Web-based Supplementary Materials for "A semiparametric missing-data-induced intensity method for missing covariate data in individually matched case-control studies"

by

Mulugeta Gebregziabher and Bryan Langholz

Web Appendix A: Derivation of the CCA from the missing data induced model

This approach, which elaborates on the discussion in Section 2 of the paper, is based on choosing the appropriate (stratified) partial likelihood based on the induced model. So, consider ϕ completely unstructured in t and z. The induced model is

$$\begin{aligned}\lambda(t,x,z,m;\alpha(.),\beta_1,\beta_2;\phi(.,.)) &= \alpha(t)\exp(\beta_1 x(1-m)+\beta_2 z+\phi(t,z)m) \\ &= \begin{cases}\alpha_{(1-m)}(t)\exp(\beta_1 x+\beta_2 z) & \text{for non-missing} \\ \alpha_m(t,z)\exp(\beta_2 z) & \text{for missing}\end{cases} \qquad \text{(A.1)}\end{aligned}$$

where $\alpha_m(t,z)=\alpha(t)\phi(t,z)$. In forming a partial likelihood, it is natural to define separate baseline hazards for parts of the model that are unstructured functions of t. Thus, for the induced model (A.1), stratification is first on missing status, with non-missing in one stratum (with parametric terms for Z and X). Among the missing, stratification is also on z. Although the model for the missing subject includes a $\exp(\beta_2 z)$ factor, the likelihood contributions from the missing are "matched" on z so that such factors will cancel from numerator and denominator, i.e., subjects with missing X can be dropped from the analysis, equivalent to the CCA.

Empirical observations from our simulation studies also indicate that if we saturate the model with interactions of M with t and Z, the parameter estimates from MMI tend to be similar to CCA estimates.

Web Appendix B: Asymptotic information calculations

Expression for β asymptotic Fisher information from the SMI partial likelihood with dichotomous X

We start with a theorem for the consistency and asymptotic normality of and, in particular, the asymptotic information for, the SMI partial likelihood

THEOREM 1: *Let (N_i, Y_i, X_i, M_i), $i = 1, 2, \ldots$ be independent replicates of (N, Y, X, M) as defined in Section 2 with $\mathcal{G}$-intensity of the form $\lambda(t|\mathcal{G}) = \lambda_0(t)\exp(\beta_0 X(t)(1 - M(t)) + \eta_0 M(t))$ and let $U = \{1, \ldots, m\}$. For 1:$(m-1)$ individually matched data, assuming the conditional independence assumption for the missingness and Conditions 1-6 of Goldstein and Langholz (1992), $\hat{\beta}, \hat{\eta}$ are consistent and asymptotically normal with per subject asymptotic information given by*

$$\Gamma = E\left\{\int_0^\tau \frac{1}{m}\sum_{j\in U}\exp(\beta_0 X_j(t)(1 - M_j(t)) + \eta_0 M_j(t))v_U(t)p(t)\lambda_0(t)\,dt\right\},$$

where $p(t) = Pr(Y_i(t) = 1)$ and v_r is a 2×2 information matrix which is a function of set r with components,

$$\begin{aligned}
v_{\beta\beta,r}(t) &= \sum_{i\in U} P_i(t;\beta_0,\eta_0)[X_i(t)(1 - M_i(t)) - E_{\beta,r}(t;\beta_0,\eta_0)]^2 \\
v_{\beta\eta,r}(t) &= \sum_{i\in U} P_i(t;\beta_0,\eta_0)[X_i(t)(1 - M_i(t)) - E_{\beta,r}(t;\beta_0,\eta_0)][M_i(t) - E_{\eta,r}(t;\beta_0,\eta_0)] \\
v_{\eta\eta,r}(t) &= \sum_{i\in U} P_i(t;\beta_0,\eta_0)[M_i(t) - E_{\eta,r}(t;\beta_0,\eta_0)]^2
\end{aligned}$$

and

$$\begin{aligned}
P_{i,r}(t;\beta,\eta) &= \frac{Y_i(t)\exp(\beta X_i(t)(1 - M_i(t)) + \eta M_i(t))}{\sum_{j\in r} Y_j(t)\exp(\beta X_j(t)(1 - M_j(t)) + \eta M_j(t))} \\
P_i(t;\beta,\eta) &= \frac{Y_i(t)\exp(\beta X_i(t)(1 - M_i(t)) + \eta M_i(t))}{\sum_{j=1}^m Y_j(t)\exp(\beta X_j(t)(1 - M_j(t)) + \eta M_j(t))} \\
E_{\beta,r}(t;\beta,\eta) &= \sum_{j\in r} X_j(t)(1 - M_j(t))P_j(t;\beta,\eta) \\
E_{\eta,r}(t;\beta,\eta) &= \sum_{j\in r} M_j(t)P_j(t;\beta,\eta)
\end{aligned}$$

