Johns Hopkins University, Dept. of Biostatistics Working Papers

11-5-2014

CROSS-DESIGN SYNTHESIS FOR EXTENDING THE APPLICABILITY OF TRIAL EVIDENCE WHEN TREATMENT EFFECT IS HETEROGENEOUS-I. METHODOLOGY

Ravi Varadhan

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

Carlos Weiss

Department of Family Medicine, Michigan State University

Cross-Design Synthesis for Extending the Applicability of Trial Evidence when Treatment Effect is Heterogeneous
I. Methodology

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

ABSTRACT:

Randomized controlled trials (RCTs) provide reliable evidence for approval of new treatments, informing clinical practice, and coverage decisions. The participants in RCTs are often not a representative sample of the larger at-risk population. Hence it is argued that the average treatment effect from the trial is not generalizable to the larger at-risk population. An essential premise of this argument is that there is significant heterogeneity in the treatment effect (HTE). We present a new method to extrapolate the treatment effect from a trial to a target group that is inadequately represented in the trial, when HTE is present. Our method integrates trial and observational data (cross-design synthesis). The target group is assumed to be well-represented in the observational database. An essential component of the methodology is the estimation of calibration adjustments for unmeasured confounding in the observational sample. The estimate of treatment effect, adjusted for unmeasured confounding, is projected onto the target sample using a weighted G-computation approach. We present simulation studies to demonstrate the methodology for estimating the marginal treatment effect in a target sample that differs from the trial sample to varying degrees. In a companion paper, we demonstrate and validate the methodology in a clinical application.

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

I. Introduction

Randomized controlled trials ("RCTs" or "trials") provide the most reliable evidence regarding efficacy of interventions for approval of new treatments, informing clinical practice, and coverage decisions. Although trials provide reliable evidence regarding treatment effects, their evidence is usually narrow in the sense that it comes from participants who are carefully screened and selected. It is generally argued that the participants in trials are a select group that is not a representative sample of the larger at-risk population, and hence that the average treatment effects from the trials are not generalizable to the larger at-risk population. Trials

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

cannot hope to avoid selection entirely. Selection criteria are useful to reduce the number of trial participants who are either too unlikely to benefit or too likely to be harmed to maintain scientific equipoise and ethical acceptability. Further, a trial can experience unplanned selection such as due to healthy volunteerism.[1] However, selection by a trial becomes a source of limited external validity when it occurs according to treatment effect modifiers causing significant heterogeneity in the treatment effects (HTE), since if there were no HTE the evidence should be generalizable.

The limited external validity of trial evidence is a cause for concern to the extent that two conditions are present. The first is selection in trials or non-random sampling. The second condition is heterogeneity of treatment effect (HTE) or explanatory variability of treatment effect across individuals. HTE occurs when a variable that substantially modifies treatment effect is understood.[2] The presence of selection by a trial is not by itself a threat to external validity in the absence of known HTE. (Admittedly, selection may be some cause for concern even in the absence of HTE because it suggests that there is an increased likelihood of selection according to an unknown source of HTE.) The trial evidence becomes less relevant to a specific target population when selection and HTE are both present.[3]

Mainly due to a lack of external validity, some guideline-makers and policy-makers believe that the evidence base from trials is inadequate, i.e. that trials provide insufficient evidence to guide care for key target populations.[4,5] In recent years many have called for trials that are less exclusionary and more pragmatic.[6,7] In addition, interest has gathered around exploiting observational studies' potential to provide evidence where trials cannot. However, a fundamental limitation of observational studies is that they do not control for confounding through randomization. We could use a propensity-score based approach to estimate the treatment effect, but this is likely to be biased due to unmeasured confounding. Therefore, limited internal validity is inherent to an observational design.

There have been a few past attempts to draw information from more than one study design. Cross-design synthesis attempts to draw on respective strengths by examination of trial and observational data side by side.[8] However, no statistical methods were proposed to estimate treatment effects in the observational study. The confidence profile approach uses a Bayesian framework to combine evidence, with adjusted likelihood functions for different study designs and biases.[9] However, confidence profile, like meta-analysis, uses only data at the study level and therefore cannot rigorously account for HTE.[2] A standardization approach proposed by

Cole and Stuart[10] uses the trial treatment effect and minimum sufficient information about the joint distribution of treatment effect modifiers from the observational target sample. It can be used even without individual-level data from the target sample if sufficient information about the joint distribution of treatment effect modifiers is available. With this approach, the probability of being selected into the trial sample can be used to provide a valid estimate of the marginal treatment effect in the target sample. The Cole and Stuart approach for estimating a marginal treatment effect appears to work well in scenarios where there is sufficient overlap between the trial and target sample. Without sufficient overlap, many people within the trial will be assigned zero weights and standardization will become unstable.

We present a novel method to project (extrapolate) the treatment effect from a trial to a target group that is poorly represented in the trial, thus extending the applicability of the trial evidence. Our method uses individual-level data from a trial and an observational study that is more representative than the trial with respect to the target population. The method is based on the principle that valid (i.e., unbiased) estimates of treatment effect are most reliably obtained in trials, but observational designs are attractive for assessing and extending applicability of evidence to a different target sample. The proposed method may be viewed as an advance in cross-design synthesis, where the main idea is to integrate two different study designs in such a manner as to exploit their unique strengths while avoiding their weaknesses. We call this cross-design synthetic approach CRAM (Calibrated Risk-Adjusted Modeling).

We present two simulation examples to demonstrate the methodology. In a companion paper, we demonstrate the application of the CRAM methodology to a clinical problem and validate the results using another trial.

