Do complementary and alternative asthma treatments affect respiratory fitness?

Comparing estimates using Neyman's repeated sampling approach and Bayesian linear regression

Mao Hu

April 25, 2014

1 Introduction

In addition to evidence-based medicine, many people in the United States and Canada use complementary and alternative medicine. This type of medicine consists of a wide array of therapies not considered part of conventional medicine, including deep breathing exercises, hypnosis, and homeopathic treatment. These therapies are used by a large proportion of people - in a study conducted by the National Center for Complementary and Alternative Medicine, 38.3 % of adults in a national survey were using complementary or alternative medicines in 2007 [1]. However, it is unclear whether many forms of alternative medicine are useful for treating illnesses.

In this analyses, we examine the causal effect of using complementary and alternative medicine within the last year on respiratory fitness in a sample of patients with asthma using conventional asthma-controlling medicine. Assuming all confounding variables were collected, both the Neyman repeated sampling method and Bayesian linear regression estimate the causal effect to be near zero. Thus there is little evidence that use of complementary or alternative medicine either improves or harms respiratory fitness.

2 Data

The data comes from a sample of patients from British Columbia with a self-reported medical diagnosis of asthma recruited via random-digit dialing [2]. **Table 1** summarizes variables that were collected from each patient during a visit a survey center after being recruited. The treatment variable of interest is `cam`, an indicator for whether or not the subject used complementary or alternative asthma treatments in the past 12 months, such as acupuncture, homeopathy, or massage. The outcome variable of interest is `fev1`, a

measure of respiratory fitness obtained by asking the patient to forcibly expel air into a spirometer while at the survey center.

Other collected covariates include `age`, `ethnicity`, `income`, `sex`, and `education`. One additional covariate of interest, whether or not the patient took asthma-controlling medicine, was also collected. Unfortunately, it was not possible to determine whether this variable described the patient before or after the decision to use complementary or alternative medicine. Therefore, we restrict the analysis to asthma patients who have used asthma-controlling medicine on 80 % or more days of the past year. **Table 1** displays descriptive statistics about this restricted sample.

Variable	**Description**	**Mean (SD)**	**Type**
`fev1`	Forced expiatory volume in 1 second. A measure of respiratory fitness.	2.362 (0.816)	Outcome
`cam`	Whether or not subject used complementary or alternative asthma treatments in the past 12 months.	0.39	Treatment
`age`	Age of subject, in years.	56.89 (14.53)	Covariate
`ethnicity`	1 if subject is White, 0 if subject is Asian/other	0.84	Covariate
`income`	1 if subject's household income is more than or equal to $ 60,000, 0 if not.	0.66	Covariate
`sex`	1 if subject is male, 0 if subject is female	0.29	Covariate
`education`	1 if subject has college education or above, 0 if less than college	0.73	Covariate

Table 1: Description of variables ($n = 202$)

