Submitted to the Annals of Applied Statistics
arXiv: arXiv:0000.0000

SUPPLEMENTARY SIMULATIONS FOR DETERMINISTIC VARIABLE SELECTION FOR LOGISTIC REGRESSION MODELS WITH RELATED COVARIATES*

By Matthew D. Koslovsky †, Michael D. Swartz†,
Luis Leon-Novelo†, Wenyaw Chan†,
and Anna V. Wilkinson‡

The University of Texas Health Science Center at Houston† and Austin ‡

Supplement B: Selecting Qualitative Covariates. The purpose of this simulation is to show deterministic annealing EMVS's ability to identify continuous or qualitative covariates associated with a binary outcome from a pool of potential covariates. For each model, we simulated 500 data sets of $n = 400$ observations from a model similar to Eq. 1 in the main manuscript. Continuous covariates followed a multivariate normal distribution with mean zero, variance one, and an exchangeable covariance structure, parameterized with ρ. We set $\rho = \{0, 0.4, 0.8\}$. Qualitative covariates came from a multinomial distribution with equal probabilities set for each of the $m+1$-levels which sum to one. Qualitative covariates of size $m+1$ were reparameterized with m indicator variables, $D_l, l = 1, \ldots, m$. Each indicator variable was set to one if their corresponding covariate was the $l+1$-level of the qualitative covariate and zero otherwise. For instance a two-level qualitative covariate was reparameterized with one indicator variable D_1 that equals one if the qualitative covariate was equal to the second level and zero otherwise. The full model contained an intercept term, 12 continuous covariates $(x_{c,1}, \ldots, x_{c,12})$, 12 two-level qualitative covariates $(x_{b,13}, \ldots, x_{b,24})$, and one four-level qualitative covariate $(x_{d,25}, \ldots, x_{d,27})$. To determine the variance of inclusion, v_0, we used regularization plots, as recommended by [2]. We considered a range of v_0 so that the upper(lower) bound of the 95% prior probability of exclusion for the odds ratio spans from 1.01(0.99) to 1.15(0.87) by 0.01 to maintain interpretability. We observed that at around $v_0 = 0.0015$, equivalent to an odds ratio between $[0.93, 1.08]$, the plots stabilized. The 95% prior probability of inclusion for the odds ratio is fixed to cover $[1/4, 4]$, $v_1 = 0.5$, similar to [1]. We applied our variance adjustment to indicator variables, as described in Section 2.3 in the main manuscript. The models were compared with and without grouping for the indicator variables. Comparisons were made at the indicator level. For example, if one indicator variable in the group was truly associated, we considered any levels that were not selected by the model a false negative. For simplicity, the intercept term α_0 is set to zero in each of the true models. The following models tested our

*Supported in part by National Cancer Institute/NIH Grant R25 CA57712, National Institute of General Medical Sciences/NIH Grant T32GM074902, and the Michael & Susan Dell Foundation, Michael & Susan Dell Center for Healthy Living

method's performance:

Model 3.1.1 Null model:
$logit(\omega(\mathbf{x}_i)) = 0$
Model 3.1.2 No indicator levels associated:
$logit(\omega(\mathbf{x}_i)) = 0.65x_{c,1} - 0.5x_{c,2} - 0.65x_{b,13} + 0.5x_{b,14}$
Model 3.1.3 Same as Model 3.1.2 but run without variance adjustment
Model 3.1.4 One indicator variable associated:
$logit(\omega(\mathbf{x}_i)) = 0.65x_{c,1} - 0.5x_{c,2} - 0.65x_{b,13} + 0.5x_{b,14} - 0.6_{d,25}$
Model 3.1.5 Two indicator variables associated:
$logit(\omega(\mathbf{x}_i)) = 0.65x_{c,1} - 0.5x_{c,2} - 0.65x_{b,13} + 0.5x_{b,14} - 0.6_{d,25} - 0.5x_{d,26}$
Model 3.1.6 All indicator variables associated:
$logit(\omega(\mathbf{x}_i)) = 0.65x_{c,1} - 0.5x_{c,2} - 0.65x_{b,13} + 0.5x_{b,14} - 0.6_{d,25} - 0.5x_{d,26} + 0.4x_{d,27}$

First, we compared the method's performance using grouped indicator variables for qualitative covariates against treating them independently (Table S1 on page 4). Under the assumption that an associated indicator variable justifies the other level's inclusion, we found that grouping increased the weighted average correct association percentage and decreased the average false positive rate (FPR) and false negative rate (FNR). For all models, the overall performance weakened for higher correlation structures. We found an average FPR for the null model (3.1.1) with moderate correlation fell around 0.06. Additionally, Figures S1 and S2 show a decrease in performance for weaker effects and qualitative terms. As the number of associated terms in a group of indicator variables increased, so did the method's ability to identify the group as associated. Comparing model 3.1.2 and 3.1.3, our simulations suggest that adjusting the exclusion variance reduced the FPR for the grouped indicators.

