11-5-2014

CROSS-DESIGN SYNTHESIS FOR EXTENDING THE APPLICABILITY OF TRIAL EVIDENCE WHEN TREATMENT EFFECT IS HETEROGENEOUS. PART II. APPLICATION AND EXTERNAL VALIDATION

Carlos Weiss
Department of Family Medicine, Michigan State University

Ravi Varadhan
Division of Biostatistics and Bioinformatics, Department of Oncology, Johns Hopkins University, ravi.varadhan@jhu.edu

Cross-Design Synthesis for Extending the Applicability of Trial Evidence when Treatment Effect is Heterogeneous: Part II. Application and External Validation

Carlos O. Weiss[a] and Ravi Varadhan[b,c]

Abstract

Randomized controlled trials (RCTs) generally provide the most reliable evidence. When participants in RCTs are selected with respect to characteristics that are potential treatment effect modifiers, the average treatment effect from the trials may not be applicable to a specific target population. We present a new method to project the treatment effect from a RCT to a target group that is inadequately represented in the trial when there is heterogeneity in the treatment effect (HTE). The method integrates RCT and observational data through cross-design synthesis. An essential component is to identify HTE and a calibration factor for unmeasured confounding for the observational study relative to the RCT. The estimate of treatment effect adjusted for unmeasured confounding is projected onto the target sample using G-computation with standardization weights. We call the method Calibrated Risk-Adjusted Modeling (CRAM) and apply it to estimate the effect of angiotensin converting enzyme inhibition to prevent heart failure hospitalization or death. External validation shows that when there is adequate overlap between the RCT and the target sample, risk-based standardization is less biased than CRAM. However, when there is poor overlap between the trial and the target sample, CRAM provides superior estimates of treatment effect.

[a] Department of Family Medicine, Michigan State University, Grand Rapids, USA
[b] Division of Biostatistics and Bioinformatics, Department of Oncology, Johns Hopkins University, USA
[c] Department of Biostatistics, School of Public Health, Johns Hopkins University, USA

Keywords: observational data, unmeasured confounding, sensitivity analysis, interaction, heterogeneity, standardization, generalizability, internal and external validity

1. Introduction

Mainly due to a lack of external validity, some medical guideline-makers and policy-makers believe that the evidence base from human trials is inadequate, i.e. that trials provide insufficient evidence to guide care for key target populations.[1, 2] In recent years many have called for trials that are less exclusionary and more pragmatic.[3, 4] Selection by trials is not entirely avoidable, as is demonstrated by informed consent. Therefore, limited external validity is inherent to a trial design.

Together with a companion article that details statistical methodology and simulations, this report presents a novel method to project the treatment effect from a trial to a target group that is poorly represented in the trial, thus extending the applicability of trial evidence. Here we present a clinical application and external validation. The method does this by using individual-level data from a trial and observational data where the observational data has a broader distribution of effect modifiers such that it partially overlaps the trial sample and the target sample. Then the treatment effect from the trial can be projected to the target sample. The essential component of the method is to identify a calibration factor for unmeasured confounding for the observational study relative to the trial. The calibration factor makes it possible to estimate a treatment effect in the observational data with adjustment for unmeasured confounding.

Heart failure (HF) is the clinical example for demonstrating the proposed methods. HF prevalence is approximately 7.9% among Medicare beneficiaries and 12.0% among those dually enrolled in Medicare and Medicaid.[5] Despite this importance to policy and providers, fundamental questions remain about use of one of the most important treatments, angiotensin converting enzyme inhibition (ACEi). A major guideline concisely stated the problem with respect to the effectiveness of ACEi:[6] "…primarily men are enrolled in clinical trials of treatments for HF; however, the majority of patients with HF in the general population are women… [and] most large, multicenter trials have not included sufficient numbers of women to allow conclusions about the efficacy and safety of their treatment... Some research has suggested that women with HF, particularly with asymptomatic reduced left ventricular ejection

fractions, may not show survival benefits from ACE inhibition." A previous attempt to address the question of efficacy of ACEi in women used aggregate data meta-analysis[7] and concluded that "women with asymptomatic [left ventricular] systolic dysfunction may not achieve a mortality benefit when treated with ACE inhibitors." There is evidence that the distribution of characteristics such as gender and old age is different in the target population than in the evidence base. For example, women above age 75 make up 51% of people with HF in Medicare and 35% in Medicaid (Data are our calculations using National Health and Nutrition Examination Survey, 1999-2004) but the proportion of women above age 75 in trials is lower as the cited guidelines states and we will show. The objectives of the clinical application study were to demonstrate the utility of CRAM for estimating treatment effect in a target sample and to externally validate CRAM results.

2. Data Sources and Participants

Studies of Left Ventricular Dysfunction (SOLVD) Trials and Registry

The SOLVD trials were designed to understand whether the ACEi enalapril reduced mortality in people with a low left ventricular ejection fraction (≤0.35).[8] Designed as a pair, the treatment trial (SOLVD-T, *n*=2,569) recruited patients with an overt history of HF and the prevention trial (SOLVD-P, *n*=4,228) did not. The trial reported estimated a treatment benefit for the primary outcome of death in SOLVD-T (SOLVD-T relative risk 0.86, *P*-value 0.0036; SOLVD-P 0.92, *P*-value 0.30) and benefit for the secondary outcome of death or HF hospitalization in both trials (SOLVD-T relative risk 0.74, *P*-value 0.0001; SOLVD-P 0.80, *P*-value <0.001).[9, 10] A related registry (SOLVD-R, *n*=5,100 not also in a SOLVD trial) was created to understand the natural history of patients with a low left ventricular ejection fraction (≤0.35) or an overt history of HF.[11] We use the SOLVD-T trial as the source of evidence, SOLVD-R registry as the bridge sample and the SOLVD-P as a target sample. In practice a trial would not be chosen as the target sample but we did so in order to have the ability to provide external validation of the treatment effect projected to it, as further explained below.

There was evidence of strong selection into the trials as samples were 6.4% and 7.4% of people screened and eligible for SOLVD-T and SOLVD-P, respectively.[9] The total number of people screened and eligible was not known for SOLVD-R. However, the number

of exclusion criteria was 26 for the trials and only 5 for the registry. We reviewed reported SOLVD inclusion and exclusion criteria one at a time (Appendix Table 1).

