Published in final edited form as:

Epidemiology. 2019 January ; 30(1): 103–111. doi:10.1097/EDE.0000000000000926.

Diagnostic assessment of assumptions for external validity: An example using data in metastatic colorectal cancer

Michael A. Webster-Clark, Hanna K. Sanoff, Til Stürmer, and Sharon Peacock Hinton
Jennifer L. Lund

Department of Epidemiology, University of North Carolina, Chapel Hill, NC (Michael Webster-Clark, Til Stürmer, Sharon Peacock Hinton Jennifer L. Lund)

Department of Medicine, University of North Carolina, Chapel Hill, NC (Hanna Sanoff)

Abstract

Background: Methods developed to estimate intervention effects in external target populations assume that all important effect measure modifiers have been identified and appropriately modeled. Propensity score-based diagnostics can be used to assess the plausibility of these assumptions for weighting methods.

Methods: We demonstrate the use of these diagnostics when assessing the transportability of treatment effects from the standard of care for metastatic colorectal cancer control arm in a phase III trial (HORIZON III) to a target population of 1,942 Medicare beneficiaries aged 65+ years.

Results: In an unadjusted comparison, control arm participants had lower mortality compared with target population patients treated with the standard of care therapy [trial vs. target hazard ratio (HR)= 0.72, 95% confidence interval (CI): 0.58, 0.89]. Applying inverse odds of sampling weights attenuated the trial vs. target HR (weighted HR=0.96, 95% CI: 0.73, 1.26). However, whether unadjusted or weighted, hazards did not appear proportional. At 6 months of follow-up, mortality was lower in the weighted trial population than the target population (weighted trial vs. target risk difference (RD) = -0.07, 95% CI -0.13, -0.01), but not at 12 months (weighted RD= 0.00, 95% CI: -0.09, 0.09).

Conclusion: These diagnostics suggest direct transport of treatment effects from HORIZON III to the Medicare population is not valid. However, the proposed sampling model might allow valid transport of the treatment effects on longer-term mortality from HORIZON III to the Medicare population treated in clinical practice.

Corresponding author contact information: Michael Webster-Clark, PharmD, Department of Epidemiology, UNC Gillings School of Global Public Health, University of North Carolina at Chapel Hill, McGavran-Greenberg, CB #7435, Chapel Hill, NC 27599-7435, mawc@live.unc.edu, Phone: 1 919 966 7433, Fax: 1 919 966 2089.

Disclosure of potential conflicts of interest: Dr. Webster-Clark worked as a research assistant with AstraZeneca for another generalizability project assessing another medication.
No other potential conflicts of interest were reported by the authors.

Accessing data and replication: Project Datasphere™ data is freely available to all researchers who can demonstrate a compelling interest, but SEER-Medicare data requires formal submission of a data use agreement. Specific SAS code used can be furnished upon request to the University of North Carolina at Chapel Hill.

Keywords

Generalizability; transportability; external validity; randomized controlled trial; methods; oncology

Large randomized controlled trials (RCTs) are designed to generate internally valid estimates of treatment effects.¹ However, this design does not guarantee that the estimated treatment effect will be externally valid in populations of interest because the sampled populations may not represent the target populations.^{2,3} Methods for obtaining treatment effects from an RCT (or other studied population) that are externally valid in specific target populations seek to generate realistic intervention effect estimates that can be used to inform policies targeted at the population-level.⁴⁻⁷

Propensity score-based weighting methods have been developed to estimate these externally valid treatment effects in specified target populations.^{2,8,9} In addition to assumptions of external consistency (i.e. the same intervention is applied in the target and study population) and external positivity (i.e. that there was a non-zero probability for being enrolled in the trial for every covariate pattern), weighting methods assume that all patient factors associated with trial participation that modify treatment effects are included and correctly specified in the model being used. Fortunately, it is possible to check these assumptions if treatment and outcome data in the target population under study are available as shown with control arm comparisons in the field of education policy.^{10,11}

Such comparisons could be especially useful in the field of oncology, where external validity of RCT findings is an important concern. In general, enrollment onto cancer trials is low, with less than 5% of adult cancer patients participating.^{12,13} Trial participation is a complex function of clinical and structural factors, a physician's views of a patient's fitness and equipoise, and a patient's willingness to enroll and be randomized. As a result, cancer treatment trials tend to underrepresent certain subgroups of patients, including older adults (age 65+ years), women, racial minorities, and those with lower socioeconomic status.¹⁴ If any of these variables are associated with the outcome, any non-null treatment effects estimated in a trial are unlikely to be generalizable to the clinical practice setting, either on the relative or absolute scale.^{15,16}

Historically, assessment of the external validity of cancer treatment trials and other RCT results to target populations has involved the qualitative assessment of differences in patient characteristics among those enrolled in trials and those observed in routine care settings or simply comparing the proportions of individuals with certain demographic characteristics in trials to demographics in clinical practice settings.¹⁷⁻²³ Increasingly, however, researchers are taking a more quantitative approach to evaluating external validity by comparing the estimated treatment effects from observational cancer cohorts to those in trials.²⁴ Others have directly compared outcomes from observational cohorts to those reported in trial populations undergoing the same treatment without accounting for differences in patient characteristics between the observational and trial patient populations.²⁵

Here we demonstrate a different approach to evaluate assumptions for external validity using data from a phase III trial of chemotherapy for metastatic colorectal cancer by comparing all-cause mortality in the trial control arm (representing the current standard of care) to that of similarly treated patients in routine care using propensity score weighting methods as a diagnostic tool, as introduced by Stuart et al.^{10,11}

METHODS

Data sources and study populations

We obtained RCT data from Project Data Sphere™. Project Data Sphere is an independent, not-for-profit initiative that serves as an online digital library of de-identified data from a variety of trials in cancer.²⁶ Data are cleaned and stored as SAS data sets with comprehensive data dictionaries and are freely accessible for download by interested researchers.