Supplement C: Selecting Interaction Terms. In the following simulations, our aim was to accommodate heredity constraints for interaction terms. For each model, we simulated 500 data sets of $n = 200$ observations from the quadratic model similar to [3] for linear regression models. Here, parent terms, $x_{c,1}, x_{c,2}$, and $x_{c,3}$, followed the same distribution as the continuous covariates above. The full model comprised an intercept term and 9 possible covariates: 3 parent terms $(x_{c,1}, x_{c,2}, x_{c,3})$, 3 pairwise interactions $(x_{c,1}x_{c,2}, x_{c,1}x_{c,3}, x_{c,2}x_{c,3})$, and 3 squared terms $(x^2_{c,1}, x^2_{c,2}, x^2_{c,3})$. For these simulations, we show how our method could incorporate prior knowledge to achieve a research objective, such as an odds ratio between a specific range being clinically irrelevant. Here, each model was tuned so that the 95% prior probability of exclusion for the odds ratio covers $[0.95, 1.05]$, $v_0 = 0.00062$ and the 95% prior probability of inclusion for the odds ratio covers $[1/4, 4]$. We restricted quadratic terms' inclusion with $\mathbf{q} = (0, 1)$ and applied a strong, $\mathbf{a} = (0, 0, 0, 1)$, and a weak, $\mathbf{a} = (0, 1, 1, 1)$, heredity constraint for interaction terms. These parameterizations indicate which combinations of parental terms' inclusion and exclusion permited the consideration of an interaction term's inclusion. The constrained models were compared to a model with no heredity constraints (i.e., iid case), $\mathbf{a} = (1, 1, 1, 1)$, which ignored the relations between covariates. We constructed three models to test our method's ability to accommodate heredity constraints.

Model 3.2.1 The true model followed a strong heredity constraint:
$logit(\omega(\mathbf{x}_i)) = 0.65x_{c,1} - 0.65x_{c,2} + 0.5x_{c,1}x_{c,2}$
Model 3.2.2 The true model followed a weak heredity constraint:
$logit(\omega(\mathbf{x}_i)) = 0.65x_{c,1} - 0.65x_{c,2} + 0.5x_{c,1}x_{c,3}$
Model 3.2.3 The true model was not hierarchically well formulated:
$logit(\omega(\mathbf{x}_i)) = 0.65x_{c,1} - 0.65x_{c,2} + 0.5x_{c,3}^2$

Overall, the method's sensitivity to correlation structure for the weighted average correct association percentage was similar to Supplement B (Table S2 on page 5). Regardless of the true model's formulation, the strong heredity constraint favored a sparse model and has a lower average FPR and a higher weighted average correct association percentage. The strong heredity constraint experienced a higher average FNR when the true model followed strong heredity (Table S1 on page 4). Figures S3, S4, and S5 show the marginal results for the interaction models. Our simulations show that the method performed better controlling the FPR for pairwise interactions under the strong heredity constraint. However, the FPR for non-associated squared terms when the true model followed weak heredity (i.e., model 3.2.3) was increased under the strong heredity constraint. Additionally, the FNR for an associated squared term, $x_{c,3}^2$, was increased under both heredity constraints. By setting an intuition-based, exclusion variance prior, we reduced the average number of false negatives, consequently increasing the average number of false positives compared with a model tuned solely with regularization plots (results not shown). However in these simulations, the intuition-driven parameterization performed better in terms of the weighted average correct association percentage.

References.

[1] Liu, C., Ma, J., and Amos, C. I. (2015). Bayesian variable selection for hierarchical gene–environment and gene–gene interactions. *Human genetics* **134**, 1, 23–36.

[2] Ročková, V. and George, E. I. (2014). Emvs: The em approach to bayesian variable selection. *Journal of the American Statistical Association* **109**, 506, 828–846.