The main outcome chosen for this study was HF hospitalization or death at 1 year, an outcome that was pre-specified by trial designers. The combined outcome provided more power to examine HTE and was valid because the direction of treatment effect was the same for each component outcome.[12] The time horizon of 1 year created synchronization across the trials and registry. Details concerning measurement for the outcome and predictors, which are listed in Table 1, are published.[8-11]

3. Testing for heterogeneity of treatment effect in the trial

When the treatment effect is heterogeneous, the average treatment effect from the trials is neither generalizable to the larger at-risk population nor applicable to a specific target sample. Therefore, ascertaining whether HTE is present is an essential aspect of our methodology. Various patient characteristics can modify the effect of the treatment. We follow a large literature demonstrating that baseline risk of the outcome is often a potent treatment effect modifier on the relative risk scale,[13-20] However, any variable found to be a strong treatment effect modifier could be used in place of baseline risk for the proposed method. We collected external evidence regarding which patient characteristics predict the outcome by performing a literature search and asking local experts to identify validated equations. In this way we identified 37 externally-validated predictors (Appendix Tables 2 and 3).[21-27] For each predictor we examined study datasets to identify corresponding data (Appendix Table 4). In some cases we employed clinical judgment regarding equivalence. E.g., serum hematocrit/3 was used when serum hemoglobin concentration was not available. Having established which predictors were present in at least one study, we examined the proportion of missing data. We use predictors with ≥85% non-missing values during risk modeling (Appendix Table 5).

To obtain baseline risk estimates we entered treatment and other predictors into a Cox proportional hazards model.[28] There was a separate model for the trial and bridge sample, which was then applied to the target sample. This was done to imitate practice settings where outcomes data for the target sample may not be available. The cumulative baseline hazard at 1 year and the linear

predictor from the model enabled calculation of individual baseline risk at 1 year.[29] Plots of risk calibration provided an assessment of the validity of individual risk estimates. Specifically, we made risk quintile groups with predicted risk, calculated the observed risk for each quintile with the Kaplan-Meier survival function[30] and then plotted the geometric mean of observed risk versus predicted risk for each quintile (Appendix Figure 1).

We also use the Follmann test, which tests for HTE in one step.[31] The Follmann test provides a likelihood ratio test statistic for H_0: $\beta_0 = \beta_1$ versus H_A: $\alpha\ \beta_0 = \beta_1$, $\alpha \neq 1$, where β is the regression coefficients for an underlying vector of covariates that identifies HTE. $\alpha<1$ is a negative interaction, $\alpha>1$ is a positive interaction and $\alpha=1$ is the null hypothesis. This initial step to identify HTE is done on the pooled trial and bridge samples. In this example using survival analysis, baseline hazard is allowed to vary between the two sample strata.

4. Standardization of treatment effect from trial to target sample

Two standardization approaches are compared because simulation studies have shown that neither approach is uniformly better than the other in terms of bias (see companion paper).

4.1 Covariate-based standardization

A logistic regression model for predicting study membership (S=1 denotes membership in the bridge sample and S=0 denotes membership in trial) on the basis of covariates is fitted using a boosted approach. The covariates chosen were personal characteristics known to be associated with the outcome (e.g. age, gender, chronic disease status). For each individual in the trial, the probability of being in the bridge sample, p_1 = prob(S=1), is estimated. Each person in the trial is weighted with the odds of being in the bridge sample, which is $p_1/(1-p_1)$. Treatment effect was estimated using a weighted proportional hazard model with treatment indicator as the only covariate. A robust standard error is computed using a sandwich estimator.[32] When the Follmann test is significant, simulation studies show that the next approach with risk-based standardization is less biased (see companion paper).

4.2 Risk-based standardization

Risk-based standardization is better than covariate-based standardization when baseline risk is an effect modifier and the Follmann test of interaction is significant. This approach is like covariate based standardization except that instead of covariates study membership is predicted as a function of 1-year risk of outcome. In addition a flexible logistic regression model with a spline function was used to model the baseline risk variable, fitted using the generalized additive model approach.[33]

5. Calibrated, Risk-Adjusted Modeling

For a brief conceptual overview of CRAM, please see section 2 of a companion paper. The CRAM approach utilizes a bridging sample to provide an evidentiary link between the evidence source and target sample.

5.1 The causal estimand

The goal is to estimate the effect of treatment where A = 1 for enalapril treatment and A=0 for control, on the failure time T, which is subject to right censoring. Let T_a denote the potential failure time when A=a (where a=0, 1). Here the causal parameter of interest is the marginal log-hazard ratio, defined as:

$$\psi_t = \log[-\log(\text{prob}(T_1 > t)] - \log[-\log(\text{prob}(T_0 > t)] , \quad (1)$$

The models for the hazard rate of failure time, $\lambda(t)$, which is defined as: prob $(t < T \leq t+dt \mid T \geq t)$ dt, where dt is a small increment of time. The Cox proportional hazards model is used throughout to estimate the marginal log-hazard ratio. In the trial sample, this is the standard intent-to-treat effect (without any covariate adjustment). In the observational bridge sample, ψ_t is computed using the G-computation formula given in Stitelman et al.[34] This is a causal estimate of treatment effect under the usual assumptions of consistency, no unmeasured confounding and correct model specification for the hazard rate.[35] The intent-to-treat effect is estimated, so there is the assumption that non-compliance in the trial is negligible.

[Table 1 about here]

A comparison of the proposed novel modeling approach to inverse probability of treatment weighted and standardization is presented in Table 1. Major differences among modeling approaches are shown, including the major focus on HTE and unmeasured confounding provided only by CRAM. Standardization approaches are partially cross-design synthetic approaches in the sense that they use information from both the trial and bridge sample. However, they are only partially synthetic because they do not use the treatment status and outcome data from the bridge sample. Furthermore they do not focus on HTE. CRAM employs full cross-design synthesis because it uses information about the treatment status and outcome from the bridging study to examine HTE as well as unmeasured confounding.

5.2 Accounting for heterogeneity of treatment effect and unmeasured confounding in the bridge sample

The next step is to account for unmeasured confounding by finding the parameter that optimally matches the bridge sample's treatment effect to the trial treatment effect. This is the CRAM calibration step. It can be shown that an unmeasured confounder can be captured by a single variable that when incorporated, for example in model (2), will provide an unbiased estimate for the marginal treatment effect formula (1). A survival (proportional hazards) model for the outcome that depends on treatment A (= 0,1) and underlying baseline risk R (1-year risk of outcome) was used. Since there is uncertainty about the correct model for the potential outcome, T_a, two models are used:

$$\lambda(t) = \lambda_0(t) \,.\, \exp(\beta_A A + \beta_R R + \beta_{AR} A*R + \beta_U U) \qquad (4)$$

$$\lambda(t) = \lambda_0(t) \,.\, \exp(\beta_A A + \beta_R R + \beta_{AR} A*R + \beta_U U + \beta_{AU} A*U) \qquad (5)$$

where $\lambda_0(t)$ is the baseline hazard for the outcome. Model (4) allows two-way interactions of treatment A only with baseline risk, and the unmeasured confounder U enters only as a first-order term. The unmeasured confounder U is modeled as described in the companion paper. Model (5) allows two-way interactions of treatment A with baseline risk as well as with unmeasured confounder U.