Project Data Sphere™ provided control arm data from a phase III trial of oxaliplatin, leucovorin, and 5-fluorouracil (FOLFOX) plus Bevacizumab or Cediranib as First-Line Treatment of Patients with Metastatic Colorectal Cancer (HORIZON III).²⁷ HORIZON III was a multicenter international trial enrolling patients from Europe, Africa, Asia, and North America. Patients eligible for the trial were 18 years old or older and were randomly assigned to receive FOLFOX with either (1) bevacizumab or (2) cediranib. Our analyses were restricted to the 257 older adults (age 65+ years) randomized to the control arm (FOLFOX+bevacizumab), as this treatment currently represents the standard of care for metastatic colorectal cancer and we were interested in older adults as the target population. While this trial did not demonstrate a treatment effect for cediranib, the fact that data from the bevacizumab arm are freely available makes it useful for demonstration of these methods.

Data for the target population were derived from a linkage of data from cancer registries included in the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) program and Medicare enrollment and claims data. SEER collects demographic and tumor data for individuals diagnosed with cancer residing within one of the 18 SEER regions and the linkage with Medicare enrollment and claims data allows for the identification of cancer treatments and follow-up for mortality.²⁸ The project was approved by the University of North Carolina at Chapel Hill's Institutional Review Board and funded by a grant from the University of North Carolina Lineberger Comprehensive Cancer Center.

The target population was defined based on modified HORIZON III trial inclusion/exclusion criteria. We identified patients aged 65 or older at diagnosis of a first, primary metastatic colorectal cancer between 2004 and 2011. All patients were required to have continuous Medicare Parts A and B coverage and no managed care enrollment 12 months before and six months after their month of diagnosis to ensure capture of important covariates and initial receipt of chemotherapy. Patients were required to initiate first-line chemotherapy with bevacizumab and oxaliplatin within six months from the cancer diagnosis (set to the first date of the month of diagnosis). Because oxaliplatin is always co-administered with 5-fluorouracil, we did not require claims for 5-fluorouracil to define FOLFOX.²⁹ eAppendix A

includes the Healthcare Common Procedure Coding System (HCPCS) codes recorded on physician and outpatient claims used to define specific chemotherapeutic agents.

The HORIZON III trial excluded participants with an Eastern Cooperative Oncology Group (ECOG)/World Health Organization (WHO) performance status greater than 2. As ECOG performance status is not collected in SEER-Medicare data, we attempted to identify a sub-cohort of the target population who would meet this criteria using two different claims-based algorithms, the Faurot frailty score³⁰ and the Davidoff disability status score.³¹ We explored various cut-points for identifying individuals likely to be frail or disabled based on predicted probabilities from these algorithms (i.e., >5%, >7.5%, >10%, >11.5%, and >15%).

We built our model estimating the probability of “sampling” or representation in the HORIZON III trial from the potential effect measure modifiers studied in the trial, as well as variables we expected to be associated with the outcome based upon literature review and expert opinion.¹²⁻¹⁴ The factors included in our model were participant age (modeled as a quadratic restricted spline with four knots at the 20th, 40th, 60th, and 80th percentiles), cancer site (colon vs. rectum), sex, and tumor differentiation (poor vs moderate or well differentiated). Age at the start of study follow-up was based upon randomization date in the trial and birth date and date of chemotherapy initiation in the target population. In the target population, all covariates were obtained from the cancer registry data. We focused on a complete case analysis for the purposes of this methodologic example because of the small size of the trial population.

All-cause mortality

In the HORIZON III trial, patients were followed from randomization; survival was defined as time from first treatment with bevacizumab until time of death or end of study, and all-cause mortality was defined as the complement of survival. In the target population, vital status through 31 December 2013 was derived from the Medicare enrollment files with overall survival defined as time from first claim for bevacizumab until date of death and all-cause mortality again defined as its complement.

Statistical analysis

First, we transferred the HORIZON III trial data onto a secure server containing the linked SEER-Medicare data. We harmonized all study variables and stacked the trial and the target population data into one dataset. We ran multivariable logistic regression models to calculate each individual’s probability of being in the HORIZON III trial conditional on the covariates specified above, analogous to a propensity score.^{2,32} We described the propensity score distribution for trial participation in both the trial and target population as a quantitative measure of population overlap. The greater the difference in mean propensity score between the two populations, the greater the difference with respect to the variables in the propensity score model. A difference in propensity score means of >0.25 standard deviations between the trial and target populations indicates a large amount of extrapolation and heavy reliance on sampling models.¹⁰ We also examined propensity score distributions visualized as a density plot to augment this simple summary measure.