[3] Wang, M., Sun, X., and Lu, T. (2015). Bayesian structured variable selection in linear regression models. *Computational Statistics* **30**, 1, 205–229.

Table S1

Evaluation of EMVS's selection performance in simulated qualitative covariate settings: ***FPR***, *average false positive rate;* ***FNR***, *average false negative rate;* ***WACAP***, *weighted average correct association percentage.*

		Qualitative Terms Treated as Group			**Qualitative Terms Treated as Indpendent**		
Constraint	*Correlation*	*FPR*	*FNR*	*WACAP*	*FPR*	*FNR*	*WACAP*
	0	0.061	n/a	0.932	n/a	n/a	n/a
Model 3.1.1	0.4	0.061	n/a	0.931	n/a	n/a	n/a
	0.8	0.082	n/a	0.906	n/a	n/a	n/a
	0	0.066	0.080	0.910	0.050	0.080	0.920
Model 3.1.2	0.4	0.072	0.085	0.902	0.054	0.085	0.914
	0.8	0.097	0.138	0.847	0.078	0.138	0.860
	0	0.084	0.080	0.899	0.050	0.080	0.920
Model 3.1.3	0.4	0.091	0.085	0.891	0.054	0.085	0.914
	0.08	0.112	0.138	0.837	0.078	0.138	0.860
	0	0.047	0.105	0.902	0.045	0.249	0.803
Model 3.1.4	0.4	0.046	0.114	0.895	0.045	0.252	0.800
	0.8	0.074	0.152	0.852	0.073	0.278	0.758
	0	0.047	0.093	0.911	0.047	0.214	0.830
Model 3.1.5	0.4	0.044	0.099	0.909	0.045	0.222	0.825
	0.8	0.077	0.132	0.866	0.077	0.246	0.782
	0	0.046	0.057	0.940	0.045	0.168	0.867
Model 3.1.6	0.4	0.045	0.068	0.933	0.044	0.173	0.863
	0.8	0.079	0.106	0.888	0.079	0.202	0.818

TABLE S2

Evaluation of EMVS's selection performance in simulated interaction settings: ***FPR****, average false positive rate;* ***FNR****, average false negative rate;* ***WACAP****, weighted average correct association percentage*

		Model 3.2.1: True Model Follows Strong Heredity			**Model 3.2.2: True Model Follow Weak Heredity**			**Model 3.1.3: True Model Not Well Formulated**		
Constraint	*Correlation*	*FPR*	*FNR*	*WACAP*	*FPR*	*FNR*	*WACAP*	*FPR*	*FNR*	*WACAP*
Strong	0	0.057	0.024	0.946	0.080	0.008	0.942	0.097	0.005	0.929
	0.4	0.068	0.051	0.919	0.099	0.013	0.924	0.091	0.012	0.928
	0.8	0.145	0.157	0.778	0.170	0.069	0.823	0.160	0.069	0.825
Weak	0	0.100	0.023	0.921	0.101	0.020	0.921	0.126	0.005	0.911
	0.4	0.113	0.047	0.894	0.113	0.032	0.904	0.145	0.014	0.893
	0.8	0.191	0.150	0.754	0.192	0.144	0.754	0.225	0.066	0.781
iid	0	0.107	0.022	0.916	0.113	0.019	0.914	0.145	0.007	0.897
	0.4	0.128	0.046	0.883	0.128	0.031	0.894	0.146	0.020	0.888
	0.8	0.227	0.144	0.732	0.229	0.142	0.730	0.228	0.107	0.755

