Comparison of Methods to Generalize Randomized Clinical Trial Results Without Individual-Level Data for the Target Population

Jin-Liern Hong*, Michael Webster-Clark, Michele Jonsson Funk, Til Stürmer*, Sara E. Dempster, Stephen R. Cole, Iksha Herr, and Robert LoCasale

* Correspondence to Dr. Jin-Liern Hong, Department of Epidemiology, UNC Gillings School of Global Public Health, University of North Carolina at Chapel Hill, 135 Dauer Drive, 2101 McGavran-Greenberg Hall, CB #7435, Chapel Hill, NC 27599-7435 (e-mail: jlhongtw@email.unc.edu); or Dr. Til Stürmer, Department of Epidemiology, UNC Gillings School of Global Public Health, University of North Carolina at Chapel Hill, 135 Dauer Drive, 2101 McGavran-Greenberg Hall, CB #7435, Chapel Hill, NC 27599-7435 (e-mail: til.sturmer@post.harvard.edu).

Initially submitted January 4, 2018; accepted for publication October 5, 2018.

Our study explored the application of methods to generalize randomized controlled trial results to a target population without individual-level data. We compared 4 methods using aggregate data for the target population to generalize results from the international trial, Justification for the Use of Statins in Prevention: an Intervention Trial Evaluating Rosuvastatin (JUPITER), to a target population of trial-eligible patients in the UK Clinical Practice Research Datalink (CPRD). The gold-standard method used individual data from both the trial and CPRD to predict probabilities of being sampled in the trial and to reweight trial participants to reflect CPRD patient characteristics. Methods 1 and 2 used weighting methods based on simulated individual data or the method of moments, respectively. Method 3 weighted the trial's subgroup-specific treatment effects to match the distribution of an effect modifier in CPRD. Method 4 calculated the expected absolute benefits in CPRD assuming homogeneous relative treatment effect. Methods based on aggregate data for the target population generally yielded results between the trial and gold-standard estimates. Methods 1 and 2 yielded estimates closest to the gold-standard estimates when continuous effect modifiers were represented as categorical variables. Although individual data or data on joint distributions remains the best approach to generalize trial results, these methods using aggregate data might be useful tools for timely assessment of randomized trial generalizability.

cardiovascular diseases; external validity; generalizability; JUPITER trial; randomized clinical trial; statins

Abbreviations: CI, confidence interval; CPRD, Clinical Practice Research Datalink; hsCRP, high-sensitivity C-reactive protein; JUPITER, Justification for the Use of Statins in Prevention: an Intervention Trial Evaluating Rosuvastatin; RCT, randomized clinical trial.

There is growing interest in improving the generalizability of randomized controlled trial (RCT) results to real-world populations. RCTs provide protection against confounding; however, participants are often highly selected, leading to differences in some important characteristics between the RCT participants and the target population for whom the treatment would be indicated in the real world. If treatment effects are heterogeneous across these characteristics, the generalizability of the RCT result to the target population is questionable.

To address this sampling issue and therefore improve generalizability of RCT results, weighting methods have been proposed (1–3). Related to propensity score methods, weighting methods predict probabilities of sampling from the target population into the RCT using, for example, multivariable logistic models, and then standardize the RCT participants to be representative of a target population of interest. This approach requires individual-level data for both the RCT and the target population, however. Researchers might not have access to individual data for the target population due to cost, time, or data protection. In contrast, aggregate data for a potential target population can usually be easily retrieved from the literature or publicly available data. Thus, approaches for generalizing RCT results based

on aggregate data from the target population could be useful tools for forecasting RCT generalizability in a timely manner.

The objective of this study was to compare several methods of generalizing RCT results to a target population when only aggregate data is available for the target population. We applied these methods in the context of estimating the anticipated rosuvastatin effect for primary cardiovascular prevention in the Clinical Practice Research Datalink (CPRD) using the results from Justification for the Use of Statins in Primary Prevention: an Intervention Trial Evaluating Rosuvastatin (JUPITER). With access to individual data for both the trial and the target population, we were able to compare results using each aggregate-data method to the gold-standard method of using individual-level data.

METHODS

Study population

The JUPITER trial (ClinicalTrials.gov identifier: NCT-00239681) was a multicenter, randomized, double-blind trial. This trial was conducted in 26 countries and randomized 17,802 subjects to receive rosuvastatin 20 mg or placebo daily to prevent cardiovascular disease (4). Eligible participants were men aged ≥50 years or women aged ≥60 years who had elevated high-sensitivity C-reactive protein (hsCRP) levels but not hyperlipidemia (i.e., hsCRP of ≥2.0 mg/L, low-density lipoprotein cholesterol of <130 mg/dL, and triglycerides of <500 mg/dL). Subjects were excluded if they had a history of cardiovascular disease, diabetes, or cancer. The primary endpoint was the occurrence of a first major cardiovascular event, defined as nonfatal myocardial infarction, nonfatal stroke, hospitalization for unstable angina, arterial revascularization procedure, or confirmed death from cardiovascular causes.

The target population of interest was the population of England who would have been eligible to participate in the JUPITER trial (Web Figure 1, available at https://academic.oup.com/aje). As a proxy for this population, we used data from the CPRD linked with Hospital Episode Statistics data, and we selected all men aged ≥50 years and women aged ≥60 years with ≥2 years of registration after the practice Up-to-Standard date from January 1, 2001, through April 30, 2014. For each patient, we defined an eligibility period, starting on the date on which the patient met both age and registration-time requirements and ending with the earliest event of the following: diagnosis of cardiovascular disease, diabetes, or cancer; initiation of any lipid-lowering agents; death; migration out of general practice; or end of study. Next, 1 visit to the general practitioner during the eligibility period was randomly selected for each patient and was defined as the index date. Patients were excluded if they had cardiovascular disease, diabetes, or cancer or used any lipid-lowering agents before meeting age and registration-time requirements.

Among all eligible CPRD patients with an index date, we selected only those who had complete data for all relevant measurements and laboratory tests (2.4% of CPRD patients with an index date), including body mass index, smoking status, hsCRP, low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, and serum creatinine. To be consistent with the JUPITER trial, the final target population included those who had a low-density lipoprotein cholesterol level of <130 mg/dL and elevated levels of hsCRP (≥2.0 mg/L).

Effect modifiers

Concerns about RCT generalizability arise when there remain overt differences in important characteristics between the trial participants and the target population, and these characteristics modify the treatment effects (1, 5). Thus, the variables selected in this study are potential effect modifiers that are plausibly associated with RCT participation, not factors solely associated with RCT participation or the outcome.