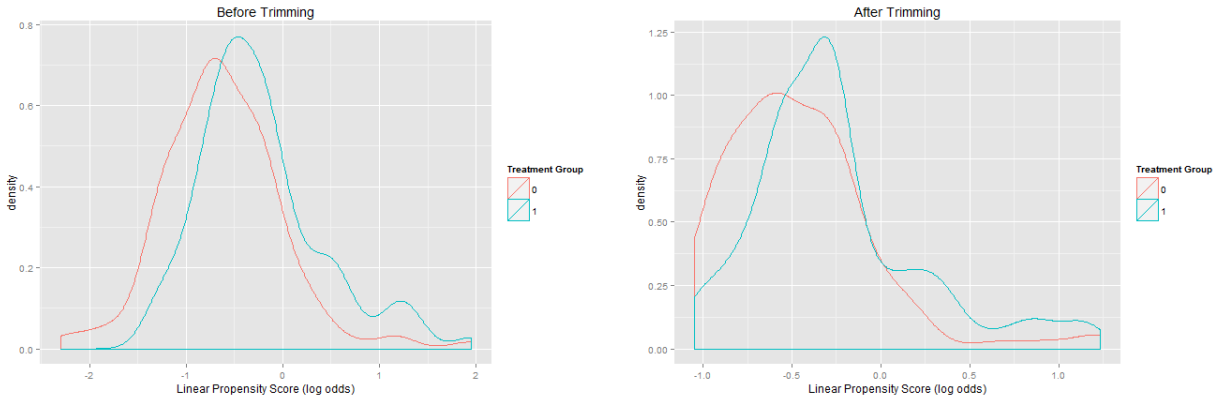
3 Improving Covariate Balance

In a randomized experiment, it is possible to make causal claims because the observed and unobserved covariates are balanced between the treatment and control groups, on average. However in observational studies, there is often covariate imbalance. Indeed, in **Figure 2** we visualize t-statistics comparing the covariates between the treatment and control groups using black dots. Notably, there is a smaller proportion of males who take alternative medicines for asthma than those who do not. The other covariates are not as strongly imbalanced. However, it is possible to obtain better causal estimates by using covariate-balancing methods, such as subclassification on the propensity score.

3.1 Estimating Propensity Scores

To obtain less confounded estimates of causal effects while analyzing observational data, Rubin suggests mimicking the design phase of a randomized experiment. The objective is to obtain groups with `cam`=1 that are similar based on their covariates to groups with `cam`=0. In an experimental design, it is possible to enforce covariate balance by "blocking" on covariates and then randomizing units in each block to the treatment and control groups [5]. Using estimated propensity scores, we can subclassify units to improve covariate balance retroactively.

To obtain estimates of the propensity scores, we fit a stepwise logistic regression of `cam` on the covariates, second-order interaction between the covariates, and the square of `age`. **Figure 1a** displays plots of the linear propensity scores fitted using the logistic regression. We then trim off control units with propensity scores below treated units and treated units with propensity scores above control units and refit the model. We repeat this process iteratively to remove units with extreme covariate values. By trimming away units with extreme propensity scores, we make the assumption of probabalistic treatment assignment more plausible. **Figure 2b** displays linear propensity scores for the trimmed data ($n = 191$) We observe better overlap between the `cam`=1 and `cam`=0 groups.



(a) Before Trimming (b) After Trimming

Figure 1: Linear Propensity Scores, `cam`=1 vs `cam`=0

3.2 Subclassification Using Propensity Scores

We then divide the dataset into subclasses based on quantiles of the propensity scores. One property of the propensity score is that balancing on the propensity scores also balances covariates. Indeed, by creating two subclasses of the sample based on propensity scores, we obtain substantial improvement in covariate balance, as visualized by red dots in **Figure 2**. Two subclasses is sufficient for this analysis - although we

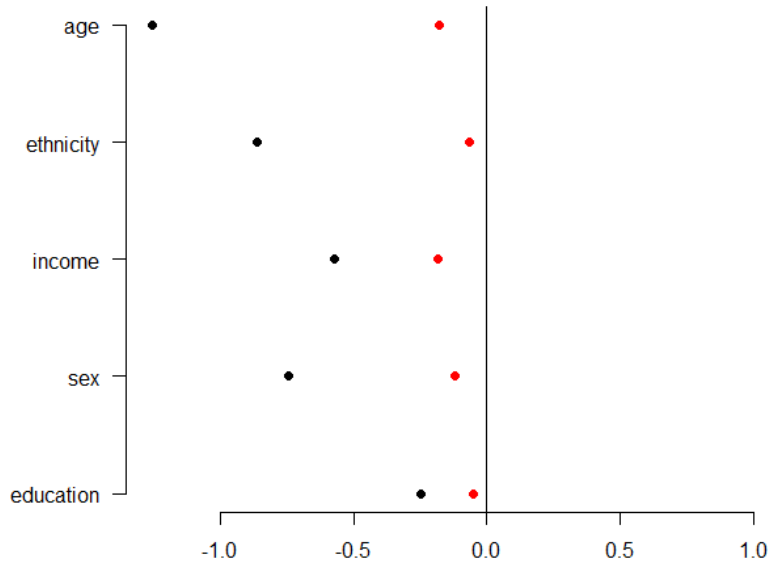


Figure 2: Improvement in covariate balance with two subgroups

might obtain better covariate balance by including more subclasses, more subclasses would also increase the variance of the estimator of the causal effect.