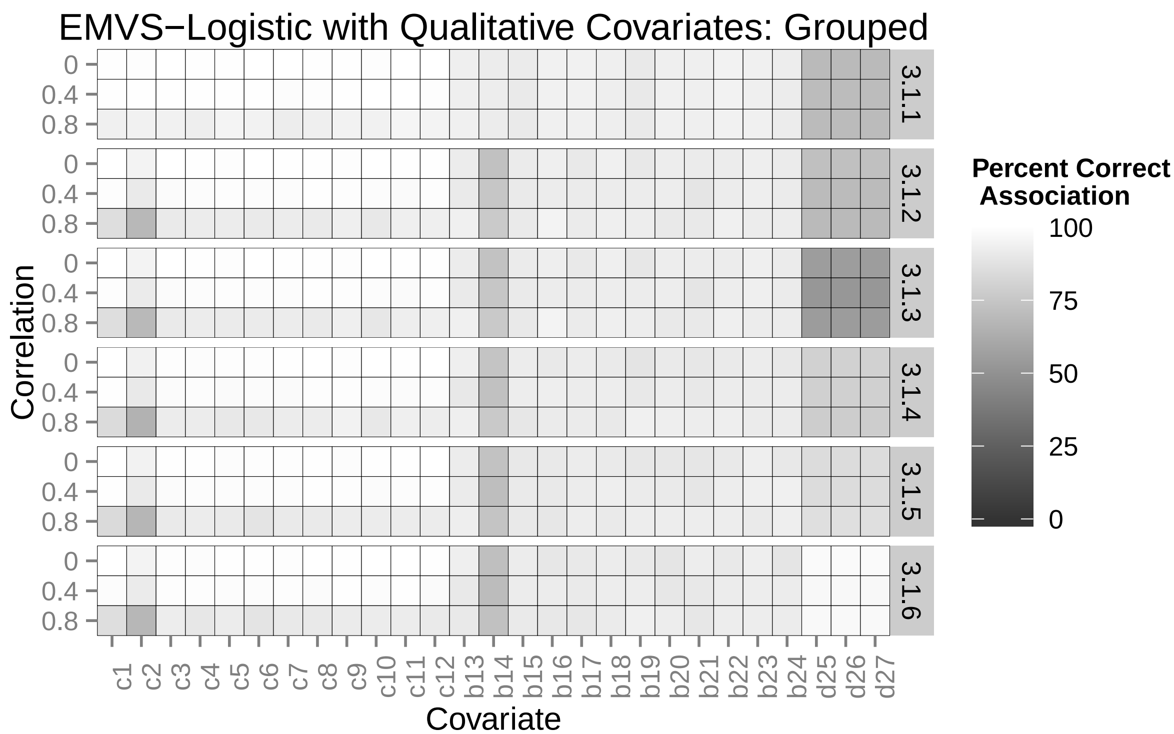

FIG S1. *Each box represents the correct association percentage for a covariate. The lighter the box, the better the model performed for that variable, averaged over all simulations.*

