

# NON-RESPONSE MODELS FOR THE ANALYSIS OF NON-MONOTONE NON-IGNORABLE MISSING DATA

JAMES M. ROBINS

*Departments of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, MA 02115, U.S.A.*

## SUMMARY

I introduce a new class of non-ignorable non-monotone missing data models. These models are useful for investigating the sensitivity of one's estimates to untestable assumptions about the missing data process. I use the new models to analyse data from a case-control study of the effect of radiation on breast cancer.

## 1. INTRODUCTION

This paper proposes a new class of non-ignorable missing data processes, the permutation missingness (PM) processes, and new analytic methods, based on this class, for studies with non-ignorable non-monotone missing data. The new methods are used to analyse data provided by the conference organizers from a case-control study of the effect of radiation therapy for an initial tumour on the development of a second breast cancer.<sup>1,2</sup> Given an ordering (permutation) of the study variables, a PM process specifies that the probability that the  $k$ th variable is observed may depend on and only on the values of variables 1 to  $k - 1$  (observed or not) and the observed values of variables  $k + 1$  onwards. The PM processes are a natural non-ignorable generalization of missing-at-random (MAR) processes. In fact, we will show that both the class of MAR (that is, ignorable) processes and the class of PM processes are special cases of an even larger class of processes – the group permutation missingness (GPM) processes. Indeed, we show in Section 7 that any GPM process can be represented as a finite sequence of nested MAR processes. Interestingly, a sequence of nested MAR processes need not itself be MAR when the sequence length exceeds one.

We will show that the class of PM processes, like the class of the ignorable missing at random processes, have the highly desirable property of being non-parametric saturated (NPS). Roughly speaking, letting  $L$ ,  $L_{\text{obs}}$ , and  $R$  denote the complete data vector, the observed data vector, and the vector of missing data indicators, we define a class of missing data processes to be NPS if, for each possible distribution  $f(R, L_{\text{obs}})$  of the observed data  $(R, L_{\text{obs}})$ , there exists a unique missing data process  $f^\Delta(R|L)$  in the class and a unique law  $f^\Delta(L)$  for the complete data  $L$  such that  $f(R, L_{\text{obs}})$  is the marginal distribution of  $(R, L_{\text{obs}})$  under the joint law  $f^\Delta(R, L) = f^\Delta(R|L)f^\Delta(L)$ . We argue that our analysis based on NPS classes is quite useful for investigating the sensitivity of one's estimate to untestable assumptions about the missing data process.

When  $L = (L_1, \dots, L_K)$  is a vector of discrete variables  $L_k$ , a necessary condition for a class to be NPS can be expressed in a more familiar way. Let  $d_i$  and  $d_0$  be the number of parameters required (that is, degrees of freedom available) to parameterize non-parametric models for the

distribution of  $L$  and of  $(R, L_{\text{obs}})$ , respectively. For example, if  $L = (L_1, L_2)'$  with  $L_1$  dichotomous and  $L_2$  trichotomous, then  $d_\ell = 2 \times 3 - 1 = 5$ . Now let  $f(R|L; \gamma)$  be a model for the law of  $R$  given  $L$  with the parameter vector  $\gamma$  of dimension  $d_r$ . Then the class of missing data processes represented by the model  $f(R|L; \gamma)$  can be an NPS class only if  $d_r = d_0 - d_\ell$ . If  $d_0 - d_\ell > d_r$ , we can always find a law  $f(R, L_{\text{obs}})$  such that there is no value of  $\gamma$  and no law  $f(L)$  for which  $f(R, L_{\text{obs}})$  is the marginal distribution of  $(R, L_{\text{obs}})$  corresponding to the joint law  $f(R|L; \gamma)f(L)$ . On the other hand, if  $d_0 - d_\ell < d_r$ , we can always find a law  $f(R, L_{\text{obs}})$  such that there are several choices of  $\gamma$  and  $f(L)$  whose marginal is  $f(R, L_{\text{obs}})$  of the observed data.

Baker *et al.*<sup>3</sup> previously proposed classes of NPS missing data processes for non-monotone missing data in the special case in which  $L = (L_1, L_2)'$  is a discrete bivariate vector (see Tables II(d)–II(f) of reference 3). In Appendix III, it is shown that the Baker *et al.*<sup>3</sup> classes differ from the GPM classes.

In contrast to Baker *et al.*'s<sup>3</sup> NPS classes, the GPM classes allow  $L$  to be of any dimension with both discrete and continuous components. If  $L$  is high-dimensional and/or has continuous components, then, due to the curse of dimensionality (that is, due to sparse data), it will usually be necessary to specify a parametric model  $f(R|L; \gamma)$  for the non-ignorable density  $f(R|L)$ . (However, see references 4 and 5 for some interesting but quite restrictive models where this is not the case.) We refer to  $f(R|L; \gamma)$  as a GPM model if, for each parameter value  $\gamma$ ,  $f(R|L; \gamma)$  is a GPM process. In Section 4, we show that for a GPM model, consistent asymptotically normal (CAN) estimators  $\hat{\gamma}$  of the true value  $\gamma^*$  of  $\gamma$  can be obtained without having to specify a model for the marginal distribution of  $L$ .