CRAM estimates that use information from the bridge sample are compared to estimates derived by standardizing the trial sample to the target sample. The standardization methods are the same as used to standardize the trial sample to the bridge sample, as discussed in section 4. Covariate based or risk based standardization can be used, depending on the results of the test for interaction or Follmann test, as discussed in section 3.

To compare CRAM to standardization estimates of treatment effect when there is little overlap between the trial and target sample, we repeated the projection procedures described above for a more "distant" target sample. This was comprised by the lowest 3 quartiles of risk in the target sample, i.e. those with the least overlap with the trial sample. We also examined the conventional target sample suggested by the clinical literature: women older than 75 years.

6. External Validation

We compare the projected log hazard ratios to the actual treatment effect in the clinically appropriate target sample, which our case happens to be the SOLVD-P trial so that the CRAM method's treatment effect can be externally validated. An additional strength of our example is that the evidence and target sample were designed together so data collection was coordinated and data measurement discrepancies were absent. We compute two treatment effects in the target sample. The first is the treatment effect that is unadjusted for any prognostic variables:

$$\lambda(t) = \lambda_0(t) \,.\, \exp(\beta_A A) \qquad (6)$$

The second is the marginal treatment effect obtained by applying the G-computation formula to the following model that conditions on baseline risk:

$$\lambda(t) = \lambda_0(t) \,.\, \exp(\beta_A A + \beta_R R) \qquad (7)$$

7. Results

[Table 2 about here]

By comparing patient characteristics one at a time (Table 2) we found that the SOLVD trials compared to the SOLVD-R had a lower proportion women compared to the observational studies (11.3-19.6% vs. 28.8-42.8%), and a markedly lower proportion of older women (0.6-1.4% vs. 8-20.8%). However, the trials had a much higher proportion history of myocardial infarction and Instrumental Activities of Daily living disability compared to the observational studies, and a higher proportion of former or current smoking, history of diabetes mellitus and fair or poor general health. The trials had a lower proportion of history of atrial fibrillation, history of chronic obstructive pulmonary disease or hospitalization in the last year and on average a lower systolic blood pressure. In both trials, the unadjusted treatment effect favored treatment however in the target sample the treatment effect was qualitatively different and suggested treatment harm. Of course, the unadjusted treatment effect in the target sample is expected to be confounded by treatment selection.

7.1. Identification of heterogeneity of treatment effect in the trial

The likelihood ratio test of HTE according the baseline risk in the trial was significant with the Cox proportional hazards model (P-value = 0.0013). The Follmann parameter was 0.68. A result <1 implies that as baseline risk increases the treatment effect expressed as a log hazard ratio becomes more negative, i.e. there is stronger risk reduction. Similar to the likelihood ratio test using the Cox model, the Follmann test for HTE according to a linear combination of covariates yielded a highly significant result (LRT statistic = 12.9, P-value = 0.00033).

7.2. Distributions of treatment effect modifier in the different samples

[Figure 1 about here]

Visual comparison of risk distributions showed that the SOLVD-T trial had the highest distribution of 1-yr risk of HF hospitalization or death. This distribution was shared partially with the distributions in the SOLVD-R bridge sample and SOLVD-P target sample

(Figure 1). In other words, the bridge sample and to a greater degree the target sample were selected to be not as high risk as the trial. Combined with the information from the Follmann test above, this meant that less strong treatment effect i.e. less risk reduction would be expected in SOLVD-R and SOLVD-P compared to SOLVD-T.

[Table 3 about here]

7.3 Calibration - Accounting for unmeasured confounding in the bridge sample

An unmeasured confounder U that decreases the probability of taking treatment, but increases the mean failure time (i.e. increasing levels of confounder decreases the failure rate), would account for the discrepancy between bridge sample and standardized treatment effects noted in Table 2. Conversely, a U that increases the probability of taking treatment, but decreases the mean failure time (i.e. increasing levels of confounder increases the failure rate), would have the same effect. Therefore, it was sufficient to consider only one direction. $\mu_1 < 0$ and $0 < \rho < 1$ were considered. To examine the sensitivity of results to U, CRAM was performed for two different values of μ_1, -0.5 and -1.0. For model (4) with μ_1 -0.5 and -1.0, respectively, the optimal ρ^* were 0.70 and 0.49. For model (5) with μ_1 -0.5 and -1.0, respectively, the optimal ρ^* were 0.72 and 0.56. As expected for both models, when the effect of the unmeasured confounder on treatment selection was stronger (μ_1 had a larger absolute value), the effect of the unmeasured confounder on failure time needed to calibrate to the same treatment effect was smaller (ρ^* was weaker). In addition, when the confounded treatment effect was larger than the trial treatment effect (to the right on a log scale or further from the null on a hazard ratio scale), μ_1 and ρ^* have different signs.

[Tables 4 and 5 about here]

7.4 Standardization and CRAM results

As shown in Table 1, if the treatment effect from the evidence source were assumed to apply to the target sample, ignoring sample selection and HTE, the true treatment effect (using Eq. 1) was a log hazard ratio of -0.51 (standard error=0.08). Covariate-based standardization of the trial to the target sample estimated the treatment effect as a log hazard ratio of -0.50 (standard error=0.22; Z-

test statistic 2.27) and risk-based standardization of the trial to the target sample estimated the treatment effect as a log hazard ratio of -0.31 (standard error=0.13; Z-test statistic=2.38). Using CRAM the projected treatment effect was -0.45 or -0.43 (standard error=0.13 or 0.16, respectively) using models (4) and (5), as shown in Table 3. The standard errors demonstrated that covariate-based standardization and CRAM were more efficient than risk-based standardization.