** Covariates existing in the true model

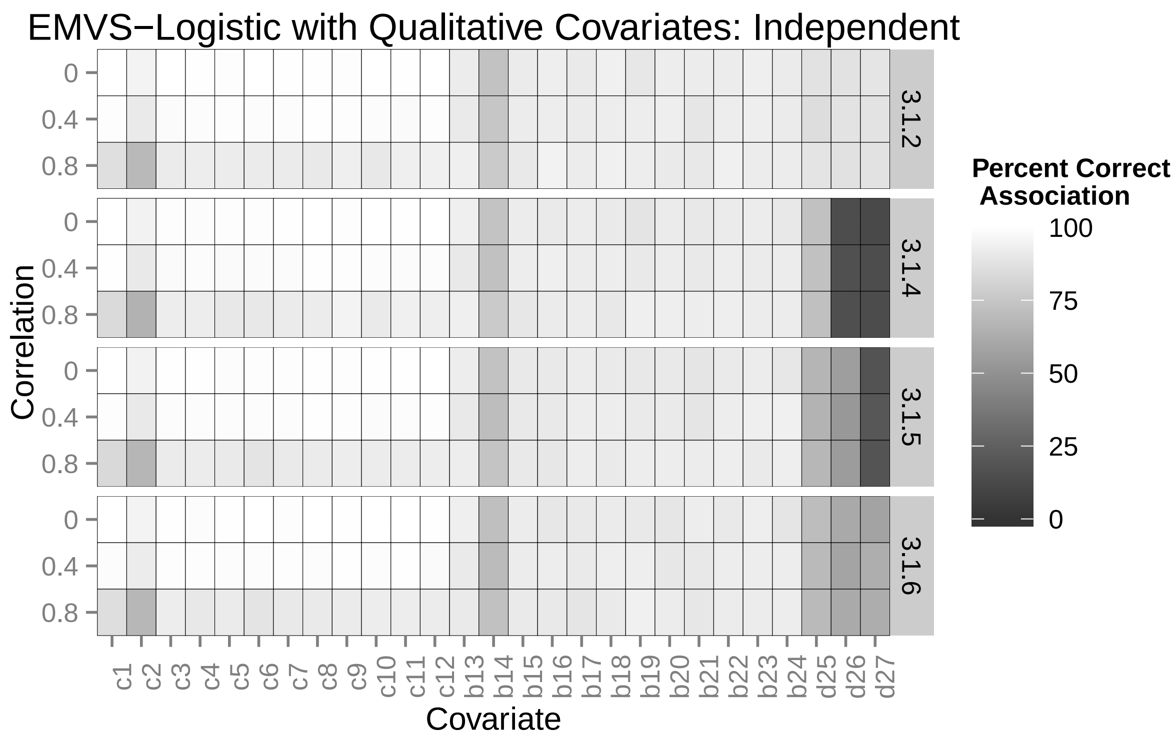

Fig S2. *Each box represents the correct association percentage for a covariate. The lighter the box, the better the model performed for that variable, averaged over all simulations.*

** Covariates existing in the true model

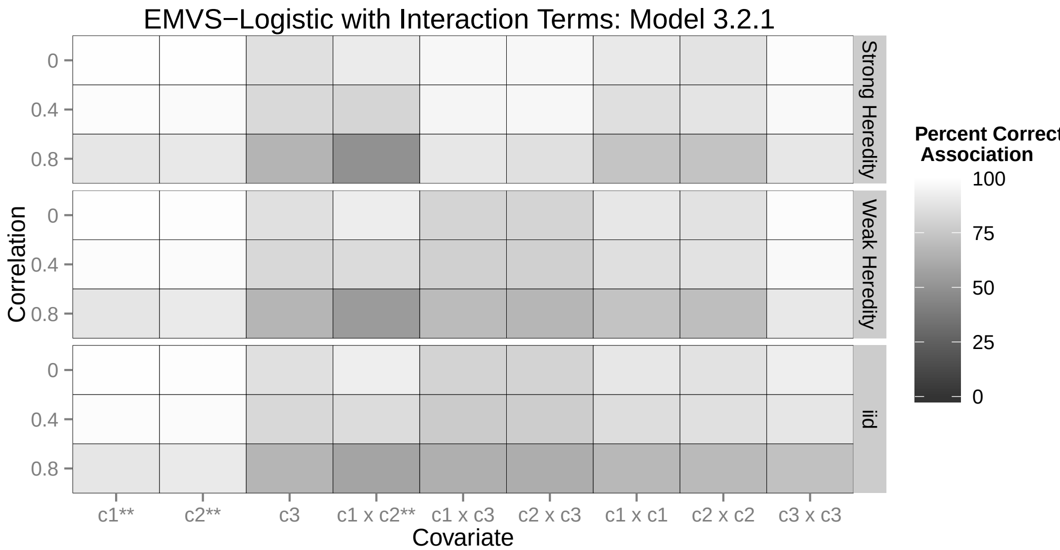

Fig S3. *Each box represents the correct association percentage for a covariate. The lighter the box, the better the model performed for that variable, averaged over all simulations. Whether or not a covariate should be included depends on the heredity constraint given. For example: if covariate c_1 is associated with the outcome, but c_2 is not, a strong constraint would exclude their interaction but a weak or no heredity constraint should include it.*