To understand the importance of this last point, suppose  $L = (X', Y)'$  where the outcome  $Y$  is dichotomous, the vector  $X$  of regressors is multivariate with continuous components, and we wish to estimate the parameter vector  $\alpha$  of the logistic model  $\text{pr}[Y = 1 | X] = \alpha X$ . As discussed in Section 3 and in references 7 and 8, p. 118, if a CAN estimator  $\hat{\gamma}$  of  $\gamma^*$  is available, then we can use inverse probability (of avoiding missing data) weighted logistic regression to estimate  $\alpha$  without specifying a parametric model for the marginal distribution of the regressors  $X$ , a distribution of no scientific interest. It follows that if  $f(R|L; \gamma)$  is a GPM model, we can estimate  $\alpha$  using inverse probability weighted logistic regression without having to model the marginal distribution of the regressors.

In contrast, the standard approach to making inferences about the logistic regression parameter  $\alpha$  with non-ignorable missing data is to use the EM algorithm to fit, by maximum likelihood, a parametric model  $f(Y|X; \alpha)f(X; \theta)f(R|L; \gamma)$  to the data  $(R, L_{\text{obs}})$ .<sup>9–13</sup> This parametric approach has the serious drawback that if the model  $f(X; \theta)$  for the marginal distribution of  $X$  is misspecified, the MLE of  $\alpha$  will be inconsistent even if the missingness model  $f(R|L; \gamma)$  is correctly specified. Recently, Robins *et al.*<sup>8</sup> (p. 118) and Rotnitzky and Robins<sup>7</sup> have described an alternative approach in which one jointly estimates the logistic parameter  $\alpha$  and the parameter  $\gamma$  of the missingness model by using augmented inverse probability weighted estimating equations without having to model the marginal distribution of the regressors. Their approach applies to any parametric missing data model, whether ignorable or non-ignorable, NPS or non-NPS, GPM or non-GPM. However, a drawback to their method is that no off-the-shelf software packages are as yet available that would allow the statistically unsophisticated user to implement the method. In contrast, in Section 4, it is shown that when  $f(R|L; \gamma)$  is a PM model,  $\hat{\alpha}$  and  $\hat{\gamma}$  can both be estimated using a canned logistic regression package that allows user-supplied weights.

Unfortunately, the class of GPM processes itself has a serious drawback that prevents it from serving as an all-purpose class of missing data processes with which to model non-ignorable missingness. Specifically, GPM processes do not allow the probability that a particular variable

is missing to depend on the value of that variable, although this probability can depend on the values of other unobserved variables. Since, based on subject matter considerations, one frequently wants to consider non-ignorable processes where the probability that a variable is missing does depend on the (possibly unobserved) value of that variable, a thorough sensitivity analysis must include missing data models other than GPM models. If, as described above, the augmented inverse probability weighted estimating equations described in references 7 and 8 are used to fit these other missing data models, one can still avoid having to specify a model for the marginal distribution of the regressors.

## 2. THE DATA AND FULL DATA MODEL

Data on 529 second breast cancer cases and 529 controls (drawn from women whose initial breast cancer was not followed by the development of a second cancer) were obtained from the conference organizers. Details of data collection and other substantive considerations are described in references 1 and 2. In this section we shall analyse the association of radiation therapy with second breast cancer while adjusting for the potential confounding variables of family history and the weight/height<sup>2</sup> by fitting the logistic model

$$\text{pr}[Y = 1|X] = \text{expit}(\alpha' X) \quad (1)$$

where  $\text{expit}(z) = e^z / \{1 + e^z\}$ ,  $Y$  is the second breast cancer (case-control) indicator,  $X = (X_0, X_1, X_2, X_3)'$ ,  $\alpha' = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)'$ . Here  $X_0 \equiv 1$ ;  $X_1 = 1$  if a subject received radiation therapy,  $X_1 = 0$  otherwise;  $X_2 = 1$  if a subject had a family history of breast cancer,  $X_2 = 0$  otherwise;  $X_3$  is an ordinal variable taking the values 1, 2, 3, or 4 as  $10^4 \times \text{weight/height}^2$  was  $< 20, 20-25, 25-30, > 30$ . The outcome  $Y$  is available on all subjects. However,  $X_1$  was missing on 3 per cent of the subjects,  $X_2$  on 49 per cent, and  $X_3$  on 43 per cent. Let  $R_j = 1$  if  $X_j$  was observed, and  $R_j = 0$  otherwise. Table I gives the joint distribution of  $R_j$ 's. Reading from the last row of Table I, we see that only 33.6 per cent of the subjects had complete covariate data. Furthermore, the missingness is non-monotone.