To compare CRAM to standardization estimates of treatment effect when there is little overlap between the trial and target sample, we also repeated the projection procedures described above for modifier distant target sample, as shown in Table 4. The true treatment effect in this sample (using Eq. 1) was -0.35 (standard error=0.19). Covariate-based standardization estimated the treatment effect as a log hazard ratio of -0.55 (standard error=0.43) and risk-based standardization of the trial to the distant target sample estimated the treatment effect as a log hazard ratio of -0.11 (standard error=0.19; Z-test statistic=0.59). The projected treatment effect using CRAM was closest at -0.28 to -0.33 (standard error=0.18 to 0.21).

Finally we compared CRAM to standardization estimates of treatment effect to the predefined conventional target sample of women older than 75 years, as shown in Table 5. Risk-based standardization of the trial to women older than 75 years estimated the treatment effect as a log hazard ratio of -0.64 (standard error=0.13; Z-test statistic=5.06). For the conventional target sample the projected treatment effect using CRAM was -0.41 to -0.44 (standard error=0.08 to 0.09).

7.5 External validation of CRAM results

By using a randomized trial as the target sample in this methodology study, we are able to provide externally valid treatment effect estimates. In the target sample, the treatment effect was log hazard ratio -0.28 (SE 0.10, P-Value = 0.006). As can be shown by comparing results to Table 3, risk-based standardization performed superiorly to covariate-based standardization and CRAM. In the distant target sample, the treatment effect was log hazard ratio -0.37 (SE 0.16, P-Value = 0.023). As can be shown by comparing results to Table 4, estimates from CRAM, using models that incorporate risk-based heterogeneity, performed superiorly to either standardization approach. We could not estimate the treatment for women older than 75 years in the validation trial SOLVD-P due to

small sample size. Therefore, it was not possible to externally validate CRAM or standardization results for women older than 75 years.

8. Discussion

This report presents a novel method to project the estimation of a treatment effect from a trial or trials to people who were not in the trial using observational data. In order to draw on strengths and address limitations that are inherent to each design, the method uses individual-level data from a synthetic study design dyad with a trial and observational studies. Attention to HTE and a calibration factor for unmeasured confounding for the observational study relative to the trial make it possible to estimate a treatment effect in the observational data. In this way the method uses empirical data to inform a sensitivity analysis for unmeasured confounding of treatment effects for people not in the trial. The proposed methods will facilitate the synthesis of evidence base provided by trials, under appropriate conditions, with evidence from target populations not included by trials. The methods can be used for a trial and registry or a trial and a cohort study.

CRAM differs from standard meta-analysis, which typically combines the results of several studies with similar design, in two important ways: it integrates studies with different designs and it uses individual-level data to understand and account for HTE and biases. The ability of CRAM to remove unmeasured confounding from the registry data in the bridge sample and provide a sensible estimate of treatment effects in the target sample is promising. Importantly, this report found a qualitative discrepancy between the treatment effect estimated for the target sample using only observational data and the treatment effect estimated with CRAM. Using only observational data with adjustment for treatment selection according to measured factors, treatment was estimated to increase risk of the primary outcome; using CRAM or the standardization approach, treatment was estimated to decrease the risk. The approach we compared to CRAM, inverse probability treatment weighting, estimates treatment effect using only the target sample. This approach might yield unbiased marginal treatment effects only when there is no unmeasured confounding. Standardization is more like cross-design synthesis approaches in the sense that it uses information from both the trial and bridge sample. However, standardization is only partially synthetic it does not use the treatment status and outcome data from the bridge sample. Furthermore

standardization approaches do not explicitly examine HTE. CRAM uses information about the treatment status and outcome from the bridging study and examines both HTE and unmeasured confounding.

Inverse probability of treatment weighted (IPTW) would commonly be used to model treatment effect in observational data. An approach based on IPTW does not require a model for the outcome - just information on the outcome - which is a desirable feature since a large number of covariates and a powerful machine learning approach can be used not only to predict treatment selection but also to optimize covariate balance. However, IPTW does not focus on HTE and does not leverage information from the target sample. In our example, IPTW is not a relevant comparator because we use a randomized treatment trial as our target sample in order to make external validation possible.

The findings of our external validation were that when there is substantial overlap between the trial and the target sample, standardization approach was better than CRAM, i.e. it was less biased than CRAM. However, when there is poor overlap between the trial and the target sample, CRAM was superior to standardization and provided valid estimates of treatment effect. Why did CRAM perform superiorly to standardization when the target sample was distant from the trial sample? This is likely due to 2 reasons: first, a greater overlap between the bridge sample used by CRAM and the distant target sample, and second, trial standardization is not outcome model based, whereas CRAM projection is based on outcome model that incorporates HTE. Consequently the models which incorporate risk-based effect heterogeneity might be expected to do better when there is indeed heterogeneity due to baseline risk. A vast majority of the standardization weights from the trial to the distant target sample are 0, whereas there is a smaller proportion of zero weights when standardizing the bridge sample to the distant target sample: approximately 60% for the trial sample versus 40% for the bridge sample (Figure 2). Further, the larger non-zero weights for trial sample are larger than those of the bridge sample. This results in standardization estimates being more unstable than the CRAM estimate that uses the bridge sample to calibrate.

[Figure 2 about here]

This report also illustrates how evidence regarding older women appeared lacking in the studied trial. There were too few older women in the trials to analyze as a subgroup (Table 1). However, trial participants may be different from a target group in many ways, not just according to age group and gender. We also found that there was varying direction to the correlations between a patient characteristic and being in the trial (e.g., some co-morbid diseases were positively correlated with being in the trial, others negatively). We visually examined the distribution of women older than 75 and found that they were actually distributed across the range of baseline risk shown in Figure 1. These findings suggest that general rules regarding selection into trials (such as the idea that trials exclude older, sicker people) should be viewed with skepticism and labels such as "older women" are not very informative. Even though SOLVD-T included few older women, its results may still apply many older women to the extent that exchangeability needs to be judged according to many characteristics, not just age and gender.

There are several clinical implications of the CRAM results. We found HTE in the trial, which makes a formal assessment of applicability of varying treatment effects essential. HTE according to baseline risk of outcome in SOLVD-T has not been reported previously to our knowledge. The original trial report conducted 4 pre-specified subgroup analyses, based on tertiles: serum sodium concentration, ejection fraction and etiology of heart failure. In addition to the pre-specified analyses, due to interim information a subgroup analysis was conducted according to 4 levels of New York Heart Association classification. There was a suggestion of interaction by ejection fraction such that the tertile with the highest ejection fraction did not benefit (interaction P-value = 0.03).[9] We report highly significant HTE according to baseline risk such that people at lower risk received smaller treatment benefit. This finding is consistent with prior research showing that multivariate risk scores have greater power than univariate variables to identify HTE.[36] While standardization was useful for estimating the treatment effect for people at somewhat lower risk of the outcome, such representing the overall SOLVD-P sample, CRAM performed better for people at lower risk such as those representing a distant target sample. In the case of HF, as with other conditions,[37] people at low risk comprise the majority of people not enrolled in a trial.