2. Brief description of CRAM

Let A denote a binary treatment that is randomized in a clinical trial, sample E. Let $\beta_A(E)$ be the estimate of efficacy of intervention in E (i.e., the intent-to-treat effect). We would like to assess whether the evidence from source E is applicable to target sample T. The evidence would be applicable if the sample E is exchangeable with T, in the sense that it is reasonable to conceive of E as a random sample of T. In particular, T should have the same distribution of the treatment effect modifiers as E in order for the evidence from E to be applicable to T (Cole and Stuart 2010). How can we apply evidence from E to T when there is lack of exchangeability between E and T with respect to treatment effect modifiers? Cole and Stuart (2010) proposed

standardization on the basis of an effect modifier for establishing exchangeability between E and T when T is the entire at-risk population rather than a specific target population. In other words, they considered the *generalizability* of evidence in situations where E is a subsample of T, albeit a non-random subsample. Direct standardization of E to T might work well when there is sufficient overlap in the distribution of the effect modifier in E and T.

We consider a more relevant situation in which the source of evidence E is not necessarily a subsample of the target population T. In particular, we are interested in the applicability of the trial evidence either to a target sample that was excluded or under-sampled (e.g. older women). Here the direct standardization approach of Cole and Stuart would not work well because without sufficient overlap standardization will be unstable. Therefore, we propose a cross-design synthetic approach using an observational bridging sample B where both E and T are well represented in terms of treatment effect modifiers (see Figure 1). First, we estimate an intent-to-treat effect in E that is standardized to the distribution of the treatment response score (this is an “optimal” linear combination of treatment effect modifiers) in B. This may be viewed as the causal treatment effect for B. We then estimate the treatment effect in B using models which incorporate unmeasured confounding. We calibrate the parameters of unmeasured confounding such that the calibrated treatment effect in the observational sample is the same as the risk-standardized treatment effect estimated in the trial. The model-based calibration makes it possible to characterize unmeasured confounding. Finally, we estimate the treatment effect for the target sample that is calibrated for unmeasured confounding.

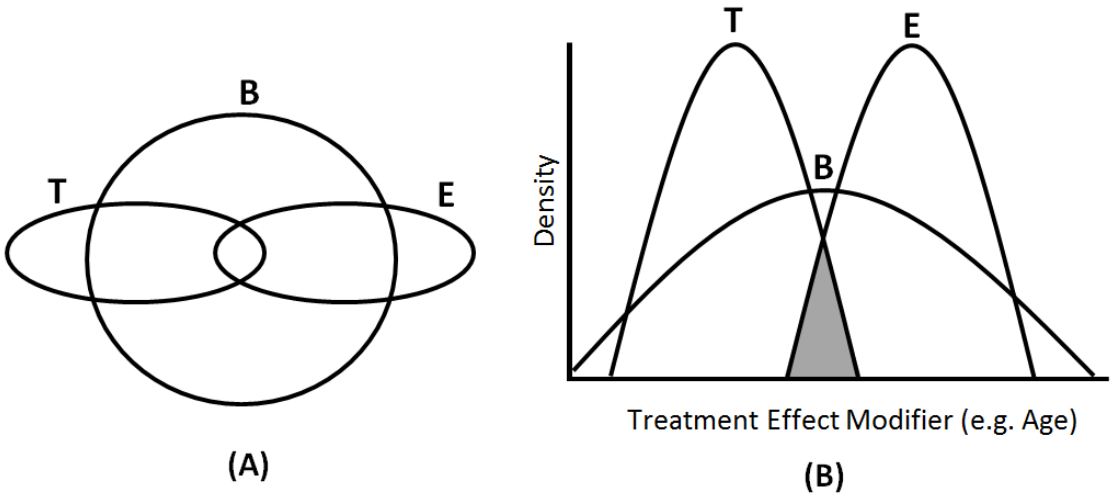


Figure 1. A schematic of a cross-design synthesis approach, calibrated risk-adjusted modeling (CRAM). We have: E = trial (evidence source sample), T = target group and B = (bridge sample). 1A is an abstract conceptualization of the overlap between the three different samples, and 1B is a concrete representation of the overlap in terms of the distribution of the treatment effect modifier (e.g., baseline risk of outcome, age). According to this particular schema, there is little overlap between the source (E) and the target (T) samples, therefore the direct standardization approach of Cole and Stuart will not be appropriate. The observation sample (B) has a substantial overlap with both E and T in terms of the treatment effect modifier. The CRAM approach utilizes B as the "bridge" that provides the evidentiary link between E and T.

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 likely to be present is an essential aspect of our methodology. Various patient characteristics can modify the effect of the treatment. There is a large literature demonstrating that baseline risk of the outcome (i.e. the probability of outcome that is independent of treatment) is often a potent treatment effect modifier on the relative risk scale.[11-18] We propose to use baseline risk of the outcome as a treatment effect modifier to model heterogeneity of treatment effects. The baseline risk of outcome is generally unknown.

We have to estimate it either using a validated external model or using a regression model developed with the data at hand. If the risk model is external, it is important to ensure that it is well calibrated for the data at hand. When a reliable external risk model is not available, it must be developed internally. This can be developed in the control arm of the trial and observational samples. Depending on the type of outcome (e.g., continuous, binary, counts, time-to-event), the risk model might be linear, generalized linear, or Cox models. We will denote r_i the predicted baseline risk of outcome for individual i with covariates X_i.