We chose 10 effect modifiers on the basis of having previously been shown to modify the effect of rosuvastatin on the cardiovascular risk (6–8) and to be plausibly associated with selection into the trial. Sex, current smoking, chronic kidney disease, and use of aspirin and antihypertensive drugs were defined as binary variables. Age, body mass index, hsCRP, high-density lipoprotein cholesterol, and low-density lipoprotein cholesterol values were considered as continuous variables as well as categorical variables, depending on the generalizing method. To simulate CPRD aggregate data similar to those that would typically be publicly available, we summarized the data as means with standard deviations for continuous variables and proportions for binary and categorical variables.

Overall and subgroup-specific treatment effects in the JUPITER trial

In the JUPITER trial, the rosuvastatin effect on the cardiovascular risk was estimated based on the intention-to-treat approach. We used Cox proportional hazard models to estimate the hazard ratio and 95% confidence interval for the comparison of major cardiovascular events between rosuvastatin and placebo groups. We also calculated risk differences and risk ratios at 1, 2, 3, and 4 years after randomization using a nonparametric method accounting for noncardiovascular competing causes of death and obtained 95% confidence intervals based on standard deviations of estimates from 200 bootstraps (9). We also repeated these analyses in the various subgroups stratified by effect modifiers of interest to obtain subgroup-specific treatment effects.

Cardiovascular risk in the CPRD

The occurrence of cardiovascular disease was defined as hospitalization with the primary discharge diagnosis being for cardiovascular disease or with procedures of arterial revascularization, using data from the linked Hospital Episode Statistics data. Eligible CPRD patients were followed from the index date to the first cardiovascular event. Follow-up ended at the earliest of statin initiation, death, migration out of general practice, or end of study. We used a nonparametric method to calculate the cumulative incidences of cardiovascular disease but did not account for noncardiovascular death as a competing event when estimating cumulative incidences because of lack of data on causes of death.

Methods for generalizing RCT results to a target population

We implemented 4 methods to generalize the results of the JUPITER trial to the target population based on aggregate data of the target population and compared these with the gold-standard method based on individual data of the target population (Table 1). Methods 1–2 are based on reweighting the trial population after predicting the probability of being sampled in the trial. Method 3 uses subgroup-specific estimates of treatment effect in JUPITER and aggregate data from the target population. Method 4 uses the overall relative treatment effect from the trial and absolute outcome risk estimates from the target population. Because of the properties of these methods, the gold-standard method and methods 1–3 can be used to estimate the expected relative and absolute effects of rosuvastatin on cardiovascular risk in the target population, while method 4 can be used only to estimate the absolute effect in the target. Because the relative effect is assumed to be constant, it will generalize directly from the trial to the target population.

Gold-standard method: weighting using individual data. This method used individual data from both JUPITER and CPRD and is considered the gold-standard for comparison. We combined individual data from the JUPITER trial and the target population to estimate the probability of being sampled in the JUPITER trial ($P_{S=1}$), using multivariable logistic regression models. The logistic models included all effect modifiers of interest, higher-order terms for continuous covariates, and product terms for the joint distribution. We first included age and laboratory test values as continuous variables in the model and repeated modeling after grouping them into categorical variables. After estimating the sampling probability, the sampling weights for the JUPITER trial participants were calculated as the inverse odds of the sampling probability $[(1 - P_{S=1}) \div P_{S=1}]$ (10) and scaled by multiplying with the marginal odds of being in the trial so that the pseudo-RCT population had a similar sample size (sum of weights) as the original RCT. Within the weighted JUPITER trial, we used Cox proportional hazard models with robust variance to calculate hazard ratios and 95% confidence intervals of the rosuvastatin effect. We also used a nonparametric method to estimate risk differences accounting for competing causes of death other than cardiovascular disease and obtained 95% confidence intervals based on standard deviations of estimates from 200 bootstraps (9).

Method 1: weighting using simulated individual data. This approach began with individual data from the JUPITER trial and aggregate data from the target population. Given the summary statistics of the effect modifiers and the total number of the target population ($n = 6{,}619$), we simulated hypothetical individual data for the target population under a strong assumption of no correlation between the effect modifiers. Details about the simulation setup are provided in Web Appendix 1. All 10 effect modifiers were simulated independently for a sample size of 661,900 to stabilize the sampling distribution. Continuous variables were simulated based on truncated normal distributions to cope with the inclusion/exclusion criteria of the JUPITER trial (e.g., the lower limit for age). With real individual data for the trial and simulated individual data for the target population, we used logistic models to estimate the sampling probability. We included only main effect terms in the logistic models because adding interaction terms can worsen covariate balance by making the variables appear more uncorrelated in the weighted trial than they were in the true target population (see Web Table 1). After estimating the individual sampling probabilities, we calculated sampling weights and estimated the treatment effect within the weighted JUPITER participants as in the gold-standard method.

Method 2: weighting using the method of moments. Signorovitch et al. (11, 12) proposed this method for indirect comparison between trials with limited availability of individual data. Patients in the trial with individual data were reweighted to have average values of the covariates that match those reported in the target population with aggregate data only. Given that the target population lacks individual data, the coefficients in the logistic models were estimated by the method of moments. The statistical details on using the method of moments to estimate the weights and sample R code have been described previously (12, 13). After obtaining the weights, we estimated the treatment effect within the weighted JUPITER participants as in the gold-standard method.

Method 3: poststratification. This approach applies only to the scenario with binary or categorical effect modifiers and can be used for 1 effect modifier at a time. We used the subgroup-specific treatment effect estimates in the trial and proportions of the categorical effect modifiers in the target population. We calculated the treatment effect for the target population by weighting the average of the stratum-specific treatment effects according to proportions of a given effect modifier in the target population. The 95% confidence interval was obtained based on the pooled standard deviations across strata. We repeated the calculation for each effect modifier separately, and then summarized the poststratification estimates of treatment effect among all effect modifiers by taking the (unweighted) average of the estimated treatment effects. Calculations of relative effect estimates were performed on the natural logarithm scale.

Method 4: expected absolute risk reduction. This method has been described previously in detail (14–16). We first obtained the cardiovascular risks in the JUPITER-eligible CPRD patients who were naive to statins. We treated these cardiovascular risks as the baseline risks in unexposed target-population patients. Next, assuming a uniform relative effect of rosuvastatin, we multiplied the cardiovascular risks in the target population by the relative risk from the trial at each time point to obtain the expected absolute risk in the target population if they were exposed to rosuvastatin. The absolute risk reduction associated with rosuvastatin in the target population was then calculated as the absolute difference between the observed risk in the unexposed patients and the expected risk in the exposed patients. For example, given a 1-year cardiovascular risk of 1.4% in the target population and a 1-year risk ratio of 0.61 observed for rosuvastatin in JUPITER, the expected risk was $0.61 \times 1.4\% = 0.85\%$ if the target population was exposed to rosuvastatin. Thus, the expected risk reduction at 1 year was $0.85\% - 1.4\% = -0.55$ percentage points. The 95% confidence intervals were obtained based on standard deviations of estimates from 200 bootstraps of the JUPITER data.