Further, the AFI for β, accounting for the estimation of η_0 is the inverse of the β, β corner of the inverse expected information matrix,

$$AFI_{SMI} = (\Gamma^{-1}_{\beta\beta})^{-1} = \Gamma_{\beta\beta} - \Gamma_{\beta\eta}[\Gamma_{\eta\eta}]^{-1}\Gamma_{\eta\beta}. \tag{A.2}$$

The proof of consistency and asymptotic normality, and the expression for the asymptotic information per subject Γ is a direct application of Theorem 3 of Goldstein and Langholz (1992) using the missing data induced intensity that results under simple random sampling of controls and the conditional independence assumption. We note that the conditions are standard and mild, the main requirement that the integral of the expected information exists. The expression for AFI_{SMI} is the Fisher information for β when estimating η_0 as a nuisance parameter and is based on a well known expression for computing a corner of the inverse of a matrix.

Special case used for efficiency comparisons

We refer to the conditions and notation given in section 2.1 of the main paper. For full (non-missing) data from $1{:}m-1$ data, the asymptotic information for β was derived in Goldstein & Langholz (1992), expression (22) and can be written as

$$AFI_{full} = \frac{1}{m} \sum_{(m_0,m_1):m_0+m_1=m} \binom{m}{m_0\ m_1} \pi^{m_1}(1-\pi)^{m_0} \frac{m_1 e^{\beta_0} m_0}{m_0 + m_1 e^{\beta_0}}.$$

where we are using $\binom{m}{m_0\ m_1}$ in a non-standard way, that is consistent with the multinomial coefficient, to represent the binomial coefficient $\frac{m!}{m_0! m_1!}$.

For CCA partial likelihood estimation analysis, contributions to the asymptotic information are only the subject of members in risk set with non-missing data. Thus, we view the problem computing the AFI for the sampling distribution from the "non-missing status stratum" when the control selection was simple randomly sampling from the entire risk set. Conditional on the number of non-missing in the sample $m - m_2$, the controls are a simple random sample of $m - m_2 - 1$ from the controls not missing X in the risk set. So, noting that

the probability of exposure conditional on non-missing is $\text{pr}(X=1|M=0)=\pi_1/(1-q)$

$$AFI(m-m_2)=\frac{1}{m-m_2}\sum_{(m_0,m_1):m_0+m_1=m-m_2}\binom{m-m_2}{m_0\ m_1}\left(\frac{\pi_0}{1-q}\right)^{m_0}\left(\frac{\pi_1}{1-q}\right)^{m_1}\frac{m_1e^{\beta_0}m_0}{m_0+m_1e^{\beta_0}}$$

Thus, applying Theorem 4 of Borgan, Goldstein, and Langholz (1995), the CCA information is the mean of $AFI(m-m_2)$ over the distribution of number of missing among the controls, m_2-1. With some calculation, the CCA AFI can be shown to be

$$\begin{aligned}
AFI_{CCA} &= \frac{1}{m}\sum_{m_2=0}^{m}\binom{m}{m_2\ m-m_2}q^{m_2}(1-q)^{m-1-m_2}\ AFI(m-m_2)\\
&= \frac{1}{m}\sum_{m_2=0}^{m}\binom{m}{m_2\ m-m_2}\sum_{(m_0,m_1):m_0+m_1=m-m_2}\binom{m-m_2}{m_0\ m_1}[(1-q)-\pi_1]^{m_0}\pi_1^{m_1}\frac{m_1e^{\beta_0}m_0}{m_0+m_1e^{\beta_0}}\\
&= \frac{1}{m}\sum_{(m_0,m_1,m_2):\sum_{j=0}^{2}m_j=m}\binom{m}{m_0\ m_1\ m_2}\pi_0^{m_0}\pi_1^{m_1}q^{m_2}\frac{m_1e^{\beta_0}m_0}{m_0+m_1e^{\beta_0}}.
\end{aligned}$$