It has been shown that the baseline outcome risk is more powerful than interaction tests using one variable at a time, and that it is not susceptible to multiple testing concerns. [2] We suggest two approaches for testing whether the treatment effect varies according to baseline outcome risk. An interaction term between the treatment and the baseline risk score can be included in a regression model of outcome variable that also has first-order terms for treatment and baseline risk. HTE is present if the interaction term is statistically (e.g., Wald test) and clinically significant.

Another interesting approach has been proposed by Follmann and Proschan.[25] In this approach we do not estimate the risk and then test for interaction separately, but we do it all in one step. This is particularly useful when a validated external risk model is unavailable.

The data is $\{Y_i, A_i, X_i\}$, where Y_i is the primary endpoint, A_i the treatment indicator, and X_i a p-dimensional vector of baseline predictors of outcome. Consider the model:

$$g(E[Y_i]) = (1 - A_i)(\alpha_0 + \boldsymbol{X'}_i\boldsymbol{\beta\gamma_0}) + A_i(\alpha_1 + \boldsymbol{X'}_i\boldsymbol{\beta\gamma_1}) \qquad (1)$$

Test for interaction now is a test of H_0: $\gamma_0 = \gamma_1$ versus H_A: $\gamma_0 \neq \gamma_1$. By substituting $\boldsymbol{\beta\gamma_j} = \boldsymbol{\delta_j}$, the null and alternate hypotheses, H_0 and H_A, can be rewritten as:

H_0: $\boldsymbol{\delta}_0 = \boldsymbol{\delta}_1$ versus H_A: $k\,\boldsymbol{\delta}_0 \neq \boldsymbol{\delta}_1$, $k \neq 1$.

A likelihood ratio test for this can be constructed by estimating a common $\boldsymbol{\delta}$ under H_0 and $\boldsymbol{\delta}$ and k under H_A. The test both identifies a single index of risk and also tests for an interaction along this index. If the test rejects the null hypothesis, the conclusions are that there is a treatment interaction and that the treatment effect varies along $\boldsymbol{X'\beta}$, which Follmann and Proschan call "disease severity", and may be viewed as being analogous to the baseline risk. However, it is

more accurate to think of this as the optimal linear combination of covariates that predicts individuals' response to the treatment. Hence, we call this treatment response score.

4. Current approach: direct standardization of treatment effect from the trial to target sample

There are two approaches for projecting (or extrapolating) the intent-to-treat effect from the source sample E (i.e. trial) to the target group P. Standardization (Cole and Stuart 2010) is the direct approach. Standardization is a commonly used technique to estimate the (causal) parameter of interest in a population of interest. It is a traditional approach in survey sampling and epidemiology to extrapolate estimates from the sample at hand to a target population. Standardization involves appropriate adjustment of the estimate from available data such that the adjusted estimate reflects the differences in characteristics between the sample at hand and the target population of interest, especially those characteristics that are thought to be confounders or effect modifiers. Stratification is used when there is small number of characteristics or when there is large data. Model based approach is warranted when there are several variables. Cole and Stuart[10] proposed a model based approach for standardizing trial results to a target population. Here we briefly review their approach.

4.1 Covariate-based standardization

Let S = 1 denote membership in study sample at hand. Obviously, S=0 denotes those in the target sample not being selected into the study. For subject *i* with covariates X_i, $p_i=\Pr(S_i=1 \mid X_i)$ denotes the probability of being selected into the study.

$$\log \frac{\Pr(S_i = 1 \mid X_i)}{1 - \Pr(S_i = 1 \mid X_i)} = X_i\theta \qquad (2)$$

Cole and Stuart fitted a weighted Cox proportional model to estimate the treatment effect, where the weight assigned to each individual was the inverse probability of being selected into the study, $1/p_i$. Individuals who according to the selection model are least likely to get into the study, but actually got in, would be assigned large weights. The treatment effect from the weighted Cox model is standardized to the target population. Robust variance estimation procedure should be used to compute standard error of treatment effect estimate since weighting induces correlation in the data.

4.2 Risk-based standardization

We follow the standardization approach of Cole and Stuart. However, we depart from their proposal in that rather than use covariate-based standardization we use the baseline risk, r_i, as the standardization variable (treatment response score can also be used here). We estimate study membership S_i=1 using a flexible logistic regression model:

$$\log \frac{\Pr(S_i=1 \mid r_i)}{1-\Pr(S_i=1 \mid r_i)} = f(r_i) \qquad (3)$$

where $f(r_i)$ is a flexible cubic spline function. The model is fitted using the generalized additive model approach implemented in the R package "mgcv".[19] Let $p_i = \Pr(S_i=1)$ for individual *i*. The weights are $w_i = 1/p_i$. We estimate standardized treatment effect using a weighted outcome model (a marginal structural model): weighted linear, logistic or Cox regression model depending on the outcome. Robust variance estimation procedure is used to compute standard error of treatment effect estimate.

5. CRAM

The direct standardization approach will not work well in situations where the target group is either excluded or under-sampled in E (e.g. older women). Without sufficient overlap, many people in the trial will be assigned zero weights and standardization will be unstable. Hence, we propose a new cross-design synthetic approach (CRAM), where we make use of both the source sample E and an observational bridge sample B to obtain the treatment effect in the target group. There are three main steps involved in CRAM: (i) definition of the causal estimand in observational sample, (ii) calibrating observational sample for unmeasured confounding, and (iii) estimating treatment effect for target sample.