Estimation of the sampling probability using the method of moments was performed in R, version 3.3.2 (R Foundation for

Table 1. Description of Different Methods for Generalizing a Randomized Clinical Trial's Results to a Target Population Without Individual Data, Using Data From Justification for the Use of Statins in Primary Prevention: an Intervention Trial Evaluating Rosuvastatin, Multiple Countries, 2003–2008, and the Clinical Practice Research Datalink–Hospital Episode Statistics Linked Database, England, 2001–2014

Data Availability in JUPITER	Target Population Selected From the CPRD–Hospital Episode Statistics Linked Database	
	Individual Data	Aggregate Data
Individual data	Gold-standard (weighting using individual data): First, predicting the probability of being eligible for JUPITER using multivariable logistic models and reweighting the JUPITER participants to reflect the patient characteristics in the target population. Next, estimating absolute and relative treatment effects within the weighted JUPITER trial.	Method 1 (weighting using simulated individual data): Simulating hypothetical individual data based on aggregate data for target population and using gold-standard weighting method to estimate treatment effect. Method 2 (weighting using the method of moments): Using the methods of moments to estimate the weights and then estimating treatment effect within the weighted JUPITER.
Aggregate data	None	Method 3 (poststratification): Computing weighted treatment effect estimate by reweighting subgroup-specific treatment effects in JUPITER based on the distribution of a given effect modifier in the target population. Method 4 (expected absolute risk reduction): Multiplying the observed outcome risk in the target population who were unexposed to statins by the relative treatment effect in JUPITER to obtain the expected risk in the target population if they were exposed to rosuvastatin. Next, calculating the expected absolute treatment effect in target population by subtracting the risk in the target population who were unexposed to statins from the expected risk if the target population were exposed to statins.

Abbreviations: CPRD, Clinical Practice Research Datalink; JUPITER, Justification for the Use of Statins in Primary Prevention: an Intervention Trial Evaluating Rosuvastatin.

Statistical Computing, Vienna, Austria) (13); all other statistical analyses were performed with SAS, version 9.3 (SAS Institute, Inc., Cary, North Carolina). This study was approved by the institutional review board of the University of North Carolina at Chapel Hill, and by the Independent Scientific Advisory Committee for Medicines and Healthcare Products Regulatory Agency database research in the United Kingdom.

RESULTS

During the study period, we identified a total of 6,619 patients in the CPRD who would have been eligible for the JUPITER trial. Compared with the JUPITER participants, the CPRD patients were more likely to be female, younger, less obese, and less likely to be using aspirin but with higher hsCRP and low-density lipoprotein cholesterol levels (Tables 2 and 3). Compared with the participants in JUPITER who were randomized to placebo, the CPRD patients had slightly higher cardiovascular risks at years 1 and 2 but lower risks at years 3 and 4 (Table 4).

We show the distribution of effect modifiers after weighting by different methods in Tables 2 and 3, depending on the form of age and laboratory values considered in the analyses (i.e., continuous or categorical variables). Both the gold-standard method (weighting using individual data) and method 2 (weighting using the method of moments) reweighted the JUPITER trial participants to resemble the CPRD population with respect to effect modifiers marginally. However, method 1 (weighting using simulated individual data) had overt imbalance between the CPRD and the weighted JUPITER trial population.

Figure 1 compares the hazard ratios of the rosuvastatin effect on cardiovascular prevention after implementing different methods of generalizing RCT results. Compared with the treatment effect observed in JUPITER (hazard ratio = 0.56, 95% confidence interval (CI): 0.46, 0.69), the expected treatment effects in the target population were attenuated with the gold-standard method and were similar between the scenarios in which we included age and laboratory values as continuous (hazard ratio = 0.65, 95% CI: 0.46, 0.91) or as categorical variables (hazard ratio = 0.66, 95% CI: 0.48, 0.91). Similarly, methods 1 and 2 also yielded attenuated treatment effects when age and laboratory values were included as categorical variables, with hazard ratios of 0.63 (95% CI: 0.47, 0.86) and 0.62 (95% CI: 0.46, 0.83), respectively. In contrast, when age and laboratory tests were considered as continuous variables, the hazard ratio estimates were close to the estimates observed in the JUPITER trial. In method 3 (poststratification), the estimated hazard ratios varied according to the effect modifiers, ranging from 0.52 to 0.58, and the average hazard ratio was 0.55, which was also close to the JUPITER trial estimates.

Figures 2 and 3 show the generalized results of the absolute benefits of rosuvastatin. Based on the gold-standard method, the absolute benefit of rosuvastatin was muted in the first 2 years of follow-up and began to emerge afterwards. Method 1

Table 2. Distribution of Continuous or Binary Effect Modifiers Before and After Weighting by Different Methods in Justification for the Use of Statins in Primary Prevention: an Intervention Trial Evaluating Rosuvastatin, Multiple Countries, 2003–2008, and the Clinical Practice Research Datalink–Hospital Episode Statistics Linked Database, England, 2001–2014

Effect Modifier	CPRD (n = 6,619)		JUPITER (n = 17,802)			JUPITER Data Weighted by Different Methods[a]								
						Gold-Standard Method			Method 1			Method 2		
	%	Mean (SD)	%	Mean (SD)	ASMD[b]	%	Mean (SD)	ASMD[b]	%	Mean (SD)	ASMD[b]	%	Mean (SD)	ASMD[b]
Age, years		65 (9.5)		66 (7.7)	0.19		65 (9.4)	0.01		67 (7.8)	0.10		65 (9.5)	0.00
Male sex	53.9		61.8		0.16	53.8		0.00	50.6		0.07	53.9		0.00
BMI[c]		27.9 (5.8)		29 (5.5)	0.22		28 (5.8)	0.02		27.6 (5.5)	0.09		27.9 (5.8)	0.00
Current smoker	17.2		15.8		0.04	17.2		0.00	19.5		0.06	17.2		0.01
Antihypertensives	51.6		49.7		0.04	52.2		0.01	48.4		0.06	51.6		0.00
Aspirin	9.2		18.6		0.27	9.4		0.01	8.8		0.02	9.2		0.00
CKD	19.2		18.3		0.02	18.7		0.01	19.3		0.00	19.2		0.00
hsCRP, mg/L		5.7 (3.9)		5.3 (3.6)	0.08		5.8 (4.0)	0.02		6.2 (4.4)	0.07		5.7 (3.9)	0.01
LDL-C, mg/dL		106 (18.3)		104 (18.7)	0.10		106 (18.1)	0.01		100 (21.3)	0.03		106 (18.3)	0.00
HDL-C, mg/dL		58 (19.0)		51 (15.3)	0.37		58 (19.0)	0.00		59 (20.0)	0.04		58 (19.0)	0.00

Abbreviations: ASMD, absolute standardized mean differences; BMI, body mass index; CKD, chronic kidney disease; CPRD, Clinical Practice Research Datalink; HDL-C, high-density lipoprotein cholesterol; hsCRP, high-sensitivity C-reactive protein; JUPITER, Justification for the Use of Statins in Primary Prevention; LDL-C, low-density lipoprotein cholesterol; SD, standard deviation.

