Biostatistics (2017), **0**, 0, *pp.* 1–7
doi:10.1093/biostatistics/supp

Supplementary material to "Estimating Causal Effects from a Randomized Clinical Trial when Noncompliance is Measured with Error"

JEFFREY A. BOATMAN*[1], DAVID M. VOCK[1], JOSEPH S. KOOPMEINERS[1],

ERIC C. DONNY[2]

[1]*Division of Biostatistics, University of Minnesota, Minneapolis, MN, USA*

[2]*Department of Psychology, University of Pittsburgh, Pittsburgh, PA, USA*

jeffrey.boatman@gmail.com

Section A: EM Algorithm M Step Updates for α and ξ

Here we give details on the EM algorithm updates for α and ξ, the coefficient vector for $\rho(a, x, y, d; \alpha) = \Pr(C = 1|A = a, X = x, Y = y, D = d; \alpha)$ and $f(B|A, X, Y, D, C; \xi)$, the conditional density of B given A, X, Y, D, C, discussed in Section 3.1. We use notation from **?**. Random vectors are denoted with bold type, e.g. $\boldsymbol{A} = (A_1, \ldots, A_n)^T$.

The complete data log likelihood for $\boldsymbol{\theta} = (\xi^T, \alpha^T)^T$ is

$$L(\boldsymbol{\theta}|\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{C}) = \sum_{i=1}^{n} \{C_i \log f(B_i|A_i, X_i, Y_i, D_i, c = 1) + (1 - C_i) \log f(B_i|A_i, X_i, Y_i, D_i, c = 0)\}$$
$$+ C_i \log \rho(A_i, X_i, Y_i, D_i; \alpha) + (1 - C_i) \log \{1 - \rho(A_i, X_i, Y_i, D_i; \alpha)\}.$$

The conditional expectation of the E step is given by

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(\nu)}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{C}) &= E_{\boldsymbol{\theta}^{(\nu)}} \left\{ \log L(\boldsymbol{\theta} | \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{C}) | \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{D} \right\} \\ &= \sum_{i=1}^{n} \left\{ w_i^{(\nu)} \log f(B_i | A_i, X_i, Y_i, D_i, c = 1) + (1 - w_i^{(\nu)}) \log f(B_i | A_i, X_i, Y_i, D_i, c = 0) \right\} \\ &+ w_i^{(\nu)} \log \rho(A_i, X_i, Y_i, D_i; \alpha) + (1 - w_i^{(\nu)}) \log \left\{ 1 - \rho(A_i, X_i, Y_i, D_i; \alpha) \right\} \end{aligned}$$

where

$$\begin{aligned} w_i^{(\nu)} &= E_{\boldsymbol{\theta}^{(\nu)}}(C_i | A_i, B_i, X_i, Y_i, D_i) \\ &= \frac{f(B_i | A_i, X_i, Y_i, D_i, c = 1; \xi^{(\nu)}) \cdot \rho(A_i, X_i, Y_i, D_i; \alpha^{(\nu)})}{f(B_i | A_i, X_i, Y_i, D_i, c = 1; \xi^{(\nu)}) \cdot \rho(A_i, X_i, Y_i; \alpha^{(\nu)}) + f(B_i | A_i, X_i, Y_i, D_i, c = 0; \xi^{(\nu)}) \cdot \left\{ 1 - \rho(A_i, X_i, Y_i; \alpha^{(\nu)}) \right\}}. \end{aligned}$$

The M step update $\alpha^{(\nu+1)}$ is the solution to the score equations

$$\sum_{i=1}^{n} \frac{w_i^{(\nu)} - \rho(A_i, B_i, X_i, Y_i, D_i; \alpha)}{\rho(A_i, B_i, X_i, Y_i, D_i; \alpha) \left\{ 1 - \rho(A_i, B_i, X_i, Y_i, D_i; \alpha) \right\}} \frac{\partial \rho(A_i, B_i, X_i, Y_i, D_i; \alpha)}{\partial \alpha^T} = 0.$$

If there are no shared parameters in the conditional density of the biomarker among the compliers and noncompliers, then the M step update $\xi^{(\nu+1)}$ is the solution to the weighted score equations

$$\sum_{i=1}^{n} \frac{\partial \log f(B_i | A_i, X_i, Y_i, D_i, c = 1; \xi_C)}{\partial \xi_C^T} w_i^{(\nu)} = 0$$

$$\sum_{i=1}^{n} \frac{\partial \log f(B_i | A_i, X_i, Y_i, D_i, c = 0; \xi_{NC})}{\partial \xi_{NC}^T} (1 - w_i^{(\nu)}) = 0,$$

where ξ_C and ξ_{NC} are the parameters corresponding to the distribution of the compliers and non-compliers.