4 Estimation of Superpopulation Average Treatment Effect

The objective of the study is to estimate the superpopulation Average Treatment Effect (ATE) of using complementary and alternative asthma treatments on respiratory fitness, as measured by forced expiatory volume. Since the subjects were not randomly sampled from a larger population, we can only make inference to a superpopulation of hypothetical units with similar characteristics to the actual sample. In the Rubin causal model, this treatment effect is obtained from the potential outcomes of each unit under the treatment and the control. Let the potential outcomes $Y(1)$ and $Y(0)$ be the value of `fev1` when `cam`=1 and of `fev1` when `cam`=0 respectively. The estimand of the superpopulation ATE is the expected value of difference between $Y(1)$ and $Y(0)$:

$$\tau = E\big(Y(1) - Y(0)\big) \tag{1}$$

The fundamental problem in the Rubin causal framework is that $Y(1)$ and $Y(0)$ are never observed simultaneously. Thus we must estimate the population ATE using only observed values Y^{obs}, through an estimator $\hat{\tau}$. In Neyman's repeated sampling approach and Bayesian linear regression, the computation of $\hat{\tau}$ is treated

differently. In both cases, we make the assumption of unconfoundedness, or "no uncollected confounders:"

$$W|\boldsymbol{X}, Y(0), Y(1) = W|\boldsymbol{X}$$

In this study, the assumption of unconfoundedness is implausible. For example, one possible confounder is the patient's respiratory fitness before deciding to take complementary or alternative medicine. This prior respiratory fitness would be associated with `fev1` and `cam`. This analysis attempts to control for this specific uncollected covariate by proxy, via `age`, `gender`, and `ethnicity`, but there are other uncollected covariates which are likely associated.

Since our data now has subclassifications based on propensity scores, we can estimate (1) above using the within subclass estimators for the ATE weighted by the subclass sizes N_1, N_2, N_3 (Rubin, Chp. 17 p 10)

$$\hat{\tau}^{\text{strat}} = \frac{N_1}{N}\hat{\tau_1} + \frac{N_2}{N}\hat{\tau_2} \tag{2}$$

In the following section, we determine how to compute the point estimate (2) and interval estimates for the population ATE using Bayesian linear regression. For details on Neyman's repeated sampling approach, see Rubin's treatment [6, Chp. 9, 17].

4.1 Estimation Using Bayesian Linear Regression

Consider the joint distribution of the potential outcomes conditional on the covariates:

$$Pr(Y(0), Y(1)|\boldsymbol{X}) \tag{3}$$

According to Rubin and Li [6, Chp. 8] [4], we can rewrite (3) above as the product of joint probabilities for unit potential outcomes $Y_i(0), Y_i(1)$ conditional on a parameter θ, as a result of de Finetti's theorem.

$$Pr(Y(0), Y(1)|\boldsymbol{X}) = \prod_i Pr(Y_i(0), Y_i(1)|X_i, \theta)Pr(\theta|\boldsymbol{X}) \tag{4}$$

At this point, we are required to specify a sampling model for $Y(0), Y(1)|\boldsymbol{X}, \beta_c, \beta_t, \sigma_c, \sigma_t$. We will make several additional assumptions: first, that $Y_i(0)$ and $Y_i(1)$ conditional on $X_i, \beta_c, \beta_t, \sigma_c, \sigma_t$ are independent; second, that the association between the covariates and the potential outcomes are roughly linear; and third, that there no interaction terms between the covariates. Looking at scatterplots of the outcome versus the covariates, the second assumption does not seem implausible.

4.2 Model specification

Below is a description of a sampling model for $Y(0), Y(1)|\boldsymbol{X}, \beta_c, \beta_t, \sigma_c, \sigma_t$ and corresponding Zellner g-priors for the β_c, β_t parameters and Gamma priors for σ_c, σ_t, as described by Hoff [3].

$$1/\sigma_c^2 \sim \text{Gamma}(\nu_{c0}/2, \nu_{c0}\sigma_{c0}^2/2) \tag{5}$$

$$1/\sigma_t^2 \sim \text{Gamma}(\nu_{t0}/2, \nu_{t0}\sigma_{t0}^2/2) \tag{6}$$

$$\beta_c \sim \text{Normal}(0, g\sigma_c^2(\boldsymbol{X_c'X_c})^{-1}) \tag{7}$$

$$\beta_t \sim \text{Normal}(0, g\sigma_t^2(\boldsymbol{X_t'X_t})^{-1}) \tag{8}$$

$$Y_i(0), Y_i(1)|X_i, \theta \sim \text{independent Normal}\Big(\begin{bmatrix} X_i'\beta_c \\ X_i\beta_t \end{bmatrix}, \begin{bmatrix} \sigma_c^2 & 0 \\ 0 & \sigma_t^2 \end{bmatrix} \Big) \tag{9}$$