[a] The gold-standard method used individual data from both JUPITER and CPRD to estimate predicted probabilities of being sampled in JUPITER and reweight the JUPITER population to reflect CPRD patient characteristics. The distributions of sampling probabilities and weights based on the gold-standard method are presented in Web Figure 3 and Web Table 4. The mean (SD) of sampling weights based on the gold-standard method was 1.00 (0.99). Methods 1 and 2 used weighting methods based on simulated individual data (method 1) or the method of moments (method 2) to estimate the sampling weights.

[b] ASMD in baseline characteristics was calculated between the CPRD and the unweighted or weighted JUPITER trial, using SAS macro stddiff% (SAS Institute, Inc., Cary, North Carolina) (23).

[c] Weight (kg)/height $(m)^2$.

Table 3. Distribution of Binary or Categorical Effect Modifiers Before and After Weighting by Different Methods in Justification for the Use of Statins in Primary Prevention: an Intervention Trial Evaluating Rosuvastatin, Multiple Countries, 2003–2008, and the Clinical Practice Research Datalink–Hospital Episode Statistics Linked Database, England, 2001–2014

Effect Modifiers	CPRD (n = 6,619)		JUPITER (n = 17,802)			JUPITER Data Weighted by Different Methods[a]					
						Gold-Standard Method		Method 1		Method 2	
	No.	%	No.	%	ASMD[b]	%	ASMD[b]	%	ASMD[b]	%	ASMD[b]
Age, years											
<65	3,637	54.9	7,565	42.5	0.34	54.9	0.01	50.2	0.10	54.9	0.00
65–69	1,125	17.0	4,635	26.0		17.0		18.6		17.0	
70–74	765	11.6	3,039	17.1		11.5		12.4		11.6	
75–79	546	8.2	1,725	9.7		8.2		9.3		8.2	
≥80	546	8.2	838	4.7		8.4		9.5		8.2	
Male sex	3,565	53.9	11,001	61.8	0.16	53.5	0.01	55.0	0.02	53.9	0.00
BMI[c]											
<25	2,172	32.8	4,009	22.6	0.23	32.8	0.00	36.1	0.07	32.8	0.00
25–29	2,369	35.8	7,010	39.5		35.7		33.7		35.8	
≥30	2,078	31.4	6,721	37.9		31.5		30.2		31.4	
Current smoker	1,138	17.2	2,821	15.8	0.04	17.2	0.00	20.0	0.07	17.2	0.00
Antihypertensives	3,416	51.6	8,846	49.7	0.04	51.6	0.00	47.4	0.09	51.6	0.00
Aspirin	612	9.2	3,313	18.6	0.27	9.4	0.01	8.7	0.01	9.2	0.00
CKD	1,271	19.2	3,257	18.3	0.02	19.4	0.00	19.1	0.02	19.2	0.00
hsCRP, mg/L											
<3.5	2,402	36.3	6,750	39.9	0.31	36.4	0.00	37.2	0.03	36.3	0.00
3.5–4.9	818	12.4	3,703	21.9		12.3		12.2		12.4	
≥5.0	3,399	51.4	6,481	38.3		51.3		50.5		51.4	
LDL-C ≥ 100 mg/dL	4,686	70.8	11,841	66.6	0.09	71.1	0.01	69.4	0.06	70.8	0.00
HDL-C < 60 mg/dL	4,069	61.5	13,333	74.9	0.29	61.1	0.01	58.8	0.00	61.5	0.00

Abbreviations: ASMD, absolute standardized mean differences; BMI, body mass index; CKD, chronic kidney disease; CPRD, Clinical Practice Research Datalink; HDL-C, high-density lipoprotein cholesterol; hsCRP, high-sensitivity C-reactive protein; JUPITER, Justification for the Use of Statins in Primary Prevention: an Intervention Trial Evaluating Rosuvastatin; LDL-C, low-density lipoprotein cholesterol.

[a] The gold-standard method used individual data from both JUPITER and CPRD to estimate predicted probabilities of being sampled in JUPITER and reweight the JUPITER population to reflect CPRD patient characteristics. The distributions of sampling probabilities and weights based on the gold-standard method are presented in Web Figure 3 and Web Table 4. The mean (standard deviation) of sampling weights based on the gold-standard method was 0.99 (1.27). Methods 1 and 2 used weighting methods based on simulated individual data (method 1) or the method of moments (method 2) to estimate the sampling weights.

[b] ASMD in baseline characteristics was calculated between the CPRD and the unweighted or weighted JUPITER trial, using SAS macro stddiff % (SAS Institute, Inc., Cary, North Carolina) (23).

[c] Weight (kg)/height $(m)^2$.

with categorical laboratory and age values and method 2 also showed similar trends, but not method 1 with continuous age and laboratory values or method 3 (Figure 3). Method 4 (expected absolute risk reduction) first estimated the expected risk in the target population if they were exposed to rosuvastatin, which is presented in Table 4, and the estimated risk reduction showed stronger benefits than the JUPITER trial estimates in the first 2 years and weaker benefits afterwards (Figure 2).

DISCUSSION

Our study explored the application of weighting and nonweighting methods to generalize RCT results in the absence of individual data for the target population. These 4 methods based on aggregate data for the target population were compared with the gold-standard approach, based on individual data for the target population, to estimate the anticipated treatment effect of rosuvastatin on cardiovascular risk in the English population who would have been eligible for the JUPITER trial. The gold-standard method showed that the effects of rosuvastatin on reducing cardiovascular risk were attenuated but remained after generalizing to the target population. We found that methods based on aggregate data for the target population generally yielded results somewhere between the RCT and the gold-standard estimate. Among these methods, weighting methods using simulated individual data (method 1) and the method of moments (method 2) led to the closest estimates when considering effect modifiers as binary or categorical variables.