Next, we used the propensity score to calculate sampling weights for the trial population. We used inverse odds of sampling weights as discussed by Westreich et al² to assess transportability (i.e. our target population does not include the sampled population). This method can also be used with the inverse probability of sampling weights described by Cole et al⁸ and Stuart et al¹¹ to assess generalizability when the target population includes the sampled population. These inverse odds of sampling weights are calculated only for trial participants and create a pseudo-trial population that reflects the distribution in the target population of covariates included in the sampling model. We computed weights for each participant in the HORIZON III control arm as the inverse odds of being sampled into the trial from the SEER-Medicare population multiplied by the marginal odds of trial sampling.

$$\text{Stabilized IOSW} = \left(\frac{P(\text{HORIZON})}{1 - P(\text{HORIZON})} \right) * \left(\frac{1 - P(\text{HORIZON} \mid \text{covariates})}{P(\text{HORIZON} \mid \text{covariates})} \right)$$

This results in a mean weight of approximately one and a weighted HORIZON III trial population similar to the SEER-Medicare population sample with respect to the selected covariates. Models included all two-way interactions between each covariate.

We then compared all-cause mortality among the older HORIZON III control arm participants to that of the SEER-Medicare target population before and after applying IOSWs. We used the Nelson-Aalen estimator for cumulative hazards, a Kaplan-Meier analogue that better incorporates weights, to plot survival curves for: (1) the unweighted HORIZON III control arm population, (2) the SEER-Medicare target population, and (3) the weighted HORIZON III control arm population for visual comparison. We used Cox proportional hazards models with sandwich covariance estimators to quantitatively compare the survival curves, assessing the proportional hazards assumption by including an interaction term with log-time in the Cox proportional hazards model. Violations of the proportional hazards assumption would mean agreement between the trial and target varied over time. We calculated risk differences (RDs) in all-cause mortality at 6, 12, and 24 months to examine similarity of all-cause mortality at specific time-points with bootstrapped 95% confidence intervals (CI), estimated based on the standard deviation of 1000 bootstrapped replicates from the trial and target populations. The two populations were bootstrapped separately, with resampling with replacement before restricting to individuals with data on tumor differentiation. If the HRs or RDs estimates comparing all-cause mortality in the trial and target populations differed from the null value after applying the inverse odds of sampling weights, then any non-null treatment effect observed in the weighted trial would be biased in the target population. This bias might be present on only the relative scale, only the absolute scale, or on both scales.¹⁶ While small differences could be explained by variation in sampling, the larger the difference, the greater the possible bias. We repeated analyses targeting SEER-Medicare populations below each frailty cutpoint examined. All analyses were performed in SAS 9.3 (Cary, NC) for Windows.

Sensitivity analyses

We performed sensitivity analyses using a sampling model that excluded variables empirically unassociated with the outcome under the assumption that they could not modify

treatment effects and could amplify existing bias when conditioned upon.³² We assessed the impact of using inverse probability of censoring weights evaluated at the quartiles of drop-out times within the trial and based upon the selected covariates to deal with loss to follow-up over time, ensuring the two populations remained comparable. Since mortality ascertainment was complete for SEER-Medicare patients, these weights were only needed in the trial population. Finally, we assessed the impact of truncating inverse odds of sampling weights at the 95th percentile.

RESULTS

There were 690 patients randomized to the control arm in HORIZON III. Of the 257 patients over 65 years old, 18 (7%) were missing data for tumor differentiation, resulting in a final trial population of 232 patients. Of the 2,335 patients meeting inclusion criteria for the SEER-Medicare target population, 393 (17%) were missing data for tumor differentiation, resulting in a final target population of 1,942 patients (Figure 1). eTable 1 describes marginal distributions of modifiers in the cohort before and after restricting to complete case populations, which were similar, though some limited differences were observed with respect to type of cancer where the absolute standardized mean difference between missing and non-missing patients was 0.10.

Trial and target populations differed substantially with respect to demographic and clinical variables. Table 1 reports the characteristics of the two populations. The median age for trial participants was 70 (interquartile range (IQR): 67 and 73 years). In contrast, the target population had a median age of 72 (IQR: 69 and 77 years). In total, 44% of the trial population was 65-69 compared to 30% of the SEER-Medicare population. Thirty-one percent of the trial population had rectal rather than colon cancer compared to only 21% of the SEER-Medicare population. Sixty-three percent of the trial population was male compared to 54% of the SEER-Medicare population.

Odds ratio estimates for the full model used to generate propensity scores and inverse odds of sampling weights are listed in eTable 2. As expected, the estimated mean probability of sampling (i.e., being represented in the trial) was higher for trial participants at 15.2% than SEER-Medicare patients at 10.1% (difference in mean sampling probability = 5.1%, standardized difference = 0.48). This difference suggests that the trial population is not a random sample of the SEER-Medicare target population with respect to the covariates in the propensity score model. Thus, the trial treatment effects may not be directly extrapolated to the target population if the variables are associated with the outcome. This concern is bolstered by the kernel density plot of propensity scores in Figure 2, showing substantial differences between the two populations. This plot indicates that there are individuals in the trial population corresponding to all the propensity scores in the target population. One individual had a weight of 12.8, but all other weights were below 5. There was some extrapolation with respect to age as a continuous variable (maximum age 92 in the SEER-Medicare and 89 in HORIZON-III).