Here $\boldsymbol{X}$ is augmented by a column $\vec{1}$, for the intercept terms, and $\boldsymbol{X_c}$ and $\boldsymbol{X_t}$ are the matrices of covariates for the control and treatment units respectively. A more careful treatment of the following derivation can be found in Rubin [6, Chp. 8] and [4]. Rubin and Li focus on obtaining a posterior predictive distribution for the missing values Y^{miss}, but here we are only concerned with updating the parameters. The reason imputing Y^{mis} is unnecessary can be found by examining the estimator for population ATE, which can be written as follows using the sampling distribution of the potential outcomes:

$$\begin{aligned} \hat{\tau} =& E(X'\hat{\beta}_t - X'\hat{\beta_c}) \\ =& \bar{X}'\hat{\beta}_t - \bar{X}'\hat{\beta}_c \end{aligned} \tag{10}$$

We see that we only need to obtain estimates of the coefficients β_c, β_t to obtain an estimate of the population average treatment effect. Unfortunately, we do not have both potential outcomes $Y_i(0), Y_i(1)$ for each unit. However, thanks to the assumption of unconfoundedness, we can rewrite (9) above as:

$$\begin{aligned} &Y_i^{mis}, Y_i^{obs}|X_i, W_i, \theta \sim \\ &\qquad \text{Normal}\Big(\begin{bmatrix} X_i'B_cW_i + X_i'B_t(1-W_i) \\ X_i'B_c(1-W_i) + X_i'B_tW_i \end{bmatrix}, \begin{bmatrix} \sigma_c^2(W_i) + \sigma_t^2(1-W_i) & 0 \\ 0 & \sigma_c^2(1-W_i) + \sigma_t^2(W_i) \end{bmatrix} \Big) \end{aligned} \tag{11}$$

We also see that since $Y_i(0), Y_i(1)$ are independent conditional on X_i and the parameters, the marginal density of $Y_i^{obs}|W, \theta, X_i$ is:

$$Y_i^{obs}|W=0,\theta,X_i \sim \text{Normal}\Big(X_i'B_c, \sigma_c^2\Big) \tag{12}$$

$$Y_i^{obs}|W=1,\theta,X_i \sim \text{Normal}\Big(X_i'B_t, \sigma_t^2\Big) \tag{13}$$

We can now obtain the posterior distribution of the parameters $\beta_c, \beta_t, \sigma_c, \sigma_t$ given the observed quantities $Y^{obs}, W, \boldsymbol{X}$.

$$\begin{aligned} Pr(\beta_c,\beta_t,\sigma_c^2,\sigma_t^2|Y^{obs},W,\boldsymbol{X}) \propto & p(\theta)Pr(Y^{obs}|\theta,W,\boldsymbol{X}) && (14)\\ = & p(\theta)\times\prod_i Pr(Y_i^{obs}|W,X_i,\theta) && (15)\\ = & \prod_{i:W_i=1} N(y_i^{obs};X_i'\beta_t,\sigma_t^2)\times N(\beta_t;\beta_{t0},\Sigma_{t0})\times Ga(1/\sigma_t^2;\nu_{t0}/2,\nu_{t0}\sigma_{t0}^2/2) && (16)\\ & \times\prod_{i:W_i=0} N(y_i^{obs};X_i'\beta_c,\sigma_c^2)\times N(\beta_c;\beta_{c0},\Sigma_{c0})\times Ga(1/\sigma_c^2;\nu_{c0}/2,\nu_{c0}\sigma_{c0}^2/2) \end{aligned}$$