Table 4. The Risk of Cardiovascular Disease in the Target Population Selected from the Clinical Practice Research Datalink–Hospital Episode Statistics Linked Database, England, 2001–2014, and the Placebo Group of Justification for the Use of Statins in Primary Prevention: an Intervention Trial Evaluating Rosuvastatin, Multiple Countries, 2003–2008

Year	Observed Cardiovascular Risk				Risk Ratio in JUPITER[a]		Expected Cardiovascular Risk in Exposed CPRD Patients[a]	
	CPRD Patients		JUPITER Placebo Group					
	Risk, %	95% CI	Risk, %	95% CI	Risk Ratio	95% CI	Risk, %	95% CI
1	1.44	1.12, 1.75	1.20	0.98, 1.42	0.61	0.51, 0.74	0.88	0.61, 1.15
2	2.55	2.12, 2.97	2.44	2.09, 2.78	0.58	0.50, 0.67	1.47	1.08, 1.86
3	3.92	3.30, 4.53	4.57	3.89, 5.25	0.56	0.48, 0.65	2.18	1.57, 2.79
4	5.14	4.40, 5.89	6.19	5.17, 7.20	0.47	0.41, 0.54	2.42	1.73, 3.10

Abbreviation: CI, confidence interval; CPRD, Clinical Practice Research Datalink; JUPITER, Justification for the Use of Statins in Primary Prevention: an Intervention Trial Evaluating Rosuvastatin.

[a] The 95% confidence intervals were obtained based on the standard deviation of estimates from 200 bootstraps.

Three key assumptions need to be considered when implementing the weighting method with individual data as the gold-standard method. First, we assumed no unmeasured effect modifiers. Based on previous literature, we have captured most of the potential effect modifiers. Second, we assumed that we correctly specified the logistic models used to predict the sampling probability. Because the true models are unknown, we have included all possible 2-way interactions and have assessed the

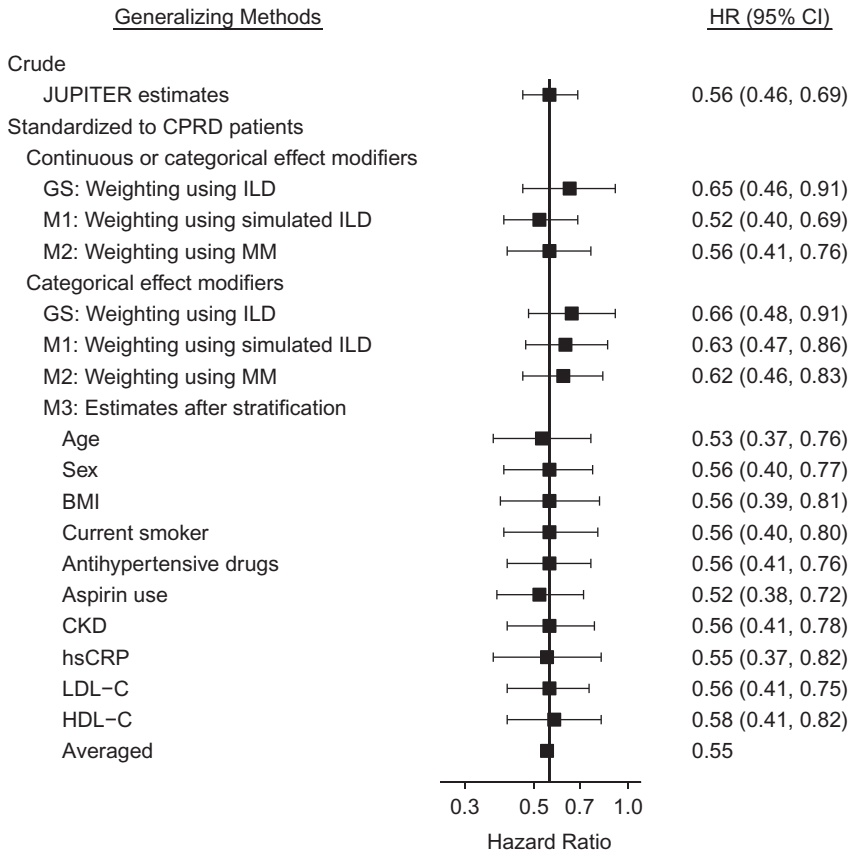

Figure 1. Hazard ratio (HR) and 95% confidence interval (CI) of the rosuvastatin effect in the primary cardiovascular prevention in Justification for the Use of Statins in Primary Prevention: an Intervention Trial Evaluating Rosuvastatin (JUPITER), multiple countries, 2003–2008, and the expected effect in a target population (Clinical Practice Research Datalink (CPRD)–Hospital Episode Statistics Linked Database, England, 2001–2014) using different methods of generalizing trial results. BMI, body mass index; CKD, chronic kidney disease; GS, gold standard; HDL-C, high-density lipoprotein cholesterol; hsCRP, high-sensitivity C-reactive protein; ILD, individual-level data; LDL-C, low-density lipoprotein cholesterol; M1, method 1; M2, method 2; M3, method 3; MM, method of moments.