Table II gives estimates of the regression coefficients,  $\alpha_1, \alpha_2, \alpha_3$  using different estimation methods. Row 1 is based on logistic regression restricted to complete cases. The complete case estimator of row 1 will be biased when the missingness process depends on both the covariates  $X$  and the outcome  $Y$ . The estimate in row 2 was obtained by Robins and Gill<sup>14</sup> using an ignorable randomized monotone missingness (RMM) model for the missing data process. It is included here for purposes of comparison with the results obtained using the non-ignorable PM models.

Results in rows 3 and 4 were obtained under non-ignorable PM models. As discussed below, the assumptions concerning the missingness process encoded in a PM model depends on the ordering (that is, a permutation) of the variables. In this data set, the estimates for all parameters in which variable 2 preceded variable 3 were essentially identical. Likewise, the estimates were nearly identical for all permutations in which variable 3 preceded variable 2. (This reflects the fact that only 3 per cent of the data were missing on variable  $X_1$ , and thus it did not matter where it was placed in the permutation ordering.)

## 3. INVERSE-PROBABILITY-WEIGHTED ESTIMATING EQUATIONS

The RMM and PM estimates in Table II are based on solving an inverse-probability-weighted estimating equation. To describe this estimating equation, let  $L = (X', Y)$  represent the full data. Let  $R = (R_1, R_2, R_3, R_4)'$  be the vector of missing data indicators. (Note in our data  $R_4$  is always 1, since data on  $Y$  is never missing. Also references to  $X_0 \equiv 1$  and  $R_0 \equiv 1$  are suppressed.) Let  $L_{(r)}$

Table I. Empirical distribution of  $(R_1, R_2, R_3)$ 

$R_1$	$R_2$	$R_3$	%
0	0	0	1.2
1	0	0	27.4
0	1	0	0.9
1	1	0	14.8
0	0	1	0.2
1	0	1	21.3
0	1	1	0.38
1	1	1	33.6

Table II. Estimation of  $\alpha$  under different missingness models

Method of estimation	$\hat{\alpha}_1$ (Rad)	$\hat{\alpha}_2$ (FH <sub>x</sub> )	$\hat{\alpha}_3$ (Wt/Ht)
Complete case	0.24	0.43	0.31
RMM	0.19	0.44	0.31
PM 123*	0.15	0.48	0.31
PM 132†	0.25	0.48	0.34

\* All permutations with 2 before 3 gave similar results

† All permutations with 3 before 2 gave similar results

be the observed components of  $L$  when  $R = r$ . For example, if  $R = (1, 0, 1, 1)$ ,  $L_{(r)} = (X_1, X_3, Y)$ . Little and Rubin<sup>9</sup> refer to  $L_{(r)}$  as  $L_{\text{obs}}$ . Let

$$\pi(r, L) \equiv \text{pr}[R = r|L]$$

be the conditional probability of observing the components  $R = r$ . Let  $\mathbf{1}$  denote a vector with each component equal to one, so  $\pi(\mathbf{1}, L)$  is the conditional probability of observing complete data given  $L$ . Throughout we assume

$$\pi(\mathbf{1}, L) > \sigma > 0 \quad (2)$$

that is, the conditional probability of having complete data is bounded away from zero, where the bound  $\sigma$  is a fixed positive constant, say, 0.1. Assumption (2) ensures that the asymptotic variances of our estimators are finite. Let  $\delta$  denote the indicator variable for complete data, so  $\delta = 1$  if and only if  $R = \mathbf{1}$ . The solution  $\tilde{\alpha}$  to the inverse probability weighted estimating equation

$$0 = \sum_i U_i(\alpha) \equiv \sum_i \{\delta_i / \pi(\mathbf{1}, L_i)\} \{Y_i - \text{expit}(\alpha' X_i)\} X_i$$

will be a consistent asymptotically normal estimator of  $\alpha_0$ .  $\tilde{\alpha}$  can be obtained by fitting model (1) to the complete cases ( $\delta = 1$ ) using a canned logistic regression program that allows for individual weights (that is,  $\pi(\mathbf{1}, L_i)^{-1}$ ).  $\sum_i U_i(\alpha)$  is an unbiased estimating function for  $\alpha$  since, like the Horvitz–Thompson estimator,<sup>15</sup> it weights each subject with complete data by the inverse of the conditional probability of having complete data. When, as will be the case in practice,  $\pi(\mathbf{1}, L_i)$  is unknown and we replace it by an  $n^{1/2}$ -consistent estimate  $\hat{\pi}(\mathbf{1}, L_i)$  obtained under a parametric model for the missingness process  $\pi(r, L) = \text{pr}[R = r|L]$ , the resulting estimator  $\hat{\alpha}$  will remain consistent and asymptotically normal for  $\alpha$ . In row 2 of Table II we used an ignorable RMM

Table III. Estimated conditional probabilities of observing complete data given  $(X_1, X_2, X_3, Y)$ 