Finally, for the SMI partial likelihood estimator, note that the (m_0,m_1,m_2) combinations have a multinomial$(\pi_0,\pi_1,q;m)$ distribution, that v_U only depends on (m_0,m_1,m_2) from the realization U and simplifying, Γ is proportional to

$$\frac{1}{m}\sum_{(m_0,m_1,m_2):\sum_{j=0}^{2}m_j=m}\binom{m}{m_0\ m_1\ m_2}\pi_0^{m_0}\pi_1^{m_1}q^{m_2}v(m_0,m_1,m_2)(m_0+m_1e^{\beta_0}+m_2e^{\eta_0})$$

where the components $v(m_0,m_1,m_2)$ are as described in the main paper. The AFI_{SMI} corresponding to β can be computed using (A.2).

$\Gamma^*(q,b)$ in the main paper yields the above expression for the values of q and b indicated.

Web Appendix C: Additional Simulation results comparing CCA, SMI, and MMI methods

Additional bias and efficiency simulation results when $n=100$ for 1:2 matching with no confounding ($\beta_Z=0$) with $\beta_X=0$ and $\beta_X=0.69$ and are given in Tables 1 to 2, respectively.

[Table 1 about here.]

[Table 2 about here.]

Web Appendix D: Simulation results comparing semi-parametric induced intensity estimators to other estimators

Using the simulation study described in Section 3 of the main paper, we compared the three induced-intensity estimators to the following missing data approaches:

Mid-Point Imputation (MPI). The mid-point of the predicted probability of $X = 1$ was calculated from a logistic regression of X on Z and D, by plugging $D = 1/2$ in the prediction model. The analysis was then implemented by replacing the missing X values by the predicted mid-points (Paik and Sacco, 2000).

Weighted Conditional Likelihood (WCL). Propensity scores were computed for X being observed (i.e. M=0) as a function of D and Z. A subject's "weight" was then defined as the log of the ratio of the propensity scores for the case and control for that subject. These weights were included as an offset term in the model (Lipsitz et al., 1998).

Multiple Imputation (MI). Five data sets with the missing X observations were imputed using a monotone logistic regression procedure that included Z and D. Then each data set was analyzed using methods described in ((Rubin, 1987)).

The results are tabulated in Tables 3 to 6 and are described in Section 3.2 of the main text.

[Table 3 about here.]

[Table 4 about here.]

[Table 5 about here.]

[Table 6 about here.]

Web Appendix E: : Comparison with other methods via data examples

Two other methods of interest that are not easy to implement using standard software are the Classic (SCC2000) and Bayesian (SCB2000) approaches of Satten and Caroll (Satten

and Carroll, 2000) as well as the Bayesian semi-parametric method (BSP2005) (Sinha et al., 2004; Sinha et al., 2005). Even though we need to undertake simulation studies to make real comparison, we provide analysis results that show overall how one gets different results for some data examples. We present results for the analysis of two data sets (the LA endometrial cancer study and the low-birth weight study that have been extensively studied these papers). Interestingly, in those examples, we clearly see the gain in efficiency when using modeled missing indicators. In practice, the simplicity of the MMI and the fact that it does not necessarily require MAR and leads to valid inference when missingness does not depend on case-control status is an advantage. The proposed method is very simple to use and can be performed using existing standard software without requiring any complex programming. It does not also require making model assumptions on the missing covariate.

[Table 7 about here.]

[Table 8 about here.]

References

Lipsitz, S., Parzen, M., and Ewell, M. (1998). Inference using conditional logistic regression with missing covariates. *Biometrics*, 54:295–303.

Paik, M. and Sacco, R. (2000). Matched case-control data analyses with missing covariates. *Applied Statistics*, 49:145–156.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons Inc, New York, NY.

Satten, G. and Carroll, R. (2000). Conditional and unconditional categorical regression models with missing covariates. *Biometrics*, 56:384–388.

Sinha, S., Mukherjee, B., and Ghosh, M. (2004). Bayesian semiparametric modeling for matched case-control studies with multiple disease states. *Biometrics*, 60:41–49.

Sinha, S., Mukherjee, B., Ghosh, M., Mallick, B., and Carroll, R. (2005). Semiparametric bayesian analysis of matched case-control studies with missing exposure. *JASA*, 100:591–601.