5.1 Causal estimand in observational bridge sample

We employ the G-computation approach in the observational bridge sample to estimate the causal effect of the treatment. This approach allows the use of observational data for the estimation of parameters that would be obtained in a randomized controlled trial. Under assumptions of consistency, no unmeasured confounding and correct model specification for the potential outcomes, these estimates can be interpreted causally.[20] Suppose the data is $\{Y_i, A_i, X_i\}$, where Y_i is the primary endpoint, A_i the treatment indicator, and X_i a p-dimensional vector of baseline predictors of outcome. We let Y_i be a binary outcome for describing the G-computation. The first step is to fit a regression model for the outcome Y on the exposure A and covariates X. The outcome model can include interactions between exposure and covariates. After fitting this model, we predict the counterfactual outcome probabilities for each individual

under each exposure, $p_{i0}=Pr(Y_i=1|A_i=0,X_i))$ and $p_{i1}=Pr(Y_i=1|A_i=1,X_i))$. The covariates are fixed at the observed values, but only the exposure status is changed. The marginal treatment effect estimate (log odds ratio) is computed as:

$$\theta = \log\left[(\tfrac{1}{n})\sum_{i=1}^{n} p_{i1}\right] - \log\left[(\tfrac{1}{n})\sum_{i=1}^{n} p_{i0}\right] \quad (5)$$

Resampling based methods, such as bootstrapping, may be used to estimate the standard error of θ.

For survival (i.e. time to event) data, the causal effect of interest is the marginal log hazard ratio, given as:

$$\psi_t = \log[-\log(\text{prob}(T_1 > t)] - \log[-\log(\text{prob}(T_0 > t)] , \quad (6)$$

where T_0 and T_1 denote the potential failure times under unexposed and exposed conditions. The Cox proportional hazards model is commonly used to estimate the marginal log-hazard ratio. The survival probability, $\text{prob}(T_a > t)$, is estimated after fitting a Cox PH model as follows:

$$\text{prob}(T_a > t) = \frac{1}{n}\sum_{i=1}^{n} S_i(t|A=a, X_i), \quad (7)$$

where $S_i(t|A=a)$ is the counterfactual survival probability at time *t* for individual *i* when he/she is given treatment A=a. The G-computation formulas (6) and (7) are given in Stitelman et al.[21]

The outcome model must be correctly specified in order for the G-computation estimate of the marginal treatment effect to be unbiased. This is a major difference between the standardization approach and the G-computation approach. The standardization approach using study selection weights does not require an outcome model, whereas the G-computation approach is critically dependent on the outcome model. On the other hand, the standardization approach requires that the study selection model be correctly specified. In addition to correct outcome model specification, the G-computation approach also requires experimental treatment assignment or positivity assumption, which means that each individual should have a non-zero probability of receiving the intervention. It also requires no unmeasured confounding, which is unlikely to hold in observational studies.

5.2. Calibration of observational bridge sample to trial

We assume that all unmeasured confounding can be captured by a single covariate U, i.e. exposure A is independent of potential outcomes {Y(A=0), Y(A=1)} given X and U. This assumption is not as unrealistic as it might seem. It can be actually proved (see Appendix). If there is knowledge about how the unmeasured confounder U is related to exposure A, outcome

Y and to covariates X, it can be used to generate realizations of U. In the absence of such knowledge, it is necessary to make strong (unverifiable) assumptions about U and sensitivity analyses are warranted. For continuous and binary outcomes Y, we take U to be normally distributed with a mean that depends on Y and A as follows:

$$E[U|Y=y, A=a] = b_0 + b_a a + b_y y; \quad var[U|Y=y, A=a] = \sigma_U^2 \qquad (8)$$

where b_a represents the impact of U on exposure status and b_y represents the impact of U on the outcome. A positive b_a implies that people with higher values of U are more likely to be exposed, and a positive b_y implies that people with higher values of U are more likely to experience the outcome.

When the outcome is the time to an event subject to right censoring, we cannot use (8) since we cannot observe the event times for all individuals due to right censoring. Let $\{Y_i, E_i\}$ be the observed data for each individual, where Y_i is the duration of follow-up and E_i is an indicator of whether the individual had the event, with $E_i=1$ indicating that the event happened. We do not know the true event times for those with $E_i=0$. We need to estimate the true failure time T given {Y, E}. We assume that censoring is uninformative and that T is log-normally distributed. Under these assumptions, the mean and variance of log T can be readily estimated from {Y, E} by fitting a censored regression model (with `survreg' function in "survival" package,[22] with a log-normal distribution). Let the estimated marginal mean and standard deviation of log T, estimated using {Y, E}, be μ_{LT} and σ_{LT}, respectively The assumption of completely uninformative censoring can be relaxed by using covariate information to model {Y, E}. We, then, generate U correlated with estimated failure time as follows.

We assume that (log U, log T) is from a bivariate normal distribution. The marginal mean of log U is μ_U and it varies according to the exposure status A as:

$$\mu_U = A*\mu_1 + (1-A)*\mu_0, \qquad (9)$$

where μ_0 and μ_1 are the means of U in the unexposed and exposed groups. Let σ_U be the standard deviation of log U, and ρ be the correlation between log U and log T. Both σ_U and ρ are assumed constant, i.e. independent of treatment group A. Now, the conditional distribution of U given T is also log-normal with mean and standard deviation given as follows:

$$\mu_{U|T} = \mu_U + Z\,\rho\,\sigma_U, \text{ where } Z = (\log Y - \mu_{LT})/\sigma_{LT} \qquad (10a)$$

$$\sigma_{U|T} = \sigma_U\,(1-\rho^2)^{1/2} \qquad (10b)$$

This completes the specification of the model for generating unmeasured confounder U that is correlated with right-censored failure time Y. Note that 4 input parameters are necessary to characterize U: μ_0, μ_1, ρ, and σ_U. We can fix μ_0 and σ_U without losing much generality, and only vary μ_1 and ρ to study the impact of unmeasured confounding. A positive μ_1 implies that people with higher values of U are more likely to be exposed, and a positive ρ implies that people with higher values of U are more likely to have a larger time to event, or alternately, positive ρ implies that higher values of U are associated with smaller rates of events.