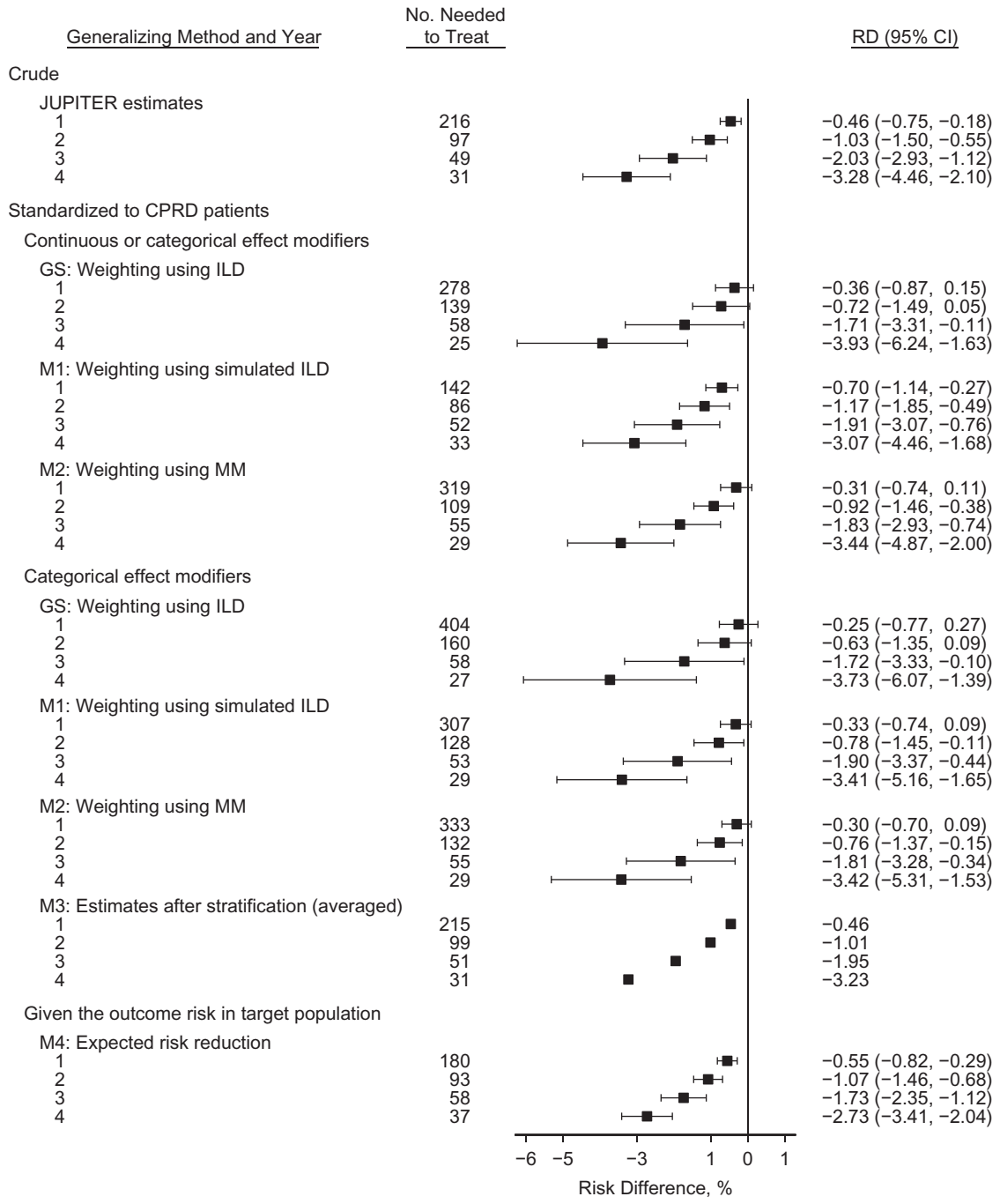

Figure 2. Risk difference (RD, %) and 95% confidence interval (CI) of the rosuvastatin effect in the primary cardiovascular prevention in Justification for the Use of Statins in Primary Prevention: an Intervention Trial Evaluating Rosuvastatin (JUPITER), multiple countries, 2003–2008, and the expected effect in a target population (Clinical Practice Research Datalink (CPRD)–Hospital Episode Statistics Linked Database, England, 2001–2014) using different methods of generalizing trial results. GS, the gold-standard; ILD, individual-level data; M1, method 1; M2, method 2; M3, method 3; M4, method 4; MM, method of moments.

models by examining marginal balance of covariates. The third assumption was that no patient in the target population had zero probability of being sampled into the RCT (i.e., positivity assumption). We acknowledge the possibility of violating this assumption. We have therefore followed the JUPITER trial's stringent inclusion and exclusion criteria to select a target population who would have been eligible for the JUPITER trial.

Weighting methods using simulated individual data (method 1) and the method of moments (method 2) showed discrepant results between the scenarios when we considered age and