Table 1
Percent relative bias (PRB), Relative efficiency(REff),Power or Type I error rate and 95% CI (95CI) for $n = 100$ *and 1:2 matching, 50% missing. Based on 1000 trials.*

		Simulation Scenario: β_X=0 and β_Z=0							
		for $\exp(\hat{\beta}_X)$				for $\exp(\hat{\beta}_Z)$			
Method	Miss Type	PRB	Reff	Power	95CI	PRB	Reff	TypeIError	95CI
FULLadj		0.01	1.00	0.04	0.96	0.01	1.00	0.06	0.94
CCA	MCAR	0.04	0.48	0.05	0.95	0.01	0.50	0.05	0.95
	MAR(Z)	0.02	0.50	0.05	0.95	0	0.50	0.07	0.93
	MAR(D)	0.04	0.39	0.03	0.97	0.02	0.40	0.05	0.95
	NI(X)	0.02	0.49	0.04	0.96	0.02	0.57	0.05	0.95
	NI(X,Z)	0.04	0.49	0.05	0.95	0.02	0.53	0.07	0.93
SMI	MCAR	0.02	0.72	0.06	0.94	0.01	1.00	0.06	0.94
	MAR(Z)	0	0.72	0.04	0.96	0.01	1.00	0.06	0.94
	MAR(D)	0.05	0.57	0.03	0.97	0.01	0.89	0.05	0.95
	NI(X)	0	0.67	0.05	0.95	0.01	1.00	0.05	0.95
	NI(X,Z)	0.01	0.70	0.05	0.95	0.01	1.00	0.06	0.94
MMI	MCAR	0.03	0.69	0.05	0.95	0	0.71	0.07	0.93
	MAR(Z)	0	0.70	0.04	0.96	0	0.71	0.07	0.93
	MAR(D)	0.06	0.54	0.04	0.96	0.01	0.57	0.05	0.95
	NI(X)	0.01	0.66	0.06	0.94	0.02	0.77	0.05	0.95
	NI(X,Z)	0.01	0.67	0.04	0.96	0	0.74	0.06	0.94
MI	MCAR	0.03	0.65	0.06	0.94	0.01	0.94	0.05	0.95
	MAR(Z)	0.01	0.67	0.06	0.94	0.01	0.94	0.06	0.94
	MAR(D)	0.04	0.53	0.06	0.94	0.02	0.89	0.04	0.96
	NI(X)	0.02	0.62	0.07	0.93	0.01	0.94	0.05	0.95
	NI(X,Z)	0.01	0.65	0.06	0.94	0.01	0.94	0.06	0.94

Table 2

Percent relative bias (PRB), Relative efficiency(REff),Power or Type I error rate and 95% CI (95CI) for $n = 100$ and 1:2 matching, 50% missing. Based on 1000 trials.

		Simulation Scenario: β_X=0.69 and β_Z=0							
		for $\exp(\hat{\beta}_X)$				for $\exp(\hat{\beta}_Z)$			
Method	Miss Type	PRB	Reff	Power	95CI	PRB	Reff	TypeIError	95CI
FULLadj		4.3	1.00	0.69	0.96	0.01	1.00	0.05	0.95
CCA	MCAR	18.8	-	0.26	0.96	0.02	0.49	0.04	0.96
	MAR(Z)	20.3	-	0.25	0.94	0.01	0.49	0.05	0.95
	MAR(D)	26.1	-	0.21	0.96	0.00	0.40	0.03	0.97
	NI(X)	13.0	-	0.24	0.96	0.00	0.56	0.05	0.95
	NI(X,Z)	17.4	-	0.27	0.97	0.01	0.53	0.05	0.95
SMI	MCAR	4.3	0.75	0.38	0.96	0.10	1.00	0.08	0.92
	MAR(Z)	2.9	0.75	0.40	0.94	0.10	1.00	0.08	0.92
	MAR(D)	2.9	0.61	0.29	0.96	0.11	0.90	0.06	0.94
	NI(X)	7.2	-	0.34	0.95	0.10	1.00	0.08	0.92
	NI(X,Z)	4.3	0.73	0.38	0.96	0.10	1.00	0.08	0.92
MMI	MCAR	4.3	0.71	0.42	0.95	0.02	0.72	0.05	0.95
	MAR(Z)	5.8	0.71	0.41	0.94	0.01	0.72	0.06	0.94
	MAR(D)	8.7	-	0.32	0.96	0.00	0.56	0.03	0.97
	NI(X)	2.9	0.68	0.37	0.96	0.00	0.75	0.03	0.97
	NI(X,Z)	7.2	-	0.40	0.96	0.00	0.75	0.04	0.96
MI	MCAR	5.8	0.65	0.39	0.94	0.01	0.90	0.05	0.95
	MAR(Z)	7.2	-	0.39	0.93	0.00	0.90	0.06	0.94
	MAR(D)	11.6	-	0.33	0.92	0.01	0.82	0.05	0.95
	NI(X)	4.3	0.64	0.36	0.95	0.04	0.95	0.04	0.96
	NI(X,Z)	7.2	-	0.39	0.93	0.02	0.95	0.04	0.96