The expected treatment effect from the trial and observational study should be the same when the two samples have the same distribution of confounders and when there is no unmeasured confounding. However, they might differ when the distributions of confounders differ. Therefore, we need to standardize the ITT effect from the trial to an observational sample. We let S=1 denote membership in observational sample and S=0 membership in the trial sample. We estimate $p_i = \Pr(S_i=1|r_i)$ for each individual *i* using (3). The weight assigned to each individual in the trial is the odds of that individual being selected into the observational sample:

$$w_i = p_i/(1-p_i), \text{ where } p_i = \Pr(S_i=1|r_i) \quad (4)$$

The individuals in observational sample receive zero weight. We then estimate standardized treatment effect using a weighted outcome model (a marginal structural model): weighted linear, logistic or Cox regression model depending on the outcome. Let β_A be the (intent-to-treat) treatment effect in the trial that is standardized to the treatment response score distribution in the observational study. In order for the standardization to provide reliable effect estimate, the baseline risk distributions in the trial and observational samples should have substantial overlap. One way to quantify the lack of overlap is to look at the proportion of subjects in the trial with relatively small weights, Eq. (4).

We can compute the corresponding marginal causal effect in the observational study using the G-computation approach described in Section II.E. For a binary outcome, we consider the outcome model:

$$\log\frac{\Pr(Y_i=1\,|A,R_i)}{1-\Pr(Y_i=1\,|A,R_i)} = \alpha_0 + \alpha_a A + \alpha_r R + \alpha_{ar} A * R \quad (11)$$

where A is binary exposure status and R is the baseline risk of outcome. Let the G-computation estimate of marginal log odds ratio, Eq. (5), be α_A. This would be the causal effect, and would equal, in expectation, the risk-standardized trial effect, β_A, when there is no unmeasured

confounding (and when the additional assumptions of consistency and positivity of treatment assignment hold). When these 2 effect estimates are different, we might suspect that unmeasured confounding is present. We might consider the following outcome model:

$$\log \frac{\Pr(Y_i=1 \mid A,R_i,U)}{1-\Pr(Y_i=1 \mid A,R_i,U)} = \alpha_0 + \alpha_a A + \alpha_r R + \alpha_{ar} A*R + \alpha_U U + \alpha_{au} A*U \qquad (12)$$

Let the corresponding G-computation estimate of marginal log odds ratio (Eq. (5)) be $\gamma_A(U)$. We use this notation to denote the dependence of the G-comp estimate on the parameters of U. We estimate unmeasured confounding by finding the parameters of U (Eqs. 10a and 10b) that optimally matches the observational sample's treatment effect to the trial treatment effect. In other words, we find b_a and b_y such that $\gamma_A(U) = \beta_A$. Even though the G-comp estimate and the standardized trial effect only have to be equal in expectation, we enforce this equality for the sample at hand.

Without loss of much generality, we may fix the following values of parameters defining U (see Eq (8)): $b_0 = 0$, $\sigma_U = 1$. The remaining two parameters b_a and b_y fully characterize the nature and degree of unmeasured confounding. There is no unmeasured confounding if either one of them is zero (or very small). A positive b_a indicates that U is positively correlated with treatment status, i.e. an increase in U increases the probability of choosing the treatment. Similarly, a positive b_y indicates that U is positively correlated with the probability of outcome. We vary $b_a < 0$ and estimate b_y such that $\gamma_A(U) = \beta_A$. We do not lose any generality by considering a negative b_a. It should be noted that a smaller magnitude of difference in U between the two treatment groups, i.e. a less negative b_a, would need to be compensated by a larger positive b_y in order to obtain the same target treatment effect; $b_y \cong 0$ would mean that there is no unmeasured confounding.

The algorithm for CRAM calibration is as follows:

1. Obtain the standardized trial treatment effect β_A to be matched.
2. Fix $b_0 = 0$, $\sigma_U = 1$ and pick a value of $b_a < 0$.
3. For some b_y, generate K realizations of U from a normal distribution with the mean and standard deviation given by (8). Each U is a vector of length N (sample size of observational data), one for each individual. This gives us K augmented observational data sets: $D_k = \{A, Y, X, U_k\}$, k=1, 2, …, K.

4. For each data set D_k, estimate the marginal log-odds ratio θ, Eq (5), corresponding to model (12).
5. Compute $\theta^*(U) = E[\theta|U] = \text{mean}(\theta^{(k)})$
6. Determine the discrepancy between estimate and target: $\Delta(U) = (\theta^*(U) - \beta_A)^2$
7. Find b_y^* that minimizes $\Delta(U)$.

The parameters (b_a=0, σ_U = 1, b_a, b_y^*) complete the description of optimal U required for calibration to eliminate any unmeasured confounding. The optimization in Step 7 is a scalar optimization involving only one parameter b_y. However, the objective function in Step 7 is noisy due to randomness of U_k. A large K is required to get good results. K = 200 is usually adequate. A sensitivity analysis for U involves repeating the CRAM algorithm for different values of b_a and estimating the optimal b_y^* for each, under model (12).