Model		Y = 0				Y = 1			
		X <sub>1</sub> = 0		X <sub>1</sub> = 1		X <sub>1</sub> = 0		X <sub>1</sub> = 1	
		X <sub>2</sub> = 0	X <sub>2</sub> = 1	X <sub>2</sub> = 0	X <sub>2</sub> = 1	X <sub>2</sub> = 0	X <sub>2</sub> = 1	X <sub>2</sub> = 0	X <sub>2</sub> = 1
PM 123	X <sub>3</sub> = 1	0.23	0.14	0.31	0.27	0.24	0.14	0.32	0.27
	X <sub>3</sub> = 2	0.32	0.19	0.30	0.26	0.33	0.19	0.31	0.26
	X <sub>3</sub> = 3	0.40	0.24	0.26	0.25	0.39	0.22	0.36	0.15
	X <sub>3</sub> = 4	0.47	0.28	0.28	0.24	0.43	0.25	0.29	0.24
PM 132	X <sub>3</sub> = 1	0.16	0.12	0.31	0.29	0.17	0.12	0.32	0.29
	X <sub>3</sub> = 2	0.29	0.21	0.30	0.28	0.29	0.20	0.31	0.27
	X <sub>3</sub> = 3	0.43	0.31	0.29	0.27	0.42	0.28	0.30	0.26
	X <sub>3</sub> = 4	0.54	0.39	0.28	0.25	0.51	0.35	0.29	0.25
RMM	X <sub>3</sub> = 1	0.37	0.32	0.33	0.28	0.36	0.30	0.39	0.28
	X <sub>3</sub> = 2	0.37	0.32	0.33	0.28	0.36	0.30	0.34	0.28
	X <sub>3</sub> = 3	0.37	0.32	0.33	0.28	0.36	0.30	0.34	0.28
	X <sub>3</sub> = 4	0.37	0.32	0.33	0.28	0.36	0.30	0.34	0.28

model, and in rows 3 and 4 we used non-ignorable PM models based on two different permutations of the variables  $X_1, X_2, X_3$ . Table III gives the estimates of  $\pi(\mathbf{1}, L)$  for each of the 32 possible values of  $L$  under each of the three aforementioned models.

Table II indicates the sensitivity of our estimates of the parameter  $\alpha$  of model (1) to assumptions about the missing data process. As discussed in the Introduction, researchers have generally carried out such sensitivity analyses using fully parametric likelihood-based approaches that require one to specify a parametric model for the marginal distribution of the regressors  $X$ , a distribution which is of no scientific interest. In contrast, the estimators of  $\alpha$  in Table II based on weighted estimating equations are semi-parametric in the sense that they will be consistent asymptotically normal for the parameter  $\alpha$  of model (1), whatever be the distribution of the regressors  $X$ , provided the assumptions about the missing data process encoded in our missing data model are correct.

In the remainder of the paper, we describe the PM missing data models and their properties and provide algorithms for fitting these models to data. RMM models were considered in Robins and Gill.<sup>14</sup>

#### 4. PERMUTATION MISSINGNESS MODELS

In this section we consider the PM models of Tables II and III. Let  $(X_1, \dots, X_K)$  denote the  $K$  variables in  $L$  which have at least some data missing. In our data set,  $K = 3$ . For each of the  $K! = 6$  permutations of  $(X_1, \dots, X_K)$  we obtain a separate class of PM models. Given a permutation, let  $(X^{(1)}, X^{(2)}, \dots, X^{(K)})$  denote the 1st, 2nd, etc. variable in our permutation. For example, the permutation 132 in row 4 of Table II corresponds to  $X^{(1)} = X_1, X^{(2)} = X_3$ , and  $X^{(3)} = X_2$ . Let  $R^k$  denote the indicator of whether variable  $X^{(k)}$  is observed. Then, for a given permutation, by definition, the class of PM processes assumes that the probability that variable  $X^{(k)}$  is missing may depend on: (i) the variables that are never missing (in our case,  $Y$ ); (ii) the permutation-specific ‘past’  $X^{(1)}, \dots, X^{(k-1)}$ ; (iii) the permutation-specific ‘observed future’  $R^{k+1}, \dots, R^K, R^{k+1}X^{(k+1)}, \dots, R^KX^{(K)}$ , but (iv) not on  $X^{(k)}$  itself. (Note that, if  $X^{(k+1)}$  is unobserved,  $R^{k+1}X^{(k+1)} = 0$  and thus does not depend on the value of  $X^{(k+1)}$ ; however,  $R^{k+1}X^{(k+1)} = X^{(k+1)}$  if

$X^{(k+1)}$  is observed. Thus  $R^{k+1}, \dots, R^K, R^{k+1}X^{(k+1)}, \dots, R^KX^{(K)}$  is the ‘observed future’. In particular, for  $k \geq 1$ , the probability that  $X^{(k)}$  is observed may depend on  $X^{(1)}, \dots, X^{(k-1)}$  whether or not they are observed. Thus a PM process can be non-ignorable. Note, however, the probability that  $X^{(1)}$  is observed only depends on the observed data. We now provide a more formal mathematical characterization of a PM model.