The weighted trial population was similar to the SEER-Medicare target population in marginal distributions of the variables used in the propensity score model with no

proportions varying by greater than 1%. Absolute standardized mean differences were well under 0.10.³³ Restricting the target population to patients with a predicted probability of frailty or disability less than 7.5% to mimic the ECOG trial exclusion did not result in large changes in marginal distributions of any covariates. This restriction removed 237 individuals from the target population and resulted in a small downward shift in age and the proportion of women included alongside a small decrease in absolute standardized mean difference for age (Table 2). Numbers excluded for other frailty thresholds are listed in eTable 3. We present the full and <7.5% cohort restricted results because the 236 additional patients excluded with a <5% frailty cutoff resulted in substantial decrease to the target population.

Figure 3A shows the survival curves for patients in the full SEER-Medicare population, the unweighted HORIZON III control arm, and the inverse odds of sampling weighted HORIZON III trial control arm. Figure 3B shows analogous survival curves after restricting to the SEER-Medicare population based on a claims-based frailty probability of <7.5%. The trial control arm survival curves were much closer to the SEER-Medicare survival curves after weighting (full cohort trial vs. target weighted HR=0.96, 95% CI: 0.73, 1.26, frailty-restricted trial vs. target weighted HR=1.01, 95% CI: 0.78, 1.30), than before (trial vs. target HR=0.72, 95% CI: 0.58, 0.89, and trial vs. target HR=0.76, 95% CI: 0.61, 0.94, respectively), especially after the first year of follow-up.

We evaluated the proportional hazards assumption by adding an interaction term between population and log-time ($\alpha=0.20$), yielding $p=0.02$ for the term in unrestricted unweighted analysis and $p=0.15$ for the unrestricted weighted analysis. This result led us to reject the hypothesis of proportional hazards. Violation of the proportional hazards assumption was clear when analyses of the time-specific risks showed the weighting performed better at some time points than others. Table 3 reports the number of events and differences in all-cause mortality at 6, 12, and 24-months between the trial and target populations. While the 6-month weighted RDs were similar (unweighted trial vs. target RD= -0.08, 95% CI -0.12, -0.04 and weighted trial vs. target RD= -0.07, 95% CI -0.13, -0.01), there were large shifts in the 12-month RDs (unweighted trial vs. target RD= -0.10, 95% CI -0.16, -0.04 while weighted trial vs. target RD= 0.00, 95% CI -0.09, 0.09) and 24-month RDs (unweighted trial vs. target RD= -0.12, 95% CI -0.22, -0.01 while weighted trial vs. target RD= -0.04, 95% CI -0.16, 0.08) with mortality in the weighted trial population more closely resembling mortality in SEER-Medicare patients than the original HORIZON III population. Frailty restriction of the target SEER-Medicare population using various cut-points attenuated hazard ratios and risk differences towards the null (eTable 4), suggesting the restricted target population was more similar to the trial population, as expected.

Sensitivity analyses

eTable 5 presents results of the multivariable Cox proportional hazards regression of age, sex, tumor differentiation, and type of cancer on mortality. While age and tumor differentiation were strong independent predictors of mortality, type of cancer and sex were not. As a result, we conducted sensitivity analyses excluding type of cancer and sex from the estimation of inverse odds of sampling weights. This resulted in slight shifts of the weighted HR and RDs away from the null. Truncating the eleven individuals with high propensity

score weights at the 95th percentile (inverse odds of sampling weight=2.33) resulted in large shifts away from the null. Applying inverse probability of censoring weights did not substantially change results. Table 4 presents quantitative results of all sensitivity analyses.

DISCUSSION

Using propensity-score based diagnostics to assess external validity can address concerns regarding estimation of treatment effects in specific target populations. In our example we observed marked heterogeneity in several key risk factors, much like recent studies in substance use disorder treatment using similar methods.^{34,35} Fortunately, the weighted survival curves in the trial control population were much closer to those observed in the target population than the unweighted trial control arm population. By using propensity score models and inverse odds of sampling weights based upon age, sex, tumor differentiation, and type of metastatic cancer, we were able to show the HORIZON III trial could be weighted to reflect all-cause mortality observed in the SEER-Medicare target population after 1 year. This result suggests that had the trial reported a non-null treatment effect for cedarinib, these factors ought to have been included when transporting from the HORIZON III trial to target populations similar to SEER-Medicare.

We noted interesting residual differences between the weighted trial and SEER-Medicare populations during the first year of follow-up that weighting and frailty restriction did not ameliorate. In the oncologic setting, these kinds of differences may be the result of more intensive monitoring and protocol directed treatment in the trial population relative to the routine care target population, creating a violation of the consistency assumption.³⁶ Additional baseline variables better predicting duration and completion of treatment or more sophisticated modeling techniques would need to be incorporated if this were the case, as this method cannot distinguish between violations of consistency and exchangeability. The weighted trial population and SEER-Medicare survival curves came together around one year, during which time most patients were likely off protocol directed treatment (due to disease progression); this may reflect differential use of second-line chemotherapy between the weighted and unweighted trial populations. These differences could also be attributable to difficulties excluding patients with very low life expectancies using claims data rather than the detailed in-person clinical evaluations used in cancer treatment trials. Regardless of the reasons for these differences, they increase uncertainty about transporting shorter-term treatment effects from trials with low early mortality to SEER-Medicare populations.