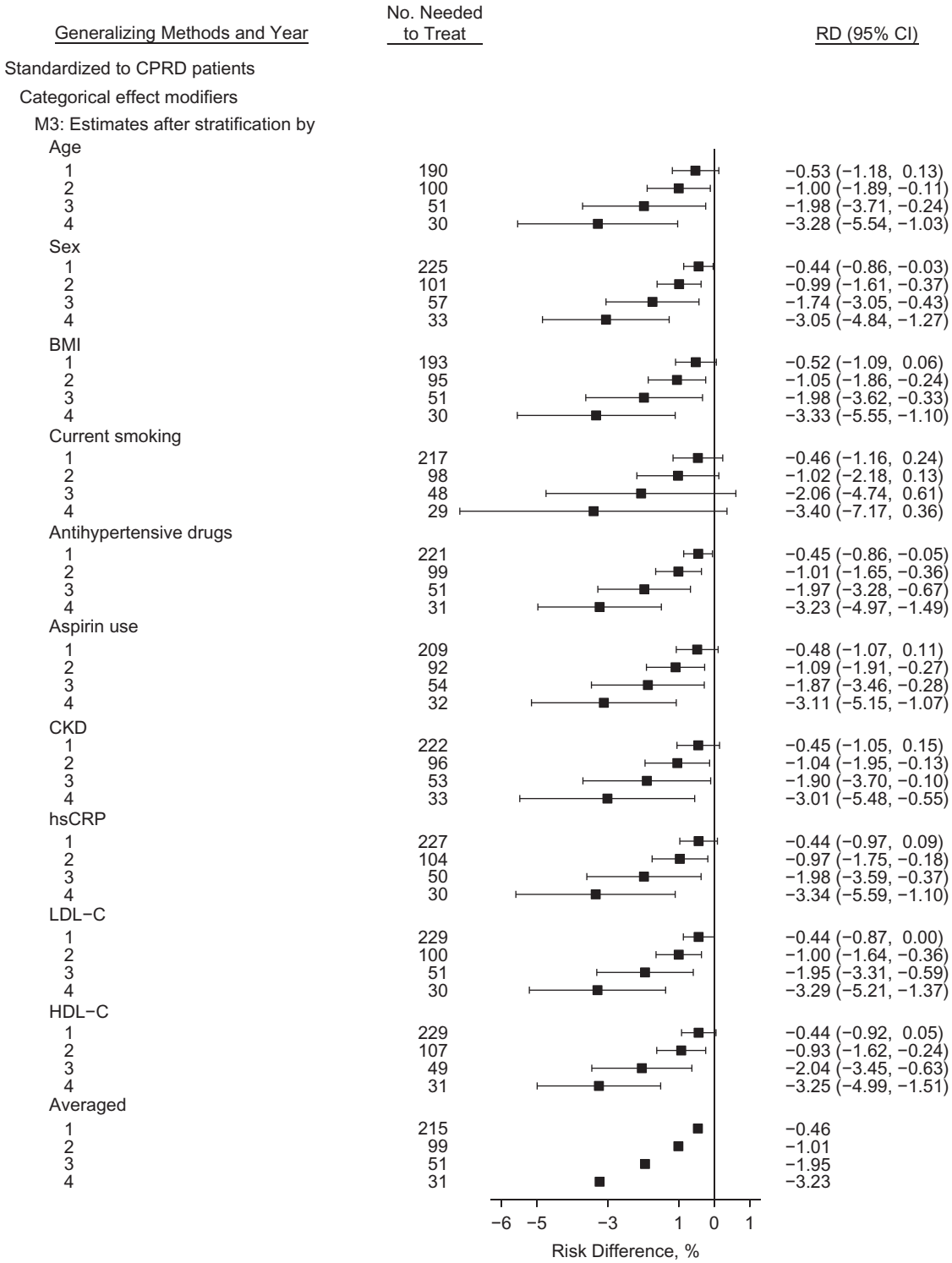

Figure 3. Risk difference (RD, %) and 95% confidence interval (CI) of the rosuvastatin effect in the primary cardiovascular prevention in Justification for the Use of Statins in Primary Prevention: an Intervention Trial Evaluating Rosuvastatin (JUPITER), multiple countries, 2003–2008, and the expected effect in a target population (Clinical Practice Research Datalink (CPRD)–Hospital Episode Statistics Linked Database, England, 2001–2014) using method 3 (M3: poststratification) according to effect modifiers. BMI, body mass index; CKD, chronic kidney disease; HDL-C, high-density lipoprotein cholesterol; hsCRP, high-sensitivity C-reactive protein; LDL-C, low-density lipoprotein cholesterol.

laboratory test results as continuous versus categorical variables in the analysis. We are not aware of literature exploring the impact of variable type on generalizing trial results, but we speculate that these differences are likely due to failure to replicate true variable distributions. Mean and standard deviation are commonly used to describe the distribution of continuous variables, but they describe the distribution well only for a normally distributed variable. Thus, for not-normally distributed variables, we are unable to reweight the trial participants to match the actual distributions in the target population in the absence of individual data, despite achieving the same mean and standard deviation between the weighted trial and target population. To investigate this further, we examined the distribution of age before and after weighting in our study and present the histograms in Web Figure 2. In the target population, age is not normally distributed, with 2 peaks at 50 and 60 years because of different age-inclusion criteria between the sexes in JUPITER. The gold-standard method based on individual data reweighted the JUPITER participants to have the same distribution of age as the target population; however, methods 1 and 2 failed to do so. In contrast, the discrete distribution for a binary or categorical variable can be successfully matched between the weighted trial and the target population without individual data. Categorization of continuous variables might be an issue when there remains residual heterogeneity within the category (17). We categorized age in 5-year intervals and categorized laboratory test results based on a priori clinical input to minimize residual heterogeneity within the category.

Weighting methods based on aggregate data are inevitably limited in the ability to match multidimensional distributions of effect modifiers between the weighted trial and target population due to lack of individual data or data on joint distributions. Although weighting methods involving parametric modeling require the assumption of correct model specification (18–20), covariate balance in joint distributions is rarely assessed in confounding control. However, covariate balance in joint distributions is important because the goal here is to reweight RCT participants to the target population on all effect modifiers, including those specific to a certain covariate pattern. We examined the variable balance in subgroups stratified by sex after reweighting the RCT participants to have similar marginal distributions of variables as the target population (Web Tables 2 and 3). We found that covariates remained balanced after stratifying by sex in the gold-standard method but worsened dramatically in methods 1 and 2. Individual data or data on joint distributions of relevant effect modifiers are needed to overcome this limitation.

Under the assumption of no correlation between effect modifiers, we simulated individual data in method 1. Little is known about the impact of correlation on generalizing RCT results. Although correlation between variables might not affect certain methods of confounding control (21), correlation might create interaction in the logistic regression models of predicting sampling probabilities, affecting the estimates in generalizability of RCT results. Thus, future study is warranted to incorporate correlation between covariates in simulations. Additionally, given that hsCRP is not a common clinical test in general practice, selecting the CPRD patients with complete data as the target population might have led to biased estimates of the generalized treatment effect (22).

Poststratification (method 3) can be considered another form of weighting to assess the generalizability of RCT results by reweighting the subgroup-specific treatment effects to match the distribution of those subgroups in the target population. The major limitation for this method is that it works only for categorical variables. In addition, while lacking individual data or data on joint distributions, this method standardizes only for the distribution of one effect modifier at a time. We then used the average of the estimates poststratified by effect modifiers as the overall estimate. Although we found that the estimates from poststratification varied considerably according to effect modifiers, the average was very close to the JUPITER original estimate. Our results indicate that the range of poststratified estimates across effect modifiers, rather than the average, might be more informative to understand the generalizability of RCT results when lacking individual data on the target population. This method might be particularly useful when there is a strong effect modifier or as a way to identify effect modifiers that could substantially influence generalizability of RCT results.

Of greater clinical interest, a simple method has been proposed and advocated to estimate the expected absolute treatment effect as a function of baseline risk in the target population (method 4) (14, 15). This method assumes a homogeneous treatment effect on the relative scale within and across the populations. Although it conflicts with the fundamental hypothesis of heterogeneity in our study, this method is easy to implement and to interpret but requires data on the absolute risk for the outcome of interest in the target population. It could be a quick tool for clinicians in combination with, for example, a Framingham risk score, but it might not be the optimal approach to assess the generalizability of randomized trial results. Our study using the JUPITER trial as an example provides evidence on heterogeneous treatment effect on both the relative and absolute scale. In addition, ideally, this method requires the same definition of the outcome in the trial and target population, which might be improbable when relying on previous publications to get baseline risks in the target population.

In theory, all these generalizing methods are easily implemented given wide availability of aggregate data on potential target populations. In practice, however, it will often be difficult to obtain aggregate data from the target population that match the trial's inclusion and exclusion criteria exactly, thus violating the positivity assumption. Based on our results, weighting methods using simulated individual data (method 1) and the method of moments (method 2) are preferred but require more complex programming techniques. Although the performance of the generalizing methods based on aggregate data depends on the achieved covariate balance and the relationship between these effect modifiers and treatment effects, our study shows that we should avoid using continuous variables when implementing weighting methods based on aggregate data because of difficulties in understanding actual distributions without individual data. In some circumstances, where a strong effect modifier is suspected or complex programming is not possible, poststratification (method 3) and expected absolute risk reduction (method 4) should be considered.

In conclusion, our study demonstrates the possibility of generalizing trial results to target populations even in the absence of individual data on the target population using a single case study. These methods, using aggregate data about the target

population, could be useful tools for timely assessment of RCT generalizability, including to individual patients in clinical care, although use of individual data or (at a minimum) data on joint distributions remains the best approach to generalize the RCT results to target populations.

ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina (Jin-Liern Hong, Michael Webster-Clark, Michele Jonsson Funk, Til Stürmer, Stephen R. Cole); R&D Information, AstraZeneca, Waltham, Massachusetts (Sara E. Dempster); and Medical Evidence and Observational Research, AstraZeneca, Gaithersburg, Maryland (Iksha Herr, Robert LoCasale).