Given a specified permutation, let  $R^* = (R^1, \dots, R^K)'$ . Note

$$\pi(\mathbf{1}, L) = \text{pr}[R^* = \mathbf{1}|L]. \quad (3)$$

We can always write  $\text{pr}(R^*|L) \equiv \text{pr}[R^1, \dots, R^K|X \equiv (X^{(1)}, \dots, X^{(K)}), Y]$  by a ‘reverse order’ factorization (that is, starting from  $R^K$ ) as

$$\text{pr}(R^*|L) = \prod_{k=1}^K \text{pr}[R^k|R^{k+1}, \dots, R^K, X, Y] \quad (4)$$

where  $R^{K+1} \equiv 0$ .

*Definition 1:* A permutation-specific PM missing process imposes the restriction on  $\text{pr}(R^*|L)$  that each term on the right-hand side of (4) satisfies

$$\text{pr}[R^k = 1|R^{k+1}, \dots, R^K, X, Y] = \text{pr}[R^k = 1|R^{k+1}, \dots, R^K, X^{(1)}, \dots, X^{(k-1)}, R^{k+1}X^{(k+1)}, \dots, R^KX^{(K)}, Y]. \quad (5)$$

Further, we assume that the right-hand side of (5) satisfies

$$\text{right hand side of (5)} > \sigma > 0 \quad (6)$$

where  $\sigma$  is an arbitrary positive constant.

*Remark:* Assumption (6) ensures a finite asymptotic variance for the estimator  $\hat{\alpha}$  described below of model (1). Equation (6) may hold for some permutations but not for others. For example, suppose that the data have a monotone missing data pattern so that whenever  $R_k = 1$ , we have that  $R_{k-1} = 1$ . Then (6) could only hold for the permutation  $X^{(1)} = X_K, X^{(2)} = X_{K-1}, \dots, X^{(K)} = X_1$ . For this particular permutation, however, the PM process (5) represents a new non-ignorable missing data process for monotone missing data.

In rows 3 and 4 of Table II, we specified linear logistic PM models. Specifically, we assumed

$$\text{right hand side of (5)} = \text{expit}[\gamma'_k V_k] \quad (7)$$

where  $\gamma_k$  is an unknown parameter vector to be estimated, and  $V_k$  is a vector of the dimension of  $\gamma_k$  that depends on the data only through

$$H_k = \{R^{k+1}, \dots, R^K, X^{(1)}, \dots, X^{(k-1)}, R^{(k+1)}X^{(k+1)}, \dots, R^KX^{(K)}, Y\}.$$

That is,  $V_k \equiv g_k[R^{k+1}, \dots, R^K, X^{(1)}, \dots, X^{(k-1)}, R^{(k+1)}X^{(k+1)}, \dots, R^KX^{(K)}, Y]$ , where  $g_k(\cdot)$  is a known vector-valued function chosen by the investigator. Note that the linear logistic PM model (7) uses distinct parameters  $\gamma_k$  for each occasion  $k$ . Note also that by varying the choice of the functions  $g_k(\cdot)$ , any PM process, that is, any process satisfying (5), can be represented by a linear logistic model of the form (7). In the analysis reported in rows 3 and 4 of Table II, we chose  $\gamma'_k V_k$  as follows:

$$\gamma'_3 V_3 = \gamma_{03} + \gamma_{13}X^{(1)} + \gamma_{23}X^{(2)} + \gamma_{33}Y + \gamma_{43}X^{(1)}X^{(2)},$$

$$\gamma'_2 V_2 = (1 - R^3)\{\gamma_{002} + \gamma_{102}X^{(1)} + \gamma_{202}Y\} + R^3\{\gamma_{012} + \gamma_{112}X^{(1)} + \gamma_{212}Y + \gamma_{312}X^{(3)} + \gamma_{412}X^{(1)}X^{(3)}\};$$

$$\begin{aligned} \gamma'_1 V_1 = & (1 - R^2)(1 - R^3) \{ \gamma_{0001} + \gamma_{1001} Y \} + R^2(1 - R^3) \{ \gamma_{0101} + \gamma_{1101} Y + \gamma_{2101} X^{(2)} \} + \\ & + R^3(1 - R^2) \{ \gamma_{0011} + \gamma_{1011} Y + \gamma_{2011} X^{(3)} \} + R^2 R^3 \{ \gamma_{0111} + \gamma_{1111} Y + \gamma_{2111} X^{(2)} \} \\ & + \gamma_{3111} X^{(3)} + \gamma_{4111} X^{(2)} X^{(3)}. \end{aligned}$$

*Remark 1:* We say a linear logistic PM model is saturated if, for each  $k$ ,  $V_k$  includes all possible functions of  $H_k$ . The model (7), as specified, is not saturated. To make it saturated,  $\gamma'_1 V_1$ ,  $\gamma'_2 V_2$ , and  $\gamma'_3 V_3$  must be expanded to include appropriate 2-way and 3-way interaction terms. For example,  $\gamma'_3 V_3$  must be expanded to include the 2-way and 3-way interaction terms  $\gamma_{53} X^{(1)} Y$ ,  $\gamma_{63} X^{(2)} Y$ , and  $\gamma_{73} X^{(1)} X^{(2)} Y$ .