Both the present example and the propensity score-based diagnostics have limitations. The SEER-Medicare data lacked some of the variables used as exclusion criteria in the HORIZON III trial, particularly ECOG status and estimated life expectancy. To counter this deficiency, we explored the impact of restricting to less frail individuals. This slightly improved the performance of the sampling model, though the gap at six months remained largely unchanged. We were also missing some data on tumor differentiation in both the trial and target populations. Our complete case analysis excluded individuals with missing data on tumor differentiation, resulting in reduced precision in our estimates and evaluation of transportability in a population with a potentially different covariate distribution than the target population defined by our inclusion criteria. Implementing multiple imputation or

weighting methods possible in larger data sets would help ensure we are assessing transportability to the correct target population provided the methods are also implemented in the transport of the trial estimates.³⁷

Another limitation of the analysis is that we relied upon expert advice to identify key covariates for the sampling model; adaptive matching or high dimensional variable selection might perform better in creating sampling models, though it may be difficult to harmonize the amounts of variables necessary to use these techniques.³⁸ The general concept of comparing weighted and unweighted outcomes would apply regardless of model selection method, however. We also relied upon straightforward sampling models, though doubly robust methods incorporating outcome modeling in addition to sampling weighting techniques are also being developed.^{9,39,40} For these doubly robust estimation techniques, only one of the models used needs to be correctly specified.^{2,9,41,42} This approach could be adapted to assess the performance of both components of the doubly robust estimator.

Additional assumptions and limitations are inherent in the method itself. First, it assumes external positivity. Thus, these methods cannot be used to transport treatment effects to populations who would have been ineligible for the trial. We can also only use this method to assess external validity in populations where we have outcome data in routine clinical practice (e.g., after receiving Food and Drug Administration approval). Furthermore, this is primarily a method to “rule-out” inadequacy of the sampling model with respect to meeting the assumptions necessary for external validity. Unfortunately, we cannot say with certainty that the sampling model chosen was the correct one because the effects of unmeasured factors on the outcome could cancel one another out in our target population.

Conclusion

The results of these analyses suggest that the direct transport of both the absolute and relative effects of bevacizumab-based treatment on mortality observed in the HORIZON III trial would likely be invalid for a contemporary target population of older adults except in the absence of treatment effect. A sampling model that weighted the HORIZON III trial population to reflect the age, sex, cancer site, and tumor differentiation of the SEER-Medicare target population appeared more appropriate for the transport of long-term effects on mortality. The method presented here represents a useful tool to identify situations in which researchers should be cautious about external validity of treatment effect estimates from RCTs in target populations of interest. Epidemiologists, trialists, and regulators should explore the use of these tools when predicting the effects of treatment for patients in routine care settings.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Financial support: The results reported herein correspond to specific aims of grant 29242-46514-420704 to investigator Jennifer Lund from the Lineberger Cancer Center.

Dr. Lund received research support from the UNC Oncology Clinical Translational Research Training Program (K12 CA120780), as well as through a Research Starter Award from the PhRMA Foundation to the UNC Department of Epidemiology.

Dr. Stürmer receives investigator-initiated research funding and support as Principal Investigator (R01/R56 AG023178, AG056479) from the National Institute on Aging (NIA), and as Co-Investigator (R01 CA174453; R01 HL118255), National Institutes of Health (NIH). He also receives salary support as Director of the Comparative Effectiveness Research (CER) Strategic Initiative, NC TraCS Institute, UNC Clinical and Translational Science Award (UL1TR001111) and as Director of the Center for Pharmacoepidemiology (current members: GlaxoSmithKline, UCB BioSciences, Merck) and research support from pharmaceutical companies (Amgen, AstraZeneca) to the Department of Epidemiology, University of North Carolina at Chapel Hill. Dr. Stürmer does not accept personal compensation of any kind from any pharmaceutical company. He owns stock in Novartis, Roche, BASF, AstraZeneca, and Novo Nordisk.