Table 3

Percent relative bias in $\exp(\hat{\beta}_X)$ in complete case analysis (CCA), single missing indicator(SMI), modeled missing indicator (MMI), weighted conditional likelihood (WCL), mid-point imputation(MPI) and multiple imputation (MI) for simulation scenarios, no versus strong-negative confounding, 1:1 design, missing data proportion =20% and 50%, number of case-control sets=400, $\exp(\beta_X)$=2,$\exp(\beta_Z)$=1.42, $pr(X = 1) = 0.5$. Based on 1000 trials.

pr($M = 1$)	Confounding	Missing type	CCA	SMI	MMI	WCL	MPI	MI
50%	no	MCAR	1.6	0.9	1.1	1.6	1.0	0.6
		MAR(Z)	1.8	0.7	0.9	1.8	0.8	0.5
		MAR(D)	3.0	2.3	2.6	3.0	0.3	1.8
		NI(X)	4.4	1.4	1.7	4.4	1.8	1.5
		NI(X,Z)	2.7	1.8	2.1	2.7	2.1	1.8
50%	strong	MCAR	2.6	-9.6	0.2	2.6	0.0	1.3
		MAR(Z)	2.0	-7.9	1.1	2.0	1.0	1.7
		MAR(D)	4.7	-8.2	1.9	4.7	1.3	1.9
		NI(X)	3.7	-5.4	1.0	3.7	-0.2	1.2
		NI(X,Z)	3.9	-3.6	1.0	3.9	0.8	2.2
20%	no	MCAR	0.6	0.4	0.5	0.6	0.3	0.4
		MAR(Z)	0.5	0.5	0.6	0.5	0.4	0.9
		MAR(D)	0.2	0.2	0.2	0.2	-0.1	0.4
		NI(X)	0.6	0.4	0.5	0.6	0.4	0.5
		NI(X,Z)	0.3	0.5	0.6	0.3	0.5	0.6
20%	strong	MCAR	1.7	-3.3	1.2	1.7	1.1	1.2
		MAR(Z)	2.4	-3.4	1.4	2.4	1.2	1.6
		MAR(D)	1.7	-1.3	1.5	1.7	1.4	1.5
		NI(X)	0.4	-1.7	0.5	0.4	0.1	1.1
		NI(X,Z)	1.2	0.9	1.0	1.1	0.8	1.5

Table 4

Percent relative bias in exp($\hat{\beta}_Z$) in complete case analysis (CCA), single missing indicator(SMI), modeled missing indicator (MMI), weighted conditional likelihood (WCL),mid-point imputation(MPI) and multiple imputation (MI) for simulation scenarios, no versus strong-negative confounding, 1:1 design, missing data proportion =20% and 50%, number of case-control sets=400, exp(β_X)=2,exp(β_Z)=1.42, pr($X = 1$) = 0.5. Based on 1000 trials.

Pr(M=1)	Confounding	Missing type	CCA	SMI	MMI	WCL	MPI	MI
50%	no	MCAR	1.9	0.8	0.7	1.9	0.8	0.5
		MAR(Z)	2.1	0.8	0.5	2.1	0.8	0.5
		MAR(D)	2.4	0.6	0.5	2.9	0.7	0.6
		NI(X)	2.7	0.9	0.9	2.7	0.8	0.6
		NI(X,Z)	1.9	2.1	0.6	-0.1	-0.8	-1.2
50%	strong	MCAR	1.5	-11	0.7	1.5	0.6	0.1
		MAR(Z)	1.0	-12	0.9	1.0	0.8	0.2
		MAR(D)	2.8	-11	1.0	2.8	1.0	0.3
		NI(X)	0.0	-7.5	0.0	-4.1	-3.1	-3.5
		NI(X,Z)	0.3	-7.4	-0.1	-5.6	-4.3	-4.7
20%	no	MCAR	1.5	0.8	1.1	1.4	0.7	0.4
		MAR(Z)	1.2	0.9	1.1	1.1	0.7	0.4
		MAR(D)	0.9	0.6	0.7	1.1	0.7	0.4
		NI(X)	1.3	0.8	1.1	1.3	0.7	0.4
		NI(X,Z)	1.2	1.1	0.9	-0.4	-1.1	-1.3
20%	strong	MCAR	1.0	-4.5	0.9	1.0	0.8	0.1
		MAR(Z)	0.9	-5.3	0.7	0.8	0.8	0.2
		MAR(D)	0.7	-2.7	0.6	0.8	0.8	0.1
		NI(X)	0.2	-2.1	0.4	-2.1	-0.2	-0.7
		NI(X,Z)	0.5	-1.8	0.7	-3.1	-1.3	-1.8