5.3. Projecting treatment effect from the bridge sample to the target sample

The last step is to project the treatment effect from the bridge sample, which is the observational study sample, to a target sample. Model (11) cannot be used directly since the participants' risk distribution is different in the target compared to the bridge sample (see Figure 1). However, the calibrated CRAM model (12) with optimal parameters (b_a=0, σ_U = 1, b_a, b_y^*) can predict treatment effect for the target sample, accounting for HTE. The G-computation formula can be applied to the calibrated model (11) with standardization weights to compute the counterfactual probabilities for each individual in the bridge sample. Formula (4) with weights is used to compute the counterfactual probabilities required for calculating the causal effect, (3):

$$\eta = \log\left[\left(\frac{1}{n}\right)\sum_{i=1}^{n} w_i p_{i1}\right] - \log\left[\left(\frac{1}{n}\right)\sum_{i=1}^{n} w_i p_{i0}\right] \qquad (13)$$

where $w_i = p_i/(1-p_i)$, and p_i is the probability that individual *i* in the bridge sample could belong to the target sample. A logistic regression model is fitted for predicting p_i = prob(S_i=1) as a function of baseline risk of outcome, where S=1 for target sample and S=0 for the observational sample. The marginal log-odds ratio (13) is the CRAM estimate of causal effect for the target sample.

6. Simulation studies

There are three different samples: the trial, the observational, and the target. Our goal is to estimate the treatment effect in the target group using two approaches: direct risk-based standardization (section II.D) and CRAM (section II.H).

The sample size was $N_0 = N_1 = 1000$ for the treated and untreated groups in trial and observational sample, and $N_t = 1000$ in the target sample. Binary outcomes were generated according to the following binomial model:

$$\text{logit}(pr[Y=1 \mid A, X_1, X_2, U]) = b_0 + b_X A + b_1 X_1 + b_2 X_2 + b_{A1} A * X_1 + b_{A2} A * X_2 + b_U U + b_{AU} A*U \quad (14)$$

where A = treatment indicator (0/1), X_1, X_2 = covariates, U = unobserved effect modifier. The outcome probability is independent of the study indicator meaning that the same outcome model applies to all samples. We sampled U and X_1 from a bivariate normal distribution: $(X_1, U) \sim N(\mu, \Sigma)$, where $\mu_0 = (1.0, 1.0)$ in the trial, $\mu_1 = (1.0, 0.5)$ in observational sample, and $\mu_t = (0, 0.5)$ in target sample. The covariance matrices for the trial (Σ_0), observational (Σ_1), and target (Σ_t) samples are:

$$\Sigma_0 = \Sigma_t = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}; \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1.5 \end{pmatrix}$$

X_2 is a binary covariate with prob(X_2=1) = 0.5 in trial, 0.8 in observational, and 0.2 in the target sample.

In the trial, the treatment A was randomly assigned with a probability of 0.5 to each participant. In the observational sample, the following assignment rule was used:

$$\text{logit}(pr[A=1 \mid X_1, X_2, U]) = -0.5 + 0.5*X_1 + 0.4 * X_2 + 0.4 * U \quad (15)$$

The outcome model parameters for (14) are: $(b_0, b_A, b_1, b_2, b_U, b_{A1}, b_{A2}, b_{AU}) = (-0.5, 0.7, 0.4, 0.4, 0.2, 0.2, 0.2, 0.1)$. The parameters were chosen such that treatment effect varies with baseline risk of outcome (independent of treatment). Specifically, treatment effect increases with increasing baseline risk of outcome.

Now, by design, the participants in trial and target samples differ in terms of their observed (X_1, X_2) and unobserved (U) covariates. Since these are also effect modifiers, the treatment effect obtained from the trial is not valid for the target sample.

We first estimate the outcome risk, which is required for both CRAM and direct standardization approach. We use the untreated subjects (A=0) from the trial and observational samples to model the outcome risk as a function of observed covariates:

$$\text{logit}(pr[Y=1 \mid A=0, X_1, X_2, S]) = a_0 + a_1 X_1 + a_2 X_2 + a_{12} X_1*X_2 + a_S S$$

We estimate outcome risk for the trial participants as $r = r(x_1, x_2, S=0) = 1 / (1 + \exp(-a_0 - a_1 x_1 - a_2 x_2 - a_{12} x_1 * x_2))$. This model is also used to estimate the baseline risk in the target sample.

[Figure 2 about here]

Figure 1 shows the baseline risk distributions for the trial, observational, and the target sample for a typical simulation. The baseline risk in the trial is on average larger than that in the target sample, while the calibration (observational) sample has a broader risk distribution that spans both the trial and target samples. This is an essential requirement for the CRAM methodology that allows the observational sample to act as the bridging sample between the trial and target samples. In these simulations, there is a great deal of overlap between the trial and target samples in terms of the baseline risk distribution. Therefore, we might expect that the direct risk-based standardization of trial result to target sample would perform as well as CRAM.

We ran 100 simulations. In each, we computed the following treatment effects:

1. The treatment effect in the trial, which we call the naïve treatment effect, because this is what would be commonly used as the estimate for the target sample. This is valid only when either the distributions of effect modifiers are the same in the trial and target, or when there is no effect modification.
2. Risk-standardized treatment effect in the trial as described in section II.D.
3. CRAM projected estimate as described in sections II.G and II.H.
4. The true treatment effect in the target sample under random treatment assignment.

[Table 1 about here]

The results are displayed in Table 1. All of the methods overestimate the truth, with the naïve treatment effect being the worst. The CRAM estimate has a greater bias than the risk-standardization estimate, but has a smaller variance. Both CRAM and risk-standardized estimates have larger standard errors than the naïve estimate.