References

- Hernán MA, Hernandez-Diaz S. Beyond the intention-to-treat in comparative effectiveness research. *Clinical Trials* (London, England). 2012;9(1):48–55.
- Westreich D, Edwards JK, Lesko CR, Stuart E, Cole SR. Transportability of trial results using inverse odds of sampling weights. *American Journal of Epidemiology*. 2017.
- Sesso HD, Gaziano JM, VanDenburgh M, Hennekens CH, Glynn RJ, Buring JE. Comparison of baseline characteristics and mortality experience of participants and nonparticipants in a randomized clinical trial: the Physicians' Health Study. *Controlled Clinical Trials*. 2002;23(6):686–702. [PubMed: 12505246]
- Westreich D, Edwards JK, Rogawski ET, Hudgens MG, Stuart EA, Cole SR. Causal Impact: Epidemiological Approaches for a Public Health of Consequence. *American Journal of Public Health*. 2016;106(6):1011. [PubMed: 27153017]
- Adimora AA, Cole SR, Eron JJ. US Black Women and Human Immunodeficiency Virus Prevention: Time for New Approaches to Clinical Trials. *Clinical Infectious Diseases*. 2017;65(2):324–327. [PubMed: 28383649]
- Westreich D, Edwards JK. Invited commentary: every good randomization deserves observation. *American Journal of Epidemiology*. 2015;182(10):857–860. [PubMed: 26487342]
- Hong J-L, Jonsson Funk M, LoCasale R, et al. Generalizing Randomized Clinical Trial Results: Implementation and Challenges Related to Missing Data in the Target Population. *American Journal of Epidemiology*. 2017:kwx287–kwx287.
- Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations the actg 320 trial. *American Journal of Epidemiology*. 2010;172(1):107–115. [PubMed: 20547574]
- Rudolph KE, Diaz I, Rosenblum M, Stuart EA. Estimating population treatment effects from a survey subsample. *American Journal of Epidemiology*. 2014;180(7):737–748. [PubMed: 25190679]
- Stuart EA, Cole SR, Bradshaw CP, Leaf PJ. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society Series A, (Statistics in Society)*. 2011;174(2):369–386.
- Stuart EA, Bradshaw CP, Leaf PJ. Assessing the generalizability of randomized trial results to target populations. *Prevention Science : the Official Journal of the Society for Prevention Research*. 2015;16(3):475–485. [PubMed: 25307417]
- Murthy VH, Krumholz HM, Gross CP. Participation in cancer clinical trials: race-, sex-, and age-based disparities. *Journal of the American Medical Association*. 2004;291(22):2720–2726. [PubMed: 15187053]
- Tejeda HA, Green SB, Trimble EL, et al. Representation of African-Americans, Hispanics, and whites in National Cancer Institute cancer treatment trials. *Journal of the National Cancer Institute*. 1996;88(12):812–816. [PubMed: 8637047]
- Unger JM, Cook E, Tai E, Bleyer A. The Role of Clinical Trial Participation in Cancer Research: Barriers, Evidence, and Strategies. *American Society of Clinical Oncology Educational Book / ASCO American Society of Clinical Oncology Meeting*. 2016;35:185–198.
- Hernán MA. Invited Commentary: Selection Bias Without Colliders. *American Journal of Epidemiology*. 2017;185(11):1048–1050. [PubMed: 28535177]

16. Pearl J, Bareinboim E. External validity: From do-calculus to transportability across populations. *Statistical Science*. 2014;29(4):579–595.
17. Moodie EEM, Karran JC, Shortreed SM. A case study of SMART attributes: a qualitative assessment of generalizability, retention rate, and trial quality. *Trials*. 2016;17:242. [PubMed: 27180047]
18. Wright JR, Bouma S, Dayes I, et al. The importance of reporting patient recruitment details in phase III trials. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*. 2006;24(6):843–845. [PubMed: 16484692]
19. Hutchins LF, Unger JM, Crowley JJ, Coltman CAJ, Albain KS. Underrepresentation of Patients 65 Years of Age or Older in Cancer-Treatment Trials. *New England Journal of Medicine*. 1999;341(27):2061–2067. [PubMed: 10615079]
20. Rothwell PM. Factors That Can Affect the External Validity of Randomised Controlled Trials. *PLoS Clinical Trials*. 2006;1(1).
21. Rothwell PM. External validity of randomised controlled trials: “to whom do the results of this trial apply?”. *Lancet*. 2005;365(9453):82–93. [PubMed: 15639683]
22. Cahan A, Cahan S, Cimino JJ. Computer-aided assessment of the generalizability of clinical trial results. *International Journal of Medical Informatics*. 2017;99:60–66. [PubMed: 28118923]
23. Moore DA, Goodall RL, Ives NJ, Hooker M, Gazzard BG, Easterbrook PJ. How generalizable are the results of large randomized controlled trials of antiretroviral therapy? *HIV Medicine*. 2000;1(3):149–154. [PubMed: 11737343]
24. Sanoff HK, Carpenter WR, Martin CF, et al. Comparative effectiveness of oxaliplatin vs non-oxaliplatin-containing adjuvant chemotherapy for stage III colon cancer. *Journal of the National Cancer Institute*. 2012;104(3):211–227. [PubMed: 22266473]
25. Lamont EB, Schilsky RL, He Y, et al. Generalizability of trial results to elderly Medicare patients with advanced solid tumors (Alliance 70802). *Journal of the National Cancer Institute*. 2015;107(1):336. [PubMed: 25432408]
26. Bertagnolli MM, Sartor O, Chabner BA, et al. Advantages of a Truly Open-Access Data-Sharing Model. *The New England Journal of Medicine*. 2017;376(12):1178–1181. [PubMed: 28328337]
27. Schmoll H-J, Cunningham D, Sobrero A, et al. Cediranib With mFOLFOX6 Versus Bevacizumab With mFOLFOX6 As First-Line Treatment for Patients With Advanced Colorectal Cancer: A Double-Blind, Randomized Phase III Study (HORIZON III). *Journal of Clinical Oncology*. 2012;30(29):3588–3595. [PubMed: 22965961]
28. Warren JL, Klabunde CN, Schrag D, Bach PB, Riley GF. Overview of the SEER–Medicare data: content, research applications, and generalizability to the United States elderly population. *Medical Care*. 2002;40(8 Suppl):Iv-3–18.
29. Lund JL, Stürmer T, Harlan LC, et al. Identifying specific chemotherapeutic agents in Medicare data: a validation study. *Medical Care*. 2013;51(5):e27. [PubMed: 22080337]
30. Faurot KR, Jonsson Funk M, Pate V, et al. Using claims data to predict dependency in activities of daily living as a proxy for frailty. *Pharmacoepidemiology and Drug Safety*. 2015;24(1):59–66. [PubMed: 25335470]
31. Davidoff AJ, Gardner LD, Zuckerman IH, Hendrick F, Ke X, Edelman MJ. Validation of disability status, a claims-based measure of functional status for cancer treatment and outcomes studies. *Medical Care*. 2014;52(6):500–510. [PubMed: 24638118]
32. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. *American Journal of Epidemiology*. 2006;163(12):1149–1156. [PubMed: 16624967]
33. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*. 2009;28(25):3083–3107. [PubMed: 19757444]
34. Susukida R, Crum RM, Stuart EA, Ebnesajjad C, Mojtabei R. Assessing Sample Representativeness in Randomized Control Trials: Application to the National Institute of Drug Abuse Clinical Trials Network. *Addiction (Abingdon, England)*. 2016;111(7):1226–1234.