Table 5

Relative efficiency in $\hat{\beta}_X$ in complete case analysis (CCA), single missing indicator(SMI), modeled missing indicator (MMI), weighted conditional likelihood (WCL), mid-point imputation(MPI) and multiple imputation (MI) for simulation scenarios, no versus strong-negative confounding, 1:1 design, missing data proportion =20% and 50%, number of case-control sets=400, $exp(\beta_X)$=2,$exp(\beta_Z)$=1.42, $pr(X = 1) = 0.5$. Based on 1000 trials.

Pr(M=1)	Confounding	Missing type	CCA	SMI	MMI	WCL	MPI	MI
50%	no	MCAR	0.48	0.71	0.71	0.48	0.71	0.67
		MAR(Z)	0.48	0.77	0.77	0.48	0.77	0.71
		MAR(D)	0.40	0.53	0.53	0.40	0.63	0.56
		NI(X)	0.44	0.67	0.67	0.44	0.67	0.50
		NI(X,Z)	0.46	0.67	0.67	0.46	0.67	0.67
50%	strong	MCAR	0.50	-	0.71	0.50	0.71	0.59
		MAR(Z)	0.53	-	0.71	0.53	0.71	0.77
		MAR(D)	0.42	-	0.59	0.42	0.71	0.63
		NI(X)	0.42	-	0.63	0.42	0.67	0.59
		NI(X,Z)	0.48	-	0.71	0.48	0.71	0.71
20%	no	MCAR	0.91	1.00	1.00	0.91	1.0	1.00
		MAR(Z)	0.91	1.00	1.00	0.91	1.0	1.00
		MAR(D)	0.91	1.00	1.00	0.91	1.0	1.00
		NI(X)	0.91	1.00	1.00	0.91	1.0	1.00
		NI(X,Z)	0.91	1.00	1.00	0.91	1.0	1.00
20%	strong	MCAR	0.83	-	0.91	0.83	0.91	0.91
		MAR(Z)	0.83	-	0.91	0.83	0.91	0.91
		MAR(D)	0.77	-	0.83	0.77	0.91	0.91
		NI(X)	0.83	-	0.91	0.83	0.91	0.91
		NI(X,Z)	0.83	-	0.91	0.83	0.91	0.91

Table 6

Relative efficiency in $\hat{\beta}_Z$ in complete case analysis (CCA), single missing indicator(SMI), modeled missing indicator (MMI), weighted conditional likelihood (WCL), mid-point imputation(MPI) and multiple imputation (MI) for simulation scenarios, no versus strong-negative confounding, 1:1 design, missing data proportion =20% and 50%, number of case-control sets=400, $exp(\beta_X)$=2,$exp(\beta_Z)$=1.42, $pr(X = 1) = 0.5$. Based on 1000 trials.