It follows from (5) and (7) that, under our linear logistic model, the probability  $\pi(\mathbf{1}, L)$  estimated in Table III equals  $\prod_{k=1}^K \text{expit}[\gamma'_k V_k]$  with  $K = 3$ . Thus, it only remains to describe how we can estimate the parameters  $\gamma_k$ ,  $k = 1, \dots, K$  from the data. We use the following recursive estimation scheme. Since  $V_1$  depends only on the observed data, we can obtain  $\hat{\gamma}_1$  by logistic regression of  $R^1$  on  $V_1$ . However, since  $V_k$  depends on  $(X^{(1)}, \dots, X^{(k-1)})$ ,  $V_k$  will only be observed for subjects with  $\delta_{k-1} = 1$  where  $\delta_{k-1}$  is the indicator variable that takes the value 1 if  $R^1 = R^2 = \dots = R^{k-1} = 1$ . However, as shown in Appendix I, given  $\hat{\gamma}_1, \dots, \hat{\gamma}_{k-1}$ , we can obtain a consistent asymptotically normal estimator  $\hat{\gamma}_k$  of  $\gamma_k$  by weighted logistic regression of  $R^k$  on  $V_k$  among subjects with  $\delta_{k-1} = 1$  if we use inverse weights  $\prod_{m=1}^{k-1} \text{expit}[\hat{\gamma}'_m V_m]$ .

Formally,  $\hat{\gamma}_1$  solves

$$0 = \sum_i \{ R_i^1 - \text{expit}[\gamma'_1 V_{1i}] \} V_{1i} \quad (8)$$

and, for  $k = 2, \dots, K$ ,  $\hat{\gamma}_k$  solves

$$0 = \sum_i \left[ \delta_{(k-1)i} \left/ \left\{ \prod_{m=1}^{k-1} \text{expit}(\hat{\gamma}'_m V_{mi}) \right\} \right. \right] \{ R_i^k - \text{expit}(\gamma'_k V_{ki}) \} V_{ki}. \quad (9)$$

Standard errors can be obtained in standard fashion using the delta method, but, due to the programming task involved, we have not yet carried out the computations. Equation (9) makes clear why the probability that  $R^k$  is missing cannot be allowed to depend on  $X^{(k)}$ ; otherwise  $V_{ki}$  in (9) would need to be a function of  $X_i^{(k)}$ . As a consequence, the contribution of a subject for whom  $\delta_{(k-1)} = 1$  and  $R_i^k = 0$  to Equation (9) would not be available for data analysis (since  $X_i^{(k)}$  would be missing).

Note that even when the components of  $X$  and  $Y$  are continuous, the linear logistic PM model (7) may still be fit using canned logistic regression software.

## 5. THE CONCEPT OF NON-PARAMETRIC SATURATED PROCESSES

We now show that the class of permutation-specific PM processes is **NPS**. Suppose that the complete data  $L$  is discrete with realizations  $\ell$  taking values in a finite set  $\mathbf{L}$ . That is,  $\mathbf{L}$  is the support of  $L$ . In fact, if, as conjectured, in reference 16, SMAR processes always constitute an NPS class, our results will hold when  $L$  contains a mixture of discrete and continuous random variables. We view the data on the  $n$  study subjects as independent and identically distributed realizations of  $(R, L_{(R)})$  from an unknown true distribution  $f^*(r, \ell_{(r)})$  generating the data.

*Definition:* We define a class (set)  $\{f(r|\ell)\}$  of missing data processes to be non-parametric saturated (NPS) over a class  $\mathbf{F}$  of observed data laws  $\{f(r, \ell_{(r)})\}$  if, for each law  $f(r, \ell_{(r)}) \in \mathbf{F}$ , there exists a unique member of the class, say  $f^\Delta(r|\ell)$  and a unique complete data law for  $L$ , say  $f^\Delta(\ell)$ , such that  $f(r, \ell_{(r)})$  is the marginal distribution of  $(R, L_{(R)})$  corresponding to the joint law  $f^\Delta(r, \ell) \equiv f^\Delta(r|\ell)f^\Delta(\ell)$ . For discrete  $L$ ,

$$f(r, \ell_{(r)}) = \sum_{\{\ell^\dagger \in L; \ell_{(r)}^\dagger = \ell_{(r)}\}} f^\Delta(r|\ell^\dagger)f^\Delta(\ell^\dagger). \quad (10)$$