According to Hoff (Hoff) corresponding posterior conditional distributions are:

$$\begin{aligned} \beta_c|\beta_t,\sigma_c^2,\sigma_t^2,Y^{obs},W,\boldsymbol{X} \sim & N\big(\beta_{c0*},\Sigma_{c0}^*\big) && (17)\\ \beta_{c0}^* = & \frac{g}{g+1}(\boldsymbol{X_c'X_c})^{-1}\boldsymbol{X_c'}Y_c^{obs}\\ \Sigma_{c0}^* = & \frac{g}{g+1}\sigma_c^2(\boldsymbol{X_c'X_c})^{-1} \end{aligned}$$

$$\begin{aligned} \beta_t|\beta_c,\sigma_c^2,\sigma_t^2,Y^{obs},W,\boldsymbol{X} \sim & N\big(\beta_{t0*},\Sigma_{t0}^*\big) && (18)\\ \beta_{t0}^* = & \frac{g}{g+1}(\boldsymbol{X_t'X_t})^{-1}\boldsymbol{X_t'}Y_t^{obs}\\ \Sigma_{t0}^* = & \frac{g}{g+1}\sigma_t^2(\boldsymbol{X_t'X_t})^{-1} \end{aligned}$$

$$\begin{aligned} \frac{1}{\sigma_c^2}|\beta_c,\beta_t,\sigma_t^2,Y^{obs},W,\boldsymbol{X} \sim & Ga([\nu_{c0}+n_c]/2,[\nu_{c0}\sigma_{c0}^2+SSR_c(\beta_c)]/2) && (19)\\ SSR_c(\beta_c) = & Y_c^{obs\ \prime}Y_c^{obs}-2\beta_c'\boldsymbol{X_c'}Y_c^{obs}+\beta_c'\boldsymbol{X_c'X_c}\beta_c \end{aligned}$$

$$\frac{1}{\sigma_t^2}|\beta_c, \beta_t, \sigma_c^2, Y^{obs}, W, \boldsymbol{X} \sim Ga([\nu_{t0} + n_t]/2, [\nu_{t0}\sigma_{t0}^2 + SSR_t(\beta_t)]/2) \tag{20}$$

$$SSR_t(\beta_t) = Y_t^{obs\ \prime} Y_t^{obs} - 2\beta_t' \boldsymbol{X_t'} Y_t^{obs} + \beta_t' \boldsymbol{X_t'} \boldsymbol{X_t} \beta_t$$

We can now obtain draws from the joint posterior distribution of $\beta_c, \beta_t, \sigma_c, \sigma_t$ conditional on $Y^{obs}, \boldsymbol{X}, W$ by iteratively sampling from these posterior conditional distributions (17), (18), (19), (20). At the end of each iteration, we compute the estimator (10) above to obtain draws from the posterior distribution of the superpopulation ATE.

In this analysis, we have broken up the units into two subgroups. Here we decided to treat each subgroup as its own analysis, estimating the posterior distribution for the parameters in each subgroup. Within each subgroup, we set g to be the subgroup size, $\nu_{c0} = \nu_{t0} = 1$, $\beta_{c0} = \beta_{t0} = \vec{0}$, and σ_c, σ_t to be equal to the ordinary least squares estimate of σ from a linear regression of `fev1` on `cam` and the other covariates. This choice of prior is relatively dispersed over plausible values of the parameters. In this analysis, at each iteration we combine the draws from the subgroup-specific superpopulation ATE using (2) above to obtain draws from the posterior distribution of the superpopulation ATE.

5 Results

Table 2 compares the point and interval estimates using Neyman's repeated sampling method and Bayesian linear regression for the superpopulation average treatment effect of using complementary or alternative medicine on forced expiratory volume in one second. The point estimates of the superpopulation average treatment effect, -0.040 for Neyman's repeated sampling method and -0.074 for Bayesian linear regression respectively are small in magnitude compared to the overall scale of `fev1`. Therefore, the effect of complementary and alternative medicine usage on respiratory fitness is minimal.

Comparing the interval estimates of the ATE from the Neyman style inference and Bayesian style inference, we see that they are largely similar relative to the scale of `fev1`. From the Neyman-style analysis, we are 95 % confident that the superpopulation average treatment effect of `cam=1` on `fev1` is between -0.31 and 0.23. Similarly, our 95 % credible interval for $\hat{\tau}$ is -0.28 and 0.13.

The assumption of unconfoundedness made in this analysis is most likely implausible. For example, a patient's respiratory fitness before using complementary or alternative medicine is a plausible confounder between `fev1` and `cam`. Since the treatment effect of `cam` is weak or non-existent, this analysis is sensitive to uncollected confounding variables. In addition, the population itself is not a random sample from a larger population, so inference to a population is speculative.