Section B: Directed Acyclic Graph assumed in Simulation and Application

Figure 1 shows one possible Directed Acyclic Graph (DAG) for the CENIC-p1 data and other possible applications. The arrows between B and Y are dashed because CENIC-p1 is a unique trial in that the DAG has the causal relationship $Y \rightarrow B$, i.e., the outcome, the number of cigarettes smoked, causes the biomarker, TNE. The arrow from X to B is dotted because, in our simulation and application, we assume that this arrow does not exist; if this arrow is removed, the DAG implies that X and B are conditionally independent given A, Y and C. This simplifies the estimation of the conditional density of B given A, B, X, and Y, but we note that this is assumption is not required for the method in general.

In contrast, the DAG in many clinical trials would have $B \rightarrow Y$. For example, in a clinical trial investigating blood pressure-lowering medication, we might expect that B, the circulating levels of medication or a metabolite, would cause the amount of decrease in blood pressure Y. In this case, it may be more intuitive to model the conditional density of Y given A, B, X, D, C and $\Pr(C = 1|A, B, X, D)$ in estimating the numerator of the weights.

We also note that in some scenarios, it may be reasonable to assume that confounders X or the response Y cause the subject to report compliance honestly, H. The proposed method still provides consistent estimation of the causal effect.

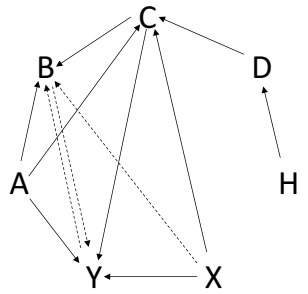

Fig. 1. one possible Directed Acyclic Graph (DAG) for the CENIC-p1 data. Dashed lines are show where assumptions may change the DAG.

SECTION C: ADDITIONAL APPLICATION TABLES

Table 1. Estimated mixture distribution coefficients and parameters (95% bootstrap percentile intervals) assuming $B|Y, C, D = 1 \sim N(\gamma_0 + \gamma_1 Y, \sigma^2)$. $B = \log(TNE)$, Y = study cigarettes smoked per day.

Component	$\hat{\gamma_0}$	$\hat{\gamma_1}$	$\hat{\sigma}$
Compliant ($C = 1$)	0.12(-0.96, 1.19)	0.02(-0.01, 0.07)	0.89(0.55, 1.34)
Non-compliant ($C = 0$)	3.33(2.79, 4.13)	0.01(-0.04, 0.04)	0.78(0.51, 1.13)

Table 2. Estimated coefficients and 95% bootstrap percentile confidence intervals for the mixture distribution logistic model $\Pr(C = 1|A = 1, X, Y, D = 1)$. MNWS: Minnesota Nicotine Withdrawal Scale, higher scores indicating greater withdrawal symptoms. QSU: Questionnaire of Smoking Urges, higher scores indicating greater urges. CES: Cigarette Evaluation Scale, higher scores indicating greater satisfaction with study cigarettes.

Coefficient	Estimate	95% CI
Intercept	1.261	(-2.396, 14.724)
Age	0.017	(-0.031, 0.083)
Baseline log(TNE)	-0.944	(-3.642, -0.440)
Baseline Cigarettes per day	-0.011	(-0.188, 0.222)
Max Symptoms Week 1	0.027	(-0.088, 0.153)
MNWS Week 5	0.086	(-0.081, 0.292)
QSU Study Cigarettes Week 5	0.045	(-0.021, 0.314)
QSU Usual Brand Cigarettes Week 5	-0.048	(-0.219, -0.005)
CES Satisfaction Week 5	0.202	(-0.469, 0.846)
Y (Study Cigarettes per day)	-0.004	(-0.251, 0.116)

Table 3. Estimated coefficients and 95% bootstrap confidence intervals for logistic model for denominator of weights. MNWS: Minnesota Nicotine Withdrawal Scale, higher scores indicating greater withdrawal symptoms. QSU: Questionnaire of Smoking Urges, higher scores indicating greater urges. CES: Cigarette Evaluation Scale, higher scores indicating greater satisfaction with study cigarettes.

Coefficient	Estimate	95% CI
Intercept	-0.707	(-4.092, 5.104)
Age	0.022	(-0.029, 0.082)
Baseline log(TNE)	-0.475	(-1.698, 0.042)
Baseline Cigarettes per day	-0.059	(-0.205, 0.022)
Max Symptoms Week 1	0.013	(-0.087, 0.123)
MNWS Week 5	0.050	(-0.109, 0.230)
QSU Study Cigarettes Week 5	0.013	(-0.063, 0.127)
QSU Usual Brand Cigarettes Week 5	-0.030	(-0.144, 0.009)
CES Satisfaction Week 5	0.418	(-0.216, 1.043)

Table 4. Baseline characteristics, confounders, and the biomarker week 6 log(TNE) for the application. Values given are n(%) for categorical variables and mean(sd) for numeric variables.