Gill *et al.*<sup>16</sup> proved for discrete  $L$  and conjectured for arbitrary  $L$  that the class of MAR (that is, ignorable) processes is NPS over

$$\mathbf{F} = \{f(\mathbf{1}, \ell) \neq 0 \text{ for all } \ell \in \mathbf{L}\} \quad (11)$$

the set of observed data laws for which there is a positive density of observing each element  $\ell$  in the support of  $L$ . Recall that  $f(r|\ell)$  is in the MAR class if and only if

$$f(r|\ell) \text{ depends on } \ell \text{ only through } \ell_{(r)} \quad (12)$$

that is, the probability that  $R = r$  depends only on the observed component  $\ell_{(r)}$  of  $\ell$ .<sup>17</sup>

In particular, Gill *et al.*<sup>16</sup> show that for the MAR class, given (11), (i)  $f^\Delta(\ell)$  is the unique law maximizing the expected log-likelihood  $\sum_{\text{all}(r, \ell_{(r)})} f(r, \ell_{(r)}) \log f^\Delta(\ell_{(r)})$  where  $f^\Delta(\ell_{(r)})$  is the marginal law of  $\ell_{(r)}$  under  $f^\Delta(\ell)$ , and (ii)  $f^\Delta(r|\ell) = f(r, \ell_{(r)})/f^\Delta(\ell_{(r)})$  is then obtained by division.

### 5.1. Implications of NPS Processes for Sensitivity Analyses

We assume that  $f^*(r, \ell_{(r)})$  is the marginal distribution of  $(R, L_{(R)})$  from a joint unknown law  $f^*(r, \ell) = f^*(r|\ell)f^*(\ell)$  where  $f^*(\ell)$  is the true but unknown distribution of the complete data  $L$  and  $f^*(r|\ell)$  is the true but unknown missing data process. Suppose (i) we model the distribution of  $L$  completely non-parametrically (that is, we choose to place no *a priori* restrictions on  $f^*(\ell)$ ) and (ii) we restrict the missing data process only by assuming it lies in a particular NPS class over a set  $\mathbf{F}$  (for example, the MAR class over (11)).

Let  $\tilde{f}(r, \ell_{(r)})$  be the empirical distribution of the observed data. For example, if, as in Section 2,  $L = (X_1, X_2, X_3, Y)$ ,  $\tilde{f}\{r = (1, 1, 0, 1), \ell_{(r)} = (1, 0, 0)\}$  is the proportion of the  $n$  study subjects with  $X_1 = 1$ ,  $X_2 = 0$ , and  $Y = 0$  observed, but  $X_3$  missing.

Then, given any empirical law  $\tilde{f}(r, \ell)$  in  $\mathbf{F}$ , there will exist a unique estimated missing data process  $\hat{f}(r|\ell)$  lying in our NPS class and a unique estimated law for  $L$ ,  $\hat{f}(\ell)$ , that reproduces the observed data *exactly*. That is, the marginal  $\hat{f}(r, \ell_{(r)}) \equiv \sum_{\{\ell^\dagger \in L; \ell_{(r)}^\dagger = \ell_{(r)}\}} \hat{f}\{r|\ell^\dagger\} \hat{f}(\ell^\dagger)$  is exactly the empirical distribution of the data  $\tilde{f}(r, \ell_{(r)})$ . (Later we give algorithms for obtaining the estimates  $\hat{f}(r|\ell)$  and  $\hat{f}(\ell)$  from the empirical distribution of the data. For the moment, we only note that, by the definition of an NPS process over  $\mathbf{F}$ , these estimates exist.) If the true law of the observed data lies in  $\mathbf{F}$ , then the empirical law  $\tilde{f}(r, \ell_{(r)})$  will lie in  $\mathbf{F}$  with probability approaching 1 as the sample size increases; furthermore, if the true missing data mechanism  $f^*(r|\ell)$  lies in our NPS class then  $\hat{f}(r|\ell)$  and  $\hat{f}(\ell)$  will consistently estimate the true laws  $f^*(r, \ell)$  and  $f^*(\ell)$ . That is, given  $f^*(r, \ell_{(r)})$  in  $\mathbf{F}$ , the law of the complete data  $f^*(\ell)$  is non-parametrically identified if we have correctly assumed that the true missing data process lies in our chosen NPS class over  $\mathbf{F}$ , for example, the MAR class over (11). In addition,  $\hat{f}(r|\ell)$  and  $\hat{f}(\ell)$  are the non-parametric maximum likelihood estimators of  $f^*(r|\ell)$  and  $f^*(\ell)$  under the sole restriction that  $f^*(r|\ell)$  lies in the given NPS class over  $\mathbf{F}$ .