** Covariates existing in the true model

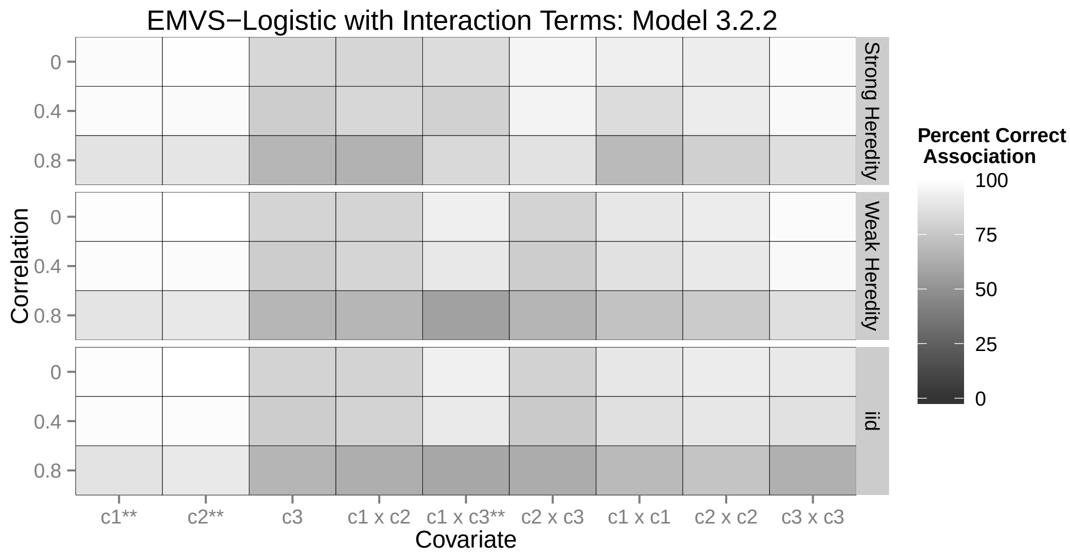

Fig S4. *Each box represents the correct association percentage for a covariate. The lighter the box, the better the model performed for that variable, averaged over all simulations. Whether or not a covariate should be included depends on the heredity constraint given. For example: if covariate c_1 is associated with the outcome, but c_2 is not, a strong constraint would exclude their interaction but a weak or no heredity constraint should include it.*

** Covariates existing in the true model

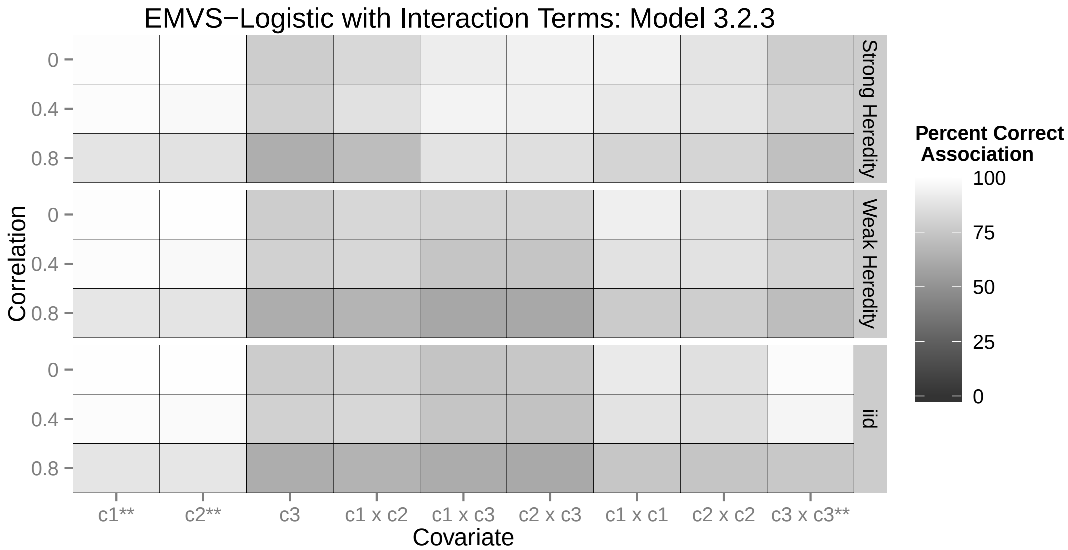

Fig S5. *Each box represents the correct association percentage for a covariate. The lighter the box, the better the model performed for that variable, averaged over all simulations. Whether or not a covariate should be included depends on the heredity constraint given. For example: if covariate c_1 is associated with the outcome, but c_2 is not, a strong constraint would exclude their interaction but a weak or no heredity constraint should include it.*

** Covariates existing in the true model

M. Koslovsky
Department of Biostatistics
UTHealth
1200 Pressler Street
Houston, TX 77030, USA
E-mail: matthew.d.koslovsky@uth.tmc.edu

M. Swartz
Department of Biostatistics
UTHealth
1200 Pressler Street
Houston, TX 77030, USA
E-mail: Michael.D.Swartz@uth.tmc.edu

L. Leon-Novelo
Department of Biostatistics
UTHealth
1200 Pressler Street
Houston, TX 77030, USA
E-mail: Luis.G.LeonNovelo@uth.tmc.edu

W. Chan
Department of Biostatistics
UTHealth
1200 Pressler Street
Houston, TX 77030, USA
E-mail: Wenyaw.Chan@uth.tmc.edu

A. Wilkinson
Department of Epidemiology
UTHealth
1616 Guadalupe Street
Austin, Texas 78701, USA
E-mail: Anna.V.Wilkinson@uth.tmc.edu