The CRAM method makes assumptions and has requirements that should be examined carefully. CRAM assumes that the settings for unmeasured confounding obtained in the bridge sample are also applicable to the target sample. This is not too unrealistic. It is

also sensible, if CRAM is viewed as an analytic device that uses trial data to narrow the otherwise wide scope of sensitivity analyses for unmeasured confounding in observational data. For projecting a treatment effect for the target sample, CRAM assumes that the same outcome model that is used in the bridge sample also holds in the target sample. The approach here assumed a non-time-varying intervention and did not deal with different patterns of adherence to treatment. CRAM relies on individual-level data for important patient characteristics. While the requirement of individual level data from trials and observational studies limits feasibility across scenarios, we believe that individual-level data is essential to validly account for HTE and treatment selection bias. A real limitation is that the method accounts for heterogeneity due to patient-related characteristics, but not for design-related features. This is because design-related features that contribute to HTE can vary from one study to another, and consequently it is much more challenging to develop a general methodology to account for all of them.

In order to perform CRAM it is necessary to have sufficient overlap between a trial and observational study to perform calibration. This will not always be available. Furthermore, when there is overlap the observational study must have sufficient people exposed and unexposed to treatment in the bridge sample. An important assumption is that a proper metric for exchangeability and calibration was used. CRAM assumes that important measurement discrepancies between studies have been addressed and that missing data is ignorable. Fortunately, the plausibility of each of these assumptions can be assessed

In summary, CRAM integrates studies with different designs and uses individual-level data to rigorously understand and account for heterogeneity of treatment effects and biases in observational (bridging) data in order to apply treatment effects from a trial to people not adequately represented in the trial. External validation showed that when there is adequate overlap between the trial and the target samples, risk-based standardization was less biased than CRAM. However, when there is poor overlap between the trial and the target samples, CRAM provided superior estimates of treatment effect.

Funding

This work was supported by the Agency for Healthcare Research and Quality, US Department of Health and Human Services, as part of the Developing Evidence to Inform Decisions about Effectiveness (DEcIDE) program [contract no. HHSA29020050034-I-TO5]. The authors of this report are responsible for its content. Statements in the report should not be construed as endorsement by the Agency for Healthcare Research and Quality or the US Department of Health and Human Services. This work used research materials obtained from the National Heart Lung and Blood Institute (NHLBI) Biologic Specimen and Data Repository Information Coordinating Center and does not necessarily reflect the opinions or views of the Studies of Left Ventricular Dysfunction or the NHLBI.

Acknowledgements

We acknowledge insightful and helpful discussions with Cynthia Boyd and Jodi Segal as part of the larger project that produced this work.

Conflict of Interest Statement

The authors declare that there is no conflict of interest.

REFERENCES

1. Atkins D. Creating and synthesizing evidence with decision makers in mind: integrating evidence from clinical trials and other study designs. *Med Care*. 2007; 45: S16-22.
2. Sox HC and Greenfield S. Comparative effectiveness research: a report from the Institute of Medicine. *Ann Intern Med*. 2009; 151: 203-5.
3. Tunis SR, Stryer DB and Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. *Jama*. 2003; 290: 1624-32.
4. Zwarenstein M, Treweek S, Gagnier JJ, et al. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. *Bmj*. 2008; 337: a2390.
5. Boyd CM, Leff B, Weiss CO, Wolff JL, Clark R and Richards T. Full Report: Clarifying Multimorbidity to Improve Targeting and Delivery of Clinical Services for Medicaid Populations. *Faces of Medicaid*. Princeton, NJ2010, p. 31.
6. Hunt SA, Abraham WT, Chin MH, et al. 2009 Focused update incorporated into the ACC/AHA 2005 Guidelines for the Diagnosis and Management of Heart Failure in Adults A Report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines Developed in Collaboration With the International Society for Heart and Lung Transplantation. *J Am Coll Cardiol*. 2009; 53: e1-e90.
7. Shekelle PG, Rich MW, Morton SC, et al. Efficacy of angiotensin-converting enzyme inhibitors and beta-blockers in the management of left ventricular systolic dysfunction according to race, gender, and diabetic status: a meta-analysis of major clinical trials. *J Am Coll Cardiol*. 2003; 41: 1529-38.
8. Studies of left ventricular dysfunction (SOLVD)--rationale, design and methods: two trials that evaluate the effect of enalapril in patients with reduced ejection fraction. *Am J Cardiol*. 1990; 66: 315-22.
9. Effect of enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure. The SOLVD Investigators. *N Engl J Med*. 1991; 325: 293-302.
10. Effect of enalapril on mortality and the development of heart failure in asymptomatic patients with reduced left ventricular ejection fractions. The SOLVD Investigattors. *N Engl J Med*. 1992; 327: 685-91.
11. Bangdiwala SI, Weiner DH, Bourassa MG, Friesinger GC, 2nd, Ghali JK and Yusuf S. Studies of Left Ventricular Dysfunction (SOLVD) Registry: rationale, design, methods and description of baseline characteristics. *Am J Cardiol*. 1992; 70: 347-53.
12. Varadhan R, Weiss C, Segal J, Wu A, Sharfstein D and Boyd CM. Evaluating Health Outcomes in the Presence of Competing Risks: A Review of Statistical Methods and Clinical Applications. *Med Care*. 2010; 48: S96-S105.
13. Kent DM and Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. *Jama*. 2007; 298: 1209-12.