Suppose now that our assumption that the true missing data process lies in our chosen NPS class over  $\mathbf{F}$  (for example, the MAR class over (11)) is false (that is, our missing data model is misspecified). Then clearly  $\hat{f}(r|\ell)$  will be inconsistent for the true law  $f^*(r|\ell)$ , since  $\hat{f}(r|\ell)$  is restricted to be in the NPS class but  $f^*(r|\ell)$  is not in the class. As a consequence,  $\hat{f}(\ell)$  will also be inconsistent for the true law  $f^*(\ell)$ .

To avoid such inconsistency, we would like to be able to test whether our model is correctly specified, that is, whether the true missing data process lies in our chosen class (for example, the MAR class). However, if  $f^*(r, \ell_{(r)})$  lies in  $\mathbf{F}$  and there are no restrictions on the law  $f^*(\ell)$ , this is not possible. No empirical test based on the observed data can ever provide evidence against the hypothesis that the true missing data process law is in our NPS class, since, as we have seen, (i)  $\hat{f}(\ell)$  and  $\hat{f}(r|\ell)$  reproduce the empirical distribution of the data exactly and (ii)  $\hat{f}(r|\ell)$  lies in our NPS class, so there can be no evidence against the hypothesis that our NPS class contains the true missing data process. Indeed, suppose we had available two disjoint classes of NPS missing data processes over classes  $\mathbf{F}_1$  and  $\mathbf{F}_2$ , respectively (for example, no missing data process in the first class is found in the second class) and the empirical law  $\tilde{f}(r|\ell_{(r)})$  lies in both  $\mathbf{F}_1$  and  $\mathbf{F}_2$ . From the first class we would obtain a set of estimates  $\hat{f}_1(r|\ell)$  and  $\hat{f}_1(\ell)$  and from the second class a set of estimates  $\hat{f}_2(r|\ell)$  and  $\hat{f}_2(\ell)$ , both of which would give marginals  $\hat{f}_1(r, \ell_{(r)})$  and  $\hat{f}_2(r, \ell_{(r)})$  that are equal to one another and reproduce the empirical distribution of the observed data exactly. As noted above, there is no empirical test to determine in which NPS class the true missing data process lies, and yet, since the NPS models are disjoint,  $\hat{f}_1(r|\ell)$  and  $\hat{f}_2(r|\ell)$  must converge to different limits as the sample size gets large and thus  $\hat{f}_1(\ell)$  and  $\hat{f}_2(\ell)$  will also converge to different limits. We conclude that if the estimator of the law of  $L$  based on one NPS class is consistent, the other must be inconsistent. To put it another way, if *a priori* we are only willing to say that the true missing data process lives in one or the other of the two NPS classes, then the marginal distribution of the complete data  $L$  is not identified. Hence analysing the data twice – first assuming that the missing data processes is in one class, and then in the other class – constitutes a sensitivity analysis. We obtain two different estimates of the marginal distribution of  $L$ , converging to different limits, both perfectly consistent with the observed data.

## 5.2. PM Classes and Sensitivity Analyses

Recall that the permutation-specific class of PM processes is the set  $\{f(r|\ell)\}$  of missing data processes satisfying (5). In the Appendix we prove the following theorem for discrete  $L$ .

*Theorem 1:* A permutation-specific class of PM models is NPS over the class  $\mathbf{F}$  described in Appendix II.

We relegate the characterization of  $\mathbf{F}$  to the Appendix, since it is somewhat unwieldy to describe. However, informally,  $\mathbf{F}$  is exactly the class of observed data laws for which the missing data process  $f^A(r|\ell)$  satisfying (10) also satisfies the restriction (6). It is (6) that allows the identification of  $f^A(\ell)$  from the observed data. We conjecture Theorem (1) holds even if  $L$  has continuous components.

It follows that we have  $3! = 6$  classes of permutation-specific PM models plus the NPS class of MAR models over the sets  $\mathbf{F}$  given in (11) and Appendix II. Since the empirical distribution of the data  $\tilde{f}(r, \ell_{(r)})$  from our breast cancer control study lies in each of these sets  $\mathbf{F}$ , we can provide non-parametric estimates of the law of  $L$  under each of the 7 NPS classes. (Note since we do not wish to restrict the distribution of  $L$  *a priori*, we no longer impose the linear logistic model (1). Rather, we shall leave both  $\text{pr}[Y = 1|X_1, X_2, X_3]$  and the marginal law of  $(X_1, X_2, X_3)$

Table IV.

NPS class	Non-parametric estimate of $P[Y = 1 X_1 = 0, X_2 = 1, X_3 = 0]$
MAR	0.703
Permutation 123	0.599
Permutation 231	0.601
Permutation 213	0.597
Permutation 132	0.704
Permutation 312	0.714
Permutation 321	0.709
Complete case analysis*	0.750

\* The complete case analysis is for comparison purposes only; it does not represent an NPS class

completely unrestricted.) For the MAR class, it is well known that  $\hat{f}(\ell)$  can be obtained by using the expectation-