This work was supported by AstraZeneca. T.S. is supported by the National Institutes of Health (grants R01/R56 AG023178, R01 AG056479, R01 CA174453, R01 HL118255, R01 MD011680, UL1 TR001111, and R21-HD080214). M.J.F. is supported in part by the National Institutes of Health (grants R01 HL118255, R01/R56 AG023178, UL1 TR001111, and R01 AG056479) and Health Resources and Services Administration (grant R40 MC29455-01-00). S.R.C. is supported in part by the National Institutes of Health (grants R01 AI100654, R24 AI067039, U01 AI103390, P30 AI050410, and DP2 HD084070).

Portions of this work were presented at the International Society for Pharmacoepidemiology 33rd Annual Meeting, August 26–30, 2017, Montreal, Quebec, Canada.

Conflict of interest: J.-L.H. is currently an employee of Takeda Pharmaceuticals International Co., but the study was conducted while she was a postdoctoral associate at University of North Carolina. R.L. and S.E.D. were employees of AstraZeneca at the time of the study. I.H. is currently employed by AstraZeneca. T.S. and M.J.F. do not accept personal compensation of any kind from any pharmaceutical company, although they receive salary support from the Center for Pharmacoepidemiology in the Department of Epidemiology, Gillings School of Global Public Health (current members: GlaxoSmithKline, UCB BioSciences, Merck & Co. Inc., and Shire). M.J.F. is a member of the Scientific Steering Committee for a postapproval safety study of an unrelated drug class funded by GlaxoSmithKline. All compensation for services provided on the Scientific Steering Committee is invoiced by and paid to University of North Carolina Chapel Hill. T.S. owns stock in Novartis, Roche, BASF, AstraZeneca, and Novo Nordisk.

REFERENCES

1. Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. *Am J Epidemiol*. 2010;172(1):107–115.
2. Stuart EA, Cole SR, Bradshaw CP, et al. The use of propensity scores to assess the generalizability of results from randomized trials. *J R Stat Soc Ser A Stat Soc*. 2011;174(2):369–386.
3. Kern HL, Stuart EA, Hill J, et al. Assessing methods for generalizing experimental impact estimates to target populations. *J Res Educ Eff*. 2016;9(1):103–127.
4. Ridker PM, Danielson E, Fonseca FA, et al. Rosuvastatin to prevent vascular events in men and women with elevated C-reactive protein. *N Engl J Med*. 2008;359(21):2195–2207.
5. Olsen RB, Orr LL, Bell SH, et al. External validity in policy evaluations that choose sites purposively. *J Policy Anal Manage*. 2013;32(1):107–121.
6. Glynn RJ, Koenig W, Nordestgaard BG, et al. Rosuvastatin for primary prevention in older persons with elevated C-reactive protein and low to average low-density lipoprotein cholesterol levels: exploratory analysis of a randomized trial. *Ann Intern Med*. 2010;152(8):488–496.
7. Ridker PM, MacFadyen J, Cressman M, et al. Efficacy of rosuvastatin among men and women with moderate chronic kidney disease and elevated high-sensitivity C-reactive protein: a secondary analysis from the JUPITER (Justification for the Use of Statins in Prevention: an Intervention Trial Evaluating Rosuvastatin) trial. *J Am Coll Cardiol*. 2010;55(12):1266–1273.
8. Ridker PM, MacFadyen JG, Fonseca FA, et al. Number needed to treat with rosuvastatin to prevent first cardiovascular events and death among men and women with low low-density lipoprotein cholesterol and elevated high-sensitivity C-reactive protein: Justification for the Use of Statins in Prevention: an Intervention Trial Evaluating Rosuvastatin (JUPITER). *Circ Cardiovasc Qual Outcomes*. 2009;2(6):616–623.
9. Cole SR, Lau B, Eron JJ, et al. Estimation of the standardized risk difference and ratio in a competing risks framework: application to injection drug use and progression to AIDS after initiation of antiretroviral therapy. *Am J Epidemiol*. 2015;181(4):238–245.
10. Westreich D, Edwards JK, Lesko CR, et al. Transportability of trial results using inverse odds of sampling weights. *Am J Epidemiol*. 2017;186(8):1010–1014.
11. Signorovitch JE, Sikirica V, Erder MH, et al. Matching-adjusted indirect comparisons: a new tool for timely comparative effectiveness research. *Value Health*. 2012;15(6):940–947.
12. Signorovitch JE, Wu EQ, Yu AP, et al. Comparative effectiveness without head-to-head trials: a method for matching-adjusted indirect comparisons applied to psoriasis treatment with adalimumab or etanercept. *Pharmacoeconomics*. 2010;28(10):935–945.
13. Phillippo DM, Ades AE, Dias S, et al. Methods for population-adjusted indirect comparisons in health technology appraisal. *Med Decis Making*. 2018;38(2):200–211.
14. Caro JJ, Migliaccio-Walle K. Generalizing the results of clinical trials to actual practice: the example of clopidogrel therapy for the prevention of vascular events. CAPRA (CAPRIE Actual Practice Rates Analysis) Study Group. Clopidogrel versus Aspirin in Patients at Risk of Ischaemic Events. *Am J Med*. 1999;107(6):568–572.
15. Collins R, Reith C, Emberson J, et al. Interpretation of the evidence for the efficacy and safety of statin therapy. *Lancet*. 2016;388(10059):2532–2561.
16. Spiegelman D, Khudyakov P, Wang M, et al. Evaluating public health interventions: 7. Let the subject matter choose the effect measure: ratio, difference, or something else entirely. *Am J Public Health*. 2018;108(1):73–76.

17. Altman DG, Royston P. The cost of dichotomising continuous variables. *BMJ*. 2006;332(7549):1080.
18. Drake C. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*. 1993:49(4): 1231–1236.
19. Tsiatis AA, Davidian M. Comment: demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Stat Sci*. 2007;22(4):569–573.
20. Smith JA, Todd PE. Does matching overcome LaLonde's critique of nonexperimental estimators? *J Econom*. 2005;125(1–2):305–353.
21. Pingel R, Waernbaum I. Correlation and efficiency of propensity score-based estimators for average causal effects. *Commun Stat Simul Comput*. 2017;46(5):3458–3478.
22. Hong JL, Jonsson Funk M, LoCasale R, et al. Generalizing randomized clinical trial results: implementation and challenges related to missing data in the target population. *Am J Epidemiol*. 2018;187(4):817–827.
23. Yang D, Dalton JE. A Unified Approach to Measuring the Effect Size Between Two Groups Using SAS. *SAS Global Forum 2012*. 2012; paper 335. http://support.sas.com/resources/papers/proceedings12/335-2012.pdf. Accessed September 17, 2018.