14. Schmid CH, Lau J, McIntosh MW and Cappelleri JC. An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. *Stat Med*. 1998; 17: 1923-42.
15. Eberly LE. Consequences of event rate heterogeneity across non-randomized study sub-groups. *Stat Med*. 2004; 23: 2023-36.
16. McIntosh MW. The population risk as an explanatory variable in research synthesis of clinical trials. *Stat Med*. 1996; 15: 1713-28.
17. Ioannidis JP and Lau J. Heterogeneity of the baseline risk within patient populations of clinical trials: a proposed evaluation algorithm. *Am J Epidemiol*. 1998; 148: 1117-26.
18. Walter SD. Variation in baseline risk as an explanation of heterogeneity in meta-analysis. *Stat Med*. 1997; 16: 2883-900.
19. Brand R and Kragt H. Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials. *Stat Med*. 1992; 11: 2077-82.
20. Sharp SJ and Thompson SG. Analysing the relationship between treatment effect and underlying risk in meta-analysis: comparison and development of approaches. *Stat Med*. 2000; 19: 3251-74.
21. Kearney MT, Nolan J, Lee AJ, et al. A prognostic index to predict long-term mortality in patients with mild to moderate chronic heart failure stabilised on angiotensin converting enzyme inhibitors. *Eur J Heart Fail*. 2003; 5: 489-97.
22. Levy WC, Mozaffarian D, Linker DT, et al. The Seattle Heart Failure Model: prediction of survival in heart failure. *Circulation*. 2006; 113: 1424-33.
23. Pocock SJ, Wang D, Pfeffer MA, et al. Predictors of mortality and morbidity in patients with chronic heart failure. *Eur Heart J*. 2006; 27: 65-75.
24. Vazquez R, Bayes-Genis A, Cygankiewicz I, et al. The MUSIC Risk score: a simple method for predicting mortality in ambulatory patients with chronic heart failure. *Eur Heart J*. 2009; 30: 1088-96.
25. Fried LP, Kronmal RA, Newman AB, et al. Risk factors for 5-year mortality in older adults: the Cardiovascular Health Study. *Jama*. 1998; 279: 585-92.
26. Lee SJ, Lindquist K, Segal MR and Covinsky KE. Development and validation of a prognostic index for 4-year mortality in older adults. *Jama*. 2006; 295: 801-8.
27. Schonberg MA, Davis RB, McCarthy EP and Marcantonio ER. Index to predict 5-year mortality of community-dwelling adults aged 65 and older using data from the national health interview survey. *J Gen Intern Med*. 2009; 24: 1115-22.
28. Therneau TM and Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. Springer, 2000, p.69, 110-9, 47-51.
29. Altman DG and Andersen PK. A Note on the Uncertainty of a Survival Probability Estimated from Cox's Regression Model. *Biometrika*. 1986; 73: 722-4.
30. Kalbfleisch JD and Prentice RL. *The Statistical Analysis of Failure Time Data.* . 2nd ed. New York, NY: Wiley & Sons, Inc., 2002.
31. Follmann DA and Proschan MA. A multivariate test of interaction for use in clinical trials. *Biometrics*. 1999; 55: 1151-5.

32. Lumley T. Survey analysis in R. 2011, p. This is the homepage for the "survey" package, which provides facilities in R for analyzing data from complex surveys.
33. Wood S. mgcv: GAMs with GCV/AIC/REML smoothness estimation and GAMMs by PQL. 2011, p. Rouutines for GAMs and other generalized ridge regression with multiple smoothing parameter selection by GCV, REML or UBRE/AIC. Also GAMMs by REML or PQL. Includes a gam() function.
34. Stitelman OM, Wester CW, De Gruttola V and van der Laan MJ. Targeted Maximum Likelihood Estimation of Effect Modification Parameters in Survival Analysis. *Int J Biostat*. 2011; 7.
35. Snowden JM, Rose S and Mortimer KM. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. *Am J Epidemiol*. 2011; 173: 731-8.
36. Hayward RA, Kent DM, Vijan S and Hofer TP. Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. *BMC Med Res Methodol*. 2006; 6: 18.
37. Ioannidis JP and Lau J. The impact of high-risk patients on the results of clinical trials. *J Clin Epidemiol*. 1997; 50: 1089-98.

TABLES AND FIGURES.

Table 1. Modeling the Treatment Effect of Enalapril on Mortality or Heart Failure Hospitalization – A Comparison of Approaches.

	Standardization, Covariate-based[a]	**Standardization, Risk-based**	**CRAM, U-adjusted**[a]	**CRAM, U Interaction-adjusted**[a]
Samples Used	Trial, Target (but does not use information on treatment status and outcome from the bridging sample)	Trial, Target (but does not use information on treatment status and outcome from the bridging sample)	Cross-design synthesis: Trial, Bridge, Target	Cross-design synthesis: Trial, Bridge, Target
Heterogeneity of Treatment Effect	Addressed indirectly	Addressed indirectly	Major focus	Major focus
Unmeasured Confounding	Not a concern	Not a concern	Major focus. Assumes unmeasured confounding transportable from bridge to target sample.	Major focus, including the possibility of interaction with an unmeasured confounder. Assumes unmeasured confounding transportable from bridge to target sample.
Overlap	Unstable when there is little overlap between samples.	Unstable when there is little overlap between samples.	Uses observational study sample as bridge to address little overlap between trial and target samples.	Uses observational study sample as bridge to address little overlap between trial and target samples.
Outcome data	Not required	Not required	Required in trial and bridge samples but not in the target sample.	Required in trial and bridge samples but not in the target sample.
Outcome model	Not required	Not required	Required	Required
Cross-design synthesis	Partial	Partial	Full	Full

Other Comments	Does not require a model for the outcome – need sufficient information from the target sample. Assumes given covariates there is no additional confounding.	Does not require a model for the outcome – need sufficient information from the target sample. Assumes given baseline risk there is no additional confounding.	Need outcome in trial and bridge samples but not in the target sample.	

[a] U stands for unmeasured confounding

Other footnotes: All models use the Cox proportional hazards model.

Table 2. Overall Participant Characteristics and Unadjusted Treatment Effects According to Study.

Sample	Evidence Source	Bridge	Target
Study	SOLVD Treatment Trial (*n*= 2,569)	SOLVD Registry (*n*= 5,100)	SOLVD Prevention Trial (*n*= 4,228)
Study Type	RCT	Observational	RCT[a]
Proportion			
Female	19.6	28.8	11.3
Age ≥75 years	5.8	16.1	4.2
Female and age ≥75 years	1.4	8.0	0.6
Smoking status			
Never	23.2	29.1	20.8
Former	54.7	48.0	55.6
Current	22.1	22.9	23.6
History of diabetes mellitus	25.8	24.6	15.3
History of myocardial infarction	65.8	76.0	80.1
History of atrial fibrillation	10.8	15.1	4.3
Dependent edema	16.8	29.0	4.4
Pulmonary edema	25.7	40.6	7.5
Lung crackles	12.1	36.3	2.6
Diuretic use	85.5	8.0	16.7
History of chronic obstructive pulmonary disease	10.0	17.7	5.4
History of stroke	7.7	8.9	5.9
General Health			
Excellent	2.3		5.2
Very good	10.2		19.2
Good	31.0		37.7
Fair	41.9		31.6
Poor	14.7		6.3
Difficulty in any Instrumental Activities of Daily Living	52.6		33.3
Hospitalization within last year	43.4	82.4	41.3
Angiotensin converting enzyme			