Pr(M=1)	Confounding	Missing type	CCA	SMI	MMI	WCL	MPI	MI
50%	no	MCAR	0.46	1.00	0.71	0.46	1.0	0.91
		MAR(Z)	0.42	1.00	0.59	0.42	1.0	0.91
		MAR(D)	0.42	0.77	0.53	0.42	1.0	0.91
		NI(X)	0.48	1.00	0.71	0.48	1.0	0.91
		NI(X,Z)	0.44	1.00	0.63	0.44	1.0	0.91
50%	strong	MCAR	0.53	-	0.71	0.53	0.91	1.00
		MAR(Z)	0.46	-	0.67	0.46	0.91	0.91
		MAR(D)	0.40	-	0.56	0.40	0.83	0.83
		NI(X)	0.56	-	0.71	0.56	-	-
		NI(X,Z)	0.50	-	0.71	0.50	-	-
20%	no	MCAR	0.91	1.00	1.00	0.91	1.0	1.00
		MAR(Z)	0.91	1.00	1.00	0.91	1.0	1.00
		MAR(D)	0.91	1.00	1.00	0.91	1.0	1.00
		NI(X)	0.91	1.00	1.00	0.91	1.0	1.00
		NI(X,Z)	0.91	1.00	1.00	0.91	1.0	1.00
20%	strong	MCAR	0.77	-	0.91	0.77	1.0	1.00
		MAR(Z)	0.77	-	0.91	0.77	1.0	1.00
		MAR(D)	0.77	-	0.83	0.77	1.0	1.00
		NI(X)	0.77	-	0.91	0.77	-	1.00
		NI(X,Z)	0.77	-	0.91	0.77	-	-

Table 7
Analysis of the risk of endometrial cancer in relation to obesity using five different missing data methods. Los Angeles, 1971-1975.

	Missing data methods[1]				
Variable	CCA	SMI	MMI	MI	SCC(2000)
OB	1.44(1.39)	1.46(1.07)	1.55(1.08)	1.79(1.40)	1.41(1.41)
GALL	3.25(1.28)	2.92(1.04)	3.15(1.10)	3.09(1.19)	3.25(1.19)
EST	3.51(1.40)	3.41(0.93)	3.47(0.93)	3.68(1.37)	3.47(1.37)
OBxGALL	-0.14(0.92)	0.24(0.81)	0.05(0.87)	0.08(0.90)	-0.19(0.89)
OBxEST	-1.10(1.39)	-0.94(1.04)	-0.97(1.04)	-1.20(1.40)	-0.88(1.39)
GALLxEST	-2.26(1.17)	-2.19(1.03)	-2.26(1.02)	-2.26(1.07)	-2.32(1.06)

CCA=complete case analysis
SMI=single missing indicator
MMI=modeled missing indicator with time and missing indicator interaction
MI=multiple imputation
SCC(2000)= Conditional logistic method, Satten and Carrol, 2000
categories are percentiles, numbers in table are log odds ratio with their standard errors

Table 8
Analysis of the low birth weight data in Hosmer and Lemeshow 2000.

	Missing data analysis results					
Method	$\hat{\beta}_1$	SE($\hat{\beta}_1$)	$\hat{\beta}_2$	SE($\hat{\beta}_2$)	$\hat{\beta}_3$	SE($\hat{\beta}_3$)
FULL	0.861	0.454	0.853	0.513	-1.128	1.359
CCA	0.923	0.547	0.941	0.625	-1.340	1.581
SMI	0.870	0.514	0.872	0.524	-1.037	1.391
MMI1	0.847	0.519	1.077	0.597	-1.017	1.385
MMI2	0.869	0.516	0.876	0.525	-1.220	1.624
WCL	0.925	0.547	0.887	0.625	-1.459	1.580
MPI	0.741	0.503	0.855	0.507	-1.270	1.347
MI	0.979	0.553	0.824	0.519	-1.140	1.391
	Missing data analysis as in Sinha et al 2005					
BSP(2005)	1.190	0.820	0.960	0.550	-1.120	1.270
SCB(2005)	0.650	0.590	0.860	0.530	-1.220	1.240
CCA	0.650	0.530	0.880	0.600	-1.060	1.620
	Full data analysis as in Sinha et al 2005					
BSP(2005)	1.210	0.560	0.910	0.490	-1.170	1.260
SCB(2005)	0.960	0.460	0.790	0.520	-1.340	1.250
CLR	0.860	0.450	0.850	0.510	-1.130	1.360

β_1 for maternal smoking, β_2 for UI and β_3 for LWT
CCA=complete case analysis adjusted for LWT and UI
FULL=Full data analysis using conditional logistic (CLR) adjusting for LWT and UI
SMI=single missing indicator, MMI=modeled missing indicator
MMI1=MMI with interaction between missing indicator and UI
MMI2=MMI with interaction between missing indicator and LWT
MPI=mid-point imputation
WCL= weighted conditional likelihood, MI=multiple imputation
BSP(2005)= Bayesian Semi-Parametric method, Sinha et al 2005
SCB(2005) is a parametric Bayes method, Sinha et al 2005
CLR is a Full data Bayes based conditional logisitc, Sinha et al 2005