	Self-Report		Estimated $\Pr(C = 1 \mid B, X, Y, D)$	
Characteristic	Non-Compliant	Compliant	<=0.5	> 0.5
n	88	137	169	49
Male	45(51%)	74(54%)	92(54%)	22(45%)
Female	43(49%)	63(46%)	77(46%)	27(55%)
White	46(52%)	75(55%)	87(51%)	28(57%)
Black	32(36%)	47(34%)	62(37%)	17(35%)
Other	10(11%)	15(11%)	20(12%)	4(8%)
Age	40.14(13.14)	42.48(13.37)	40.57(13.04)	44.90(13.83)
Baseline log(TNE)	3.71(0.94)	3.69(0.84)	3.79(0.84)	3.44(0.88)
Baseline Cigarettes per day	16.78(8.41)	14.78(6.83)	15.96(7.46)	13.83(6.40)
Max Symptoms Week 1	12.85(7.30)	12.00(7.03)	12.10(7.22)	12.31(6.73)
MNWS Week 5	7.27(5.27)	6.13(4.82)	6.38(4.86)	6.88(5.36)
QSU Study Cigarettes Week 5	19.69(13.80)	20.32(12.90)	19.46(12.96)	22.43(14.43)
QSU Usual Brand Cigarettes Week 5	29.42(16.60)	25.70(16.33)	27.62(16.70)	24.78(15.21)
CES Satisfaction Week 5	2.06(1.13)	2.79(1.46)	2.33(1.33)	3.15(1.43)
Week 6 log(TNE)*	3.43(2.93, 4.10)	2.28(0.86, 3.80)	3.47(2.93, 4.07)	0.29(-0.40, 0.99)

*Mean(1^{st} quartile, 3^{rd} quartile)

SECTION D: MEASUREMENT ERROR IN Y

Throughout the manuscript we assumed the correctly or incorrectly reporting compliance has no effect on the self-reported outcome, that is, i.e. D does not affect Y. We note that many clinical trials rely on self-reported compliance but use endpoints which are direct physiologic measures or adjudicated clinical events. For example, in CENIC-p1, ITT estimates of the effect of nicotine level on other physiologic endpoints including expired carbon monoxide were included in the primary analysis (**?**).

Furthermore, under very plausible assumptions concerning the self-reported error of cigarette consumption, we can still obtain consistent estimators of the causal effect. In particular, assume that Y is the outcome without any self-report or measurement error and we are interested in estimating $E\{Y^*(a,0)\}$, the average effect if possibly contrary to fact all subjects were assigned treatment group a, fully complied with the assigned treatment, and there was no measurement error in the response. Instead of observing Y directly, we observe $W = Y + \epsilon$, where ϵ is the self-report or measurement error. In this case, the CURE estimator becomes

$$\sum_{i=1}^{n} \frac{E\left(C_i|A_i, B_i, X_i, D_i, W_i\right)}{E\left(C_i|A_i, X_i\right)} \left\{W_i - \mu(a,0)\right\} I(A_i = a) = 0.$$

Note that this is a mean-zero estimating function provided that $E\left(\epsilon_i|C_i = 1, X_i\right) = 0$. That is, among compliers, the self-reported error is not systemic at all levels of the confounders. As in the main paper, for simplicity we consider only a single-arm trial with $a = 1$ for all participants, but the results easily generalize to multi-arm trials.

$$
\begin{aligned}
&E\left[\frac{E\left(C_i|B_i, X_i, D_i, W_i\right)}{E\left(C_i|X_i\right)}\left\{W_i - \mu(1,0)\right\}\right] \\
&= E\left[\frac{C_i}{E\left(C_i|X_i\right)}\left\{Y_i + \epsilon - \mu(1,0)\right\}\right] \\
&= E\left[\frac{C_i}{E\left(C_i|X_i\right)}\left\{Y_i - \mu(1,0)\right\} + \frac{C_i}{E\left(C_i|X_i\right)}\epsilon_i\right] \\
&= 0 + E\left[\frac{C_i}{E\left(C_i|X_i\right)}E\left(\epsilon_i|C_i, X_i\right)\right] \\
&= E\left[\frac{C_i}{E\left(C_i|X_i\right)}E\left(\epsilon_i|C_i = 1, X_i\right)\right] = 0.
\end{aligned}
$$