We ran another set of simulations, where we considered a target sample that is more distant from the trial sample. This is an important scenario because the standardization approach proposed by Cole and Stuart can be expected to be unstable when there is little overlap between the trial and target sample (a large proportion of trial participant would receive zero weights). In contrast, CRAM is designed to address the lack of overlap with the bridge sample as an evidentiary bridge. Specifically, in the second scenario the target sample had a

substantially lower baseline risk than the trial sample. We sampled U and X_1 from a bivariate normal distribution: $(X_1, U) \sim N(\mu, \Sigma)$, where $\mu_0 = (1.0, 1.0)$ in the trial, $\mu_1 = (1.0, 0.5)$ in observational sample, and $\mu_t = (0.05, -0.05)$ in target sample. The covariance matrices for the trial (Σ_0), observational (Σ_1), and target (Σ_t) samples are:

$$\Sigma_0 = \begin{pmatrix} 0.1 & 0 \\ 0 & 0.1 \end{pmatrix}; \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1.5 \end{pmatrix}; \Sigma_t = \begin{pmatrix} 0.05 & 0 \\ 0 & 0.05 \end{pmatrix}$$

X_2 is a binary covariate with prob(X_2=1) = 0.5 in trial, 0.8 in observational, and 0.05 in the target sample. All other parameters are the same as in the first set of simulations.

[Figure 3 about here]

Figure 2 shows the baseline risk distributions for the trial, observational, and the target sample for a typical simulation. The baseline risk in the trial is on average much larger than that in the target sample compared to the first set of simulations. The calibration (observational) sample once again has a broader risk distribution that spans both the trial and target samples.

[Table 2 about here]

The results are displayed in Table 2. The true effect is smaller compared to the first scenario. This makes sense since the target sample now has a smaller baseline risk on average compared to previous scenario. We also see that both the CRAM and risk-standardized estimates are smaller compared to the previous scenario. Both estimates are still positively biased as in the first scenario. An important, but expected, observation is the dramatically increased variance of the risk-standardized estimate. This is because there is much less overlap between the trial and target sample in the second scenario compared to the first scenario. The variance of the CRAM estimate also increased due to the same reason, but the estimate is much less variable compared to the risk-standardized estimate. Thus, the CRAM approach, although biased, is more reliable than risk-standardization for distant target samples as long as there is substantial overlap between the observational bridge sample and the target sample.

7. 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 of a trial and an observational study. A calibration factor for unmeasured confounding for the observational study relative to the trial makes it

possible to estimate a treatment effect in the observational data and in the target sample. 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 rigorously understand and account for HTE and biases. The ability of CRAM to remove unmeasured confounding from the registry data in the calibration interval and provide a sensible estimate of treatment effects in the projection region is promising.

The CRAM method makes several assumptions and has certain requirements. 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 we have presented a method to mainly account 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 calibration in CRAM it is necessary to have sufficient overlap in baseline risk distribution between a trial and observational study. To obtain CRAM projection for the target sample, there must be sufficient overlap between the observational sample and the target sample. Thus, the observational sample plays a critical role as a “bridge” between the trial and the target samples, facilitating the extension of trial results to the target sample. If the trial sample overlaps substantially with the target sample, a direct standardization approach might be adequate and CRAM might not be warranted. However, for target groups that are usually poorly represented in trials, including older women and older adults with multimorbidities, direct standardization might not provide reliable effect estimates. If we can find a “compatible” bridge sample, i.e. an observational study, we might be able to employ the CRAM methodology. By compatibility, we mean that there are no major discrepancies between studies in terms of measurement of treatment, outcome and prognostic variables. 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.

CRAM is computationally intensive in large data sets, mainly due to the calibration step. Therefore, computing standard errors via resampling can be time consuming. Finally, 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.

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 data in order to apply treatment effects from a trial to people not adequately represented in the trial. In a companion paper, we provide an external validation in the SOLVD prevention trial showing that the CRAM projected treatment effect is proximal to the actual treatment effect.

REFERENCES:

1. Pinsky PF, Miller A, Kramer BS, et al. Evidence of a healthy volunteer effect in the prostate, lung, colorectal, and ovarian cancer screening trial. *Am J Epidemiol* 2007;165:874-881.

2. Weiss CO, Segal JB, Boyd CM, Wu A, Varadhan R. *A Framework to Identify and Address Heterogeneity of Treatment Effect in Comparative Effectiveness Research.* Rockville, MD2010.

3. Longford NT. Selection bias and treatment heterogeneity in clinical trials. *Stat Med* 1999;18:1467-1474.

4. 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.

5. Sox HC, Greenfield S. Comparative effectiveness research: a report from the Institute of Medicine. *Ann Intern Med* 2009;151:203-205.

6. Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. *J*AMA 2003;290:1624-1632.

7. Zwarenstein M, Treweek S, Gagnier JJ, et al. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. *BMJ* 2008;337:a2390.

8. Droitcour J, Silberman G, Chelimsky E. A new form of meta-analysis for combining results from randomized clinical trials and medical-practice databases. *Int J Technol Assess Health Care*1993;9:440-449.

9. Eddy DM. The confidence profile method: a Bayesian method for assessing health technologies. *Operations Research* 1989;37:210-228.

10. Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *Am J Epidemiol* 2010;172:107-115.

11. Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. *JAMA* 2007;298:1209-1212.

12. Schmid CH, Lau J, McIntosh MW, 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;17:1923-1942.

13. Eberly LE. Consequences of event rate heterogeneity across non-randomized study sub-groups. *Stat Med* 2004;23:2023-2036.

14. McIntosh MW. The population risk as an explanatory variable in research synthesis of clinical trials. *Stat Med* Aug 30 1996;15:1713-1728.