Method	Neyman's Repeated Sampling	Bayesian Linear Regression
Point estimate $\hat{\tau}$	-0.040	-0.074
Standard error of $\hat{\tau}$	0.13	-
95 % confidence interval for $\hat{\tau}$	-0.31 to 0.23	-
Standard deviation of $\hat{\tau}$	-	0.11
95 % credible interval for $\hat{\tau}$	-	-0.28 to 0.13

Table 2: Point and interval estimates for the population average treatment effect τ

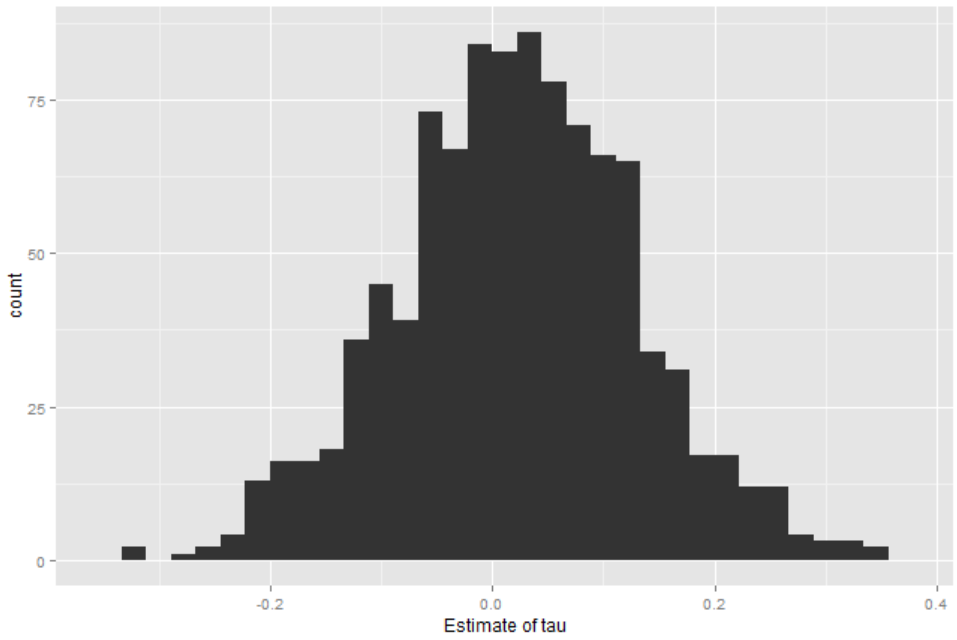


Figure 3: 995 draws from posterior predictive distribution of $\hat{\tau}$ from Bayesian linear regression

6 Conclusion

The estimated causal effect of using complementary and alternative asthma treatments on 1-second forced expiatory volume estimated to be near 0, assuming no uncollected covariates. The estimates of this causal effect from Neyman-style inference and Bayesian linear regression are similar. Although we achieved good balance between collected covariates using subclassification, this causal estimate should be used with caution, since the assumption of unconfoundedness is implausible. In addition, since the patients were not randomly sampled, inference beyond a hypothetical superpopulation with similar characteristics to the study population is speculative.

Considering the additional assumptions needed to specify a model used in Bayesian linear regression, there is little motivation to use a Bayesian approach for estimating a superpopulation causal effect in this context.

References

[1] Patricia M Barnes, Barbara Bloom, Richard L Nahin, National Center for Health Statistics (US), et al. Complementary and alternative medicine use among adults and children: United states, 2007, 2008.

[2] Wenjia Chen, J Mark FitzGerald, Roxanne Rousseau, Larry D Lynd, Wan C Tan, and Mohsen Sadatsafavi. Complementary and alternative asthma treatments and their association with asthma control: a population-based study. *BMJ open*, 3(9):e003360, 2013.

[3] Peter D Hoff. *A first course in Bayesian statistical methods*. Springer, 2009.

[4] Fan Li. Unpublished text about bayesian causal analysis, 2014.

[5] Donald B Rubin. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in medicine*, 26(1):20–36, 2007.

[6] Donald B Rubin. *Unnamed Textbook Used in STA 320*. 2014.