inhibitor exposure			
Any ACEi	n/a	37.9	n/a
Enalapril	50.0	n/a	49.9
Mean / (standard deviation)			
Age, years	60.4	62.8	58.7
	(9.9)	(12.2)	(10.3)
Left ventricular ejection fraction, %	24.9	31.9	28.3
	(6.7)	(9.7)	(5.6)
Cardiothoracic ratio	0.6	0.7	0.5
	(0.5)	(0.5)	(0.5)
Systolic blood pressure, mmHg	124.9	128.2	125.4
	(17.5)	(22.0)	(16.4)
Diastolic blood pressure, mmHg	76.8	77.0	78.0
	(10.2)	(13.0)	(9.6)
Resting pulse, beats per minute	79.9	82.1	74.9
	(13.2)	(17.2)	(12.4)
Serum creatinine, mg/dL	1.3	1.3	1.2
	(0.4)	(0.5)	(0.3)
Serum sodium, meq/dL	139.7	138.1	140.2
	(3.2)	(4.2)	(2.9)
Serum hemoglobin, g/dL	14.1		14.3
	(1.6)		(1.4)
Body weight, kg	79.7	78.2	81.2
	(16.6)	(17.7)	(14.3)
Unadjusted treatment effect, log hazard ratio (SE, P-value)	-0.51 (0.080, <0.0001)	0.47 (0.05, <0.0001)	-0.28 (0.10, 0.006)

Abbreviations: ACEi is angiotensin converting enzyme inhibitor; n/a is not applicable; RCT is randomized, controlled trial; SOLVD is Studies of Left Ventricular Dysfunction.

[a] A RCT was chosen here to externally validate the CRAM methodology. In practice treatment will not be randomized in the target sample.

Other footnotes: blanks indicate variable is not available for that study.

Table 3. Treatment Effect of Enalapril on Heart Failure Hospitalization or Death in the Target Sample (SOLVD Prevention Trial).

Model	True Effect	Standardization, Covariate-based[b]	Standardization, Risk-based[c]	CRAM, U-adjusted	CRAM, U Interaction-adjusted
Estimate $\psi_{t=1}$[a]	-0.28 (0.10)	-0.50 (0.22)	-0.31 (0.13)	--	--
μ_1=-0.5		--	--	-0.43 (0.14)[d]	-0.45 (0.15)[e]
μ_1=-1.0		--	--	-0.45 (0.16)[f]	-0.43 (0.13)[g]

Abbreviations: ATE is average treatment effect. CRAM is calibrated, risk-adjusted modeling.

Other footnotes: Numbers in parentheses are standard errors.

[a] Marginal log hazard ratio, Equation 1

[b] Based on boosted logistic regression.

[c] Based on flexible logistic regression with a spline term for baseline risk.

[d] The optimal ρ* was 0.72.

[e] The optimal ρ* was 0.70.

[f] The optimal ρ* was 0.56.

[g] The optimal ρ* was 0.49

Table 4. Treatment Effect of Enalapril on Heart Failure Hospitalization or Death in a Distant Target Sample.

Model	True Effect	Standardization, Covariate-based[b]	Standardization, Risk-based[c]	CRAM, U-adjusted	CRAM, U Interaction-adjusted
Estimate $\psi_{t=1}$[a]	-0.35 (0.19)	-0.55 (0.43)	-0.11 (0.19)	--	--
μ_1=-0.5		--	--	-0.31 (0.18)[d]	-0.31 (0.19)[e]
μ_1=-1.0		--	--	-0.28 (0.18)[f]	-0.33 (0.21)[g]

Abbreviations: ATE is average treatment effect. CRAM is calibrated, risk-adjusted modeling.

Other footnotes: Numbers in parentheses are standard errors.

[a] Marginal log hazard ratio, Equation 1

[b] Based on boosted logistic regression.

[c] Based on flexible logistic regression with a spline term for baseline risk.

[d] The optimal ρ* was 0.72.

[e] The optimal ρ* was 0.70.

[f] The optimal ρ* was 0.56.

[g] The optimal ρ* was 0.49.

Table 5. Treatment Effect of Enalapril on Heart Failure Hospitalization or Death in a Conventional Target Sample (Women Older than 75 Years).

Model	Standardization, Covariate-based[b]	Standardization, Risk-based[c]	CRAM, U-adjusted	CRAM, U Interaction-adjusted
Estimate $\psi_{t=1}$[a]	-0.094 (0.44)	-0.64 (0.13)	--	--
μ_1=-0.5	--	--	-0.43 (0.08)[d]	-0.41 (0.09)[e]
μ_1=-1.0	--	--	-0.44 (0.09)[f]	-0.42 (0.08)[g]

Abbreviations: ATE is average treatment effect. CRAM is calibrated, risk-adjusted modeling.

Other footnotes: Numbers in parentheses are standard errors. True effect could not be computed in the validation trial due to small sample size of women older than 75 years.

[a] Marginal log hazard ratio, Equation 1

[b] Based on boosted logistic regression.

[c] Based on flexible logistic regression with a spline term for baseline risk.

[d] The optimal ρ* was 0.72.

[e] The optimal ρ* was 0.70.

[f] The optimal ρ* was 0.56.

[g] The optimal ρ* was 0.49.

Figure 1. Distribution of 1-yr Risk of Heart Failure Hospitalization or Mortality According to Study.

