 2014;33:2390–2407. doi: 10.1002/sim.6100
18. Finazzi S, Poole D, Luciani D, Cogo PE, Bertolini G. Calibration belt for quality-of-care assessment based on dichotomous outcomes. *PLoS One*. 2011;6:e16110. doi: 10.1371/journal.pone.0016110
19. Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *Am J Epidemiol*. 2014;179:764–774. doi: 10.1093/aje/kwt312
20. Stekhoven DJ, Bühlmann P. MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012;28:112–118. doi: 10.1093/bioinformatics/btr597
21. Berkowitz SA. Statistical Code. https://saberkowitz.web.unc.edu/statistical-code/tmle-transport-code/. Accessed January 21, 2019.