15. Ioannidis JP, Lau J. Heterogeneity of the baseline risk within patient populations of clinical trials: a proposed evaluation algorithm. *Am J Epidemiol* 1998;148:1117-1126.

16. Walter SD. Variation in baseline risk as an explanation of heterogeneity in meta-analysis. *Stat Med* 1997;16:2883-2900.

17. Brand R, Kragt H. Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials. *Stat Med*1992;11:2077-2082.

18. Sharp SJ, Thompson SG. Analysing the relationship between treatment effect and underlying risk in meta-analysis: comparison and development of approaches. *Stat Med* 2000;19:3251-3274.

19. Wood S. mgcv: GAMs with GCV/AIC/REML smoothness estimation and GAMMs by PQL , http://cran.r-project.org/web/packages/mgcv/index.html. (2011, accessed 5 November 2011).

20. Snowden JM, Rose S, Mortimer KM. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. *Am J Epidemiol* Apr 1 2011;173:731-738.

21. Stitelman OM, Wester CW, De Gruttola V, van der Laan MJ. Targeted Maximum Likelihood Estimation of Effect Modification Parameters in Survival Analysis. *Int J Biostat* 2011;7:1-34.
22. Therneau T. survival: Survival analysis, including penalised likelihood, http://cran.r-project.org/web/packages/survival/index.html. (2011, accessed 12 November 12 2010).

TABLES AND FIGURES.

Table 1. Treatment effect estimate for the target sample and its standard deviation for different approaches (100 simulations).

Method	Mean	SD	MSE
Naïve	1.08	0.14	0.09
Risk standardized	0.94	0.36	0.15
CRAM	1.01	0.20	0.08
Truth	0.81	0.14	0.02

Abbreviations: MSE is mean squared error. SD is standard deviation.

Table 2. Treatment effect estimate for a more distant target sample and its standard deviation for different approaches: second set of simulations (100 simulations).

Method	Mean	SD	MSE
Naïve	1.09	0.14	0.09
Risk standardized	0.88	1.47	2.19
CRAM	0.93	0.46	0.26
Truth	0.72	0.12	0.01

Abbreviations: MSE is mean squared error. SD is standard deviation.

Figure 2. Distributions of baseline risk of outcome (probability of outcome independent of the treatment) in the trial, calibration, and target samples.

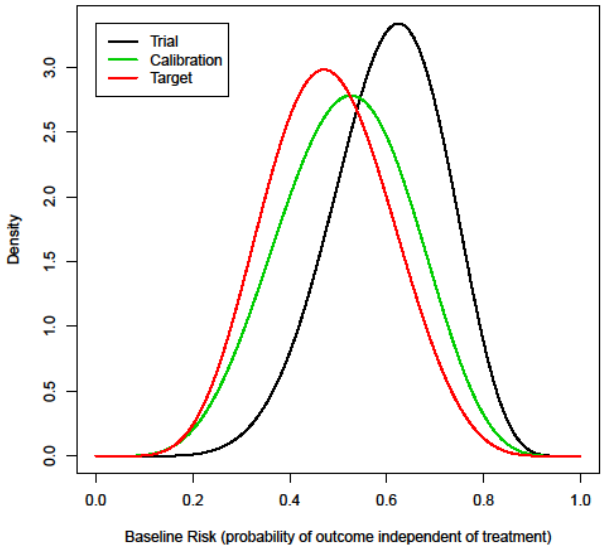
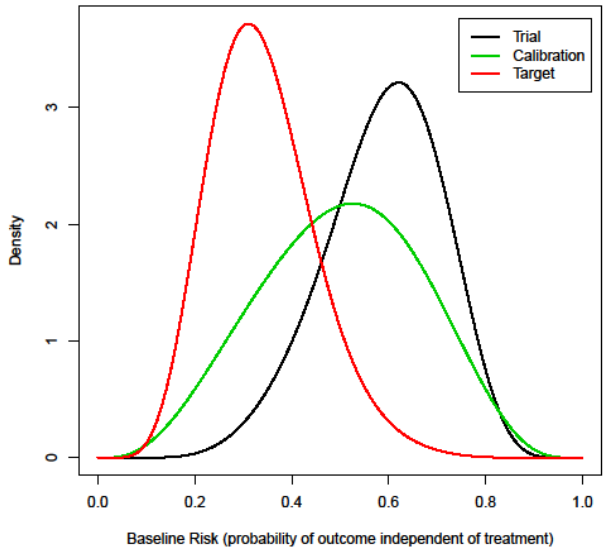


Figure 3. Distributions of baseline risk of outcome (probability of outcome independent of the treatment) in the trial, observational, and target samples in the second set of simulations for a more distant target sample.



Appendix 1: Proof that a single variable is sufficient to capture all unmeasured confounding (Courtesy: Constantine Frangakis, personal communication).

Suppose we have a binary treatment A, with potential outcomes Y(A=a), a=0,1. Ignorability (or no further confounding) after conditioning on a covariate would mean that {Y(0), Y(1) } is independent of A given a vector of covariates **X**. This implies that

{Y(0), Y(1)} is independent of A given e(**X**), (A1)

where e(**X**) = prob(A=1 | **X**).

Now, obviously {Y(0),Y(1)} is independent of A given {Y(0),Y(1)}, which, from (A1) implies that {Y(0),Y(1)} is independent of A given e({Y(0),Y(1) }), where e({Y(0), Y(1)}) = prob(A=1 | Y(0),Y(1)) =: U. The latter variable, U, is a scalar. This completes the proof.

This U is the simplest scalar that balances the covariates **X** between the treatment groups in the sense that it is a function of any other variable that creates such a balance. U can be either continuous or categorical.