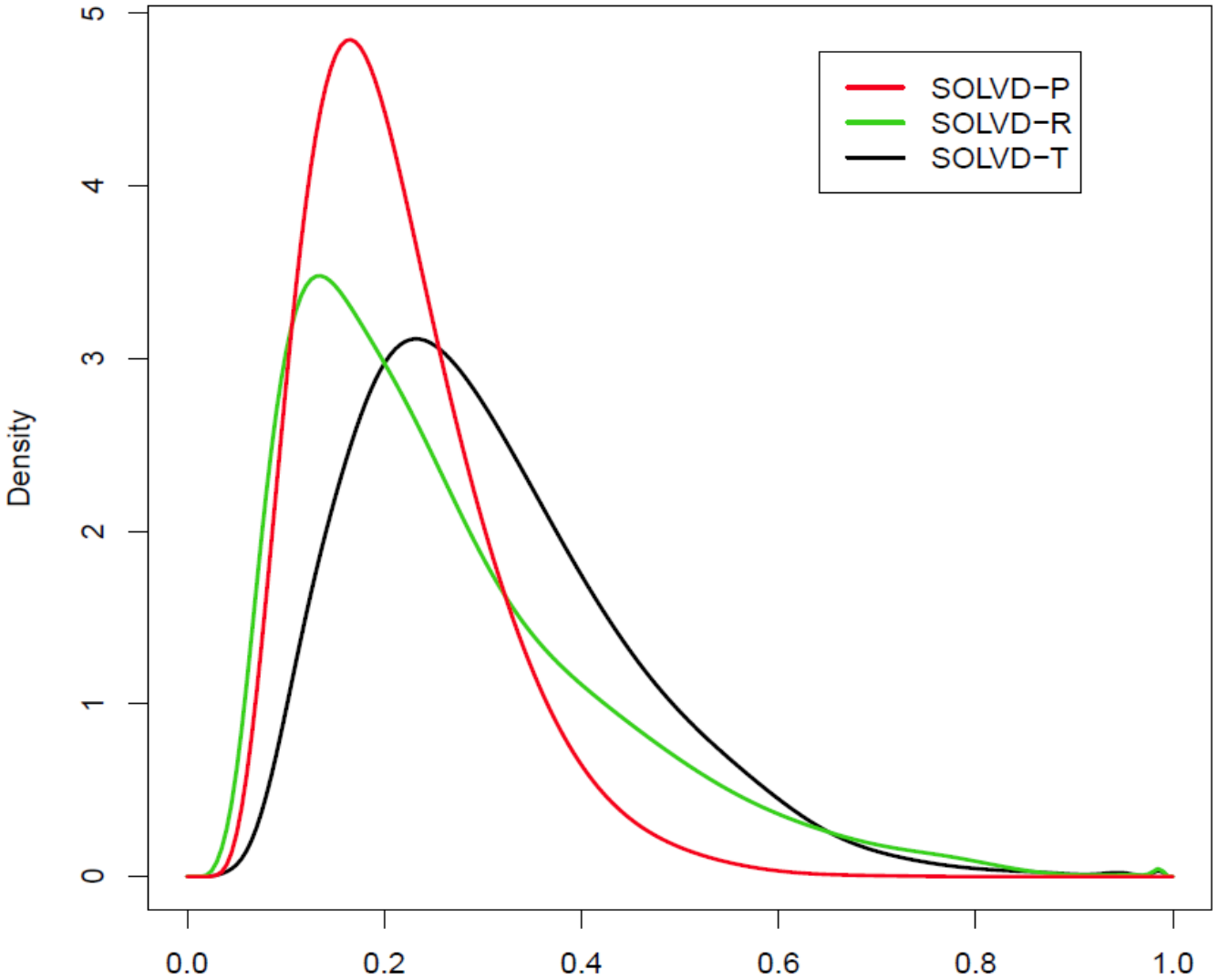

Figure 1 Legend.

Abbreviations: SOLVD is Studies of Left Ventricular Dysfunction. SOLVD-P is the SOLVD Prevention Trial; SOLVD-R is the SOLVD Registry; SOLVD-T is the SOLVD Treatment Trial.

Other footnotes: 1-yr risk of heart failure hospitalization or death is estimated with survival analysis as described in Methods.

Figure 2. Cumulative Distributions of Standardization Weights for Standardizing the Trial and Bridging samples to the Distant Target Sample.

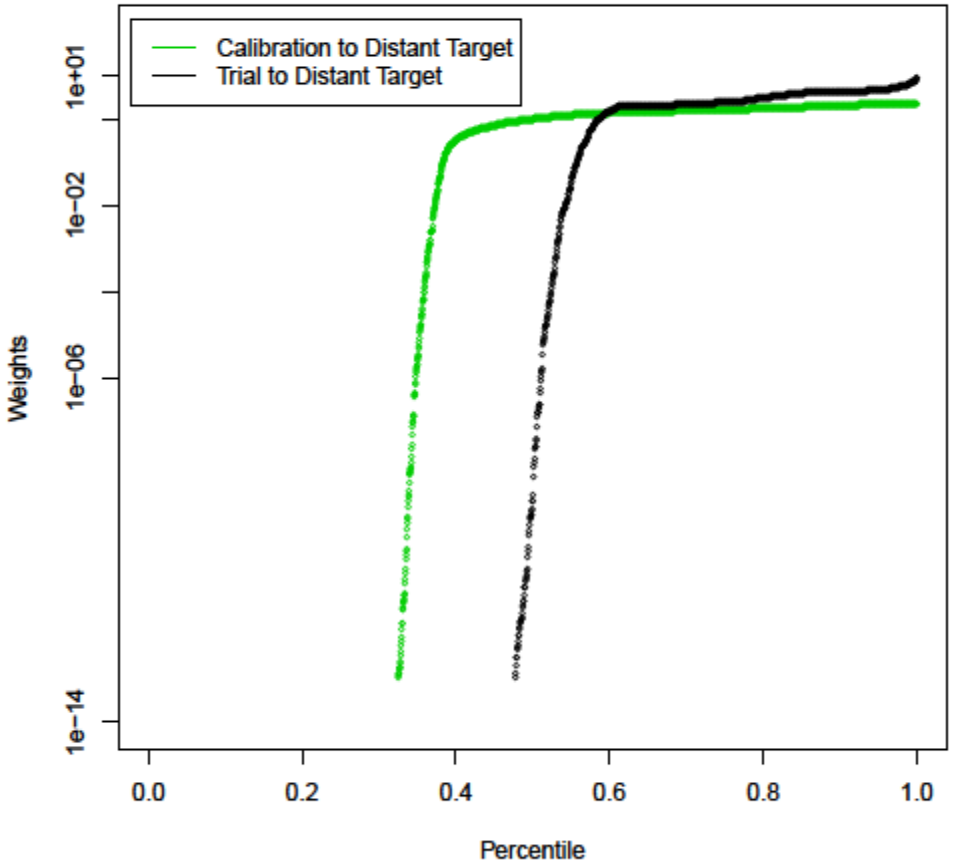

Figure 2 Legend: Green circles are weights for risk-based standardization of the bridging sample to the distant target sample; black circles are weights for risk-based standardization of trial sample to the distant target sample.