35. Susukida R, Crum RM, Ebnesajjad C, Stuart EA, Mojtabai R. Generalizability of findings from randomized controlled trials: application to the National Institute of Drug Abuse Clinical Trials Network. *Addiction* (Abingdon, England). 2017;112(7):1210–1219.
36. Basch E, Deal AM, Dueck AC, et al. Overall Survival Results of a Trial Assessing Patient-Reported Outcomes for Symptom Monitoring During Routine Cancer Treatment. *Journal of the American Medical Association*. 2017;318(2):197–198. [PubMed: 28586821]
37. Hong JL JFM, LoCasale R, Dempster S, Cole SR, and Stürmer T. Generalizing Randomized Clinical Trials: Challenges Due to Missing Data on the Target Population. *Pharmacoepidemiology and Drug Safety*. 2016;25(Supplement):23–24.
38. Najafzadeh M, Schneeweiss S. From Trial to Target Populations - Calibrating Real-World Data. *The New England Journal of Medicine*. 2017;376(13):1203–1205. [PubMed: 28355503]
39. Rudolph KE, Sofrygin O, Zheng W, van der Laan MJ. Robust and flexible estimation of data-dependent stochastic mediation effects: a proposed method and example in a randomized trial setting. *Cornell University*. arXiv:170709021. 2017.
40. Ogburn EL, Rotnitzky A, Robins JM. Doubly robust estimation of the local average treatment effect curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2015;77(2):373–396. [PubMed: 25663814]
41. Hernán MA, VanderWeele TJ. Compound treatments and transportability of causal inference. *Epidemiology*. 2011;22(3):368–377. [PubMed: 21399502]
42. Lesko CR, Buchanan AL, Westreich D, Edwards JK, Hudgens MG, Cole SR. Generalizing Study Results: A Potential Outcomes Perspective. *Epidemiology*. 2017;28(4):553–561. [PubMed: 28346267]

Figure 1: Study flow diagram for HORIZON III participants and SEER-Medicare patients with first metastatic colorectal cancer. Superscripts indicate restrictions applied for A) consistency of treatment across populations, B) access to data and variables for use in the sampling model, and C) external positivity of the populations with respect to the primary risk factor of age.

Figure 2:

Kernel density plots for the predicted probability of sampling for bevacizumab and oxaliplatin-treated participants in the HORIZON trial (solid line) and the bevacizumab and oxaliplatin-treated 2004-2011 SEER-Medicare metastatic colorectal cancer target population (dashed line). The peaks are reflective of the categorical coding of several variables. Every joint combination of covariates contained at least one trial individual.

Figure 3: Hazard ratios (HR) and survival curves comparing outcomes in the bevacizumab and oxaliplatin-treated 2004-2011 SEER-Medicare population and bevacizumab and oxaliplatin-treated HORIZON III patients before and after applying inverse odds of sampling weights. Weights were derived using sampling models containing age modeled with quadratic restricted splines with knots at the 20th, 40th, 60th, and 80th percentiles, sex, tumor differentiation (poor vs. moderate or well differentiated), and type of cancer. Panel A) shows the survival curves when targeting the full SEER-Medicare cohort. Panel B) shows the survival curves when targeting only individuals with predicted probability of frailty < 7.5%.

Table 1:

Characteristics of the Weighted^a and Unweighted HORIZON III Populations and the Bevacizumab and Oxaliplatin-Treated 2004-2011 SEER-Medicare Target Population

Variable	Trial N (%)	Full SEER-Medicare N (%)	Trial population weighted to full population N (%)	Absolute Standardized Mean Difference: Weighted trial vs. SEER-Medicare
Total N	232	1942	232.0	
Age group				
65-69	103 (44)	577 (30)	61.3 (26)	0.07
70-74	85 (37)	621 (32)	81.5 (35)	0.07
75+	44 (19)	744 (38)	88.5 (38)	0.00
Type of cancer				
Rectal	72 (31)	405 (21)	46.9 (20)	0.01
Colon	160 (69)	1537 (79)	184.4 (80)	0.01
Sex				
Male	147 (63)	1045 (54)	124.6 (54)	0.00
Female	85 (37)	897 (46)	106.6 (46)	0.00
Tumor differentiation				
Well or moderate	191 (82)	1321 (68)	154.7 (67)	0.02
Poor	41 (18)	621 (32)	76.5 (33)	0.02

^aWeighted based upon age modeled with quadratic restricted splines with knots at the 20th, 40th, 60th, and 80th percentiles, sex, tumor differentiation (poor vs. moderate or well differentiated), and type of cancer with two-way interaction terms between all variables

Table 2:

Characteristics of the Unweighted and Weighted^a HORIZON III Trial Populations and the Bevacizumab and Oxaliplatin-Treated 2004-2011 SEER-Medicare Target Populations, Restricted Based on <7.5% Predicted Probability of Frailty From the Fautot Claims-Based Frailty Algorithm

Variable	Trial N (%)	Non-frail SEER-Medicare N (%)	Trial weighted to SEER-Medicare population N (%)	Absolute Standardized Mean Difference: Weighted trial vs. SEER-Medicare
Total N	232	1705	231.9	
Age group				
65-69	103 (44)	544 (32)	67.9 (29)	0.06
70-74	85 (37)	568 (33)	81.0 (35)	0.03
75+	44 (19)	593 (35)	83.4 (36)	0.02
Type of cancer				
Rectal	72 (31)	360 (21)	47.0 (20)	0.02
Colon	160 (69)	1345 (79)	185.4 (80)	0.02
Sex				
Male	147 (63)	883 (52)	120.5 (52)	0.00
Female	85 (37)	822 (48)	111.8 (48)	0.00
Tumor differentiation				
Well or moderate	191 (82)	1156 (68)	156.00 (67)	0.01
Poor	41 (18)	549 (32)	76.3 (33)	0.01

^aWeighted based upon age modeled with quadratic restricted splines with knots at the 20th, 40th, 60th, and 80th percentiles, sex, tumor differentiation (poor vs. moderate or well differentiated), and type of cancer with two-way interaction terms between all variables

Table 3:

Risk Differences Comparing Survival at 6-, 12-, and 24-Months in the Bevacizumab and Oxaliplatin-Treated Patients with a First Metastatic Colorectal Cancer in HORIZON III and 2004-2011 Bevacizumab and Oxaliplatin-Treated SEER-Medicare Patients Before and After Applying Sampling Weights^a

Time point	Bevacizumab-treated population	Person-time (years)	Deaths	Risk	Crude trial vs. target RD	95% C.I.	Weighted ^a trial vs. target RD	95% C.I.
6 months	SEER-Medicare target	887	312	0.16	-0.08	-0.12, -0.04	-0.07	-0.13, -0.01
	HORIZON III trial	109	19	0.08				
12 months	SEER-Medicare target	1,621	667	0.34	-0.10	-0.16, -0.04	-0.00	-0.09, 0.09
	HORIZON III trial	200	52	0.24				
24 months	SEER-Medicare target	2,602	1200	0.62	-0.12	-0.22, -0.01	-0.04	-0.16, 0.08
	HORIZON III trial	251	76	0.50				

Abbreviations: C.I., confidence interval; RD, risk difference

^aWeighted based upon age modeled with quadratic restricted splines with knots at the 20th, 40th, 60th, and 80th percentiles, sex, tumor differentiation (poor vs. moderate or well differentiated), and type of cancer with two-way interaction terms between all variables

Table 4:

HRs and RDs Comparing Survival in the Bevacizumab and Oxaliplatin-treated Patients with First Metastatic Colorectal Cancer in the 2004-2011 SEER-Medicare Population with Survival in the HORIZON III Trial in Sensitivity Analyses

Analysis	HR	95% C.I.	6-month RD	95% C.I.	1-year RD	95% C.I.	2-year RD	95% C.I.
Fewer covariates ^a	0.91	0.71, 1.16	-0.06	-0.12, 0.00	-0.02	-0.10, 0.06	-0.06	-0.18, 0.05
Truncating weights ^b	0.86	0.66, 1.07 ^c	-0.07	-0.12, -0.04	-0.05	-0.13, 0.03	-0.07	-0.18, 0.04
Inverse probability of censoring weights ^d	0.95	0.67, 1.19 ^c	-0.06	-0.12, -0.01	0.01	-0.11, 0.08	-0.03	-0.18, 0.07

Abbreviations: C.I., confidence interval; HR, hazard ratio; IOSW, inverse odds of sampling weights; RD, risk difference

^aWeighted based upon age modeled with quadratic restricted splines at the 20th, 40th, 60th, and 80th percentiles and tumor differentiation (poor vs. moderate or well differentiated) with a two-way interaction term

^bAt the 95% percentile

^cBecause these methods created asymmetry in the bootstrap distribution, we used the 2.5th and 97.5th confidence limits to construct the 95% C.I.

^dIncluded age, tumor differentiation, sex, and type of cancer in the censoring model; omitted interaction terms due to convergence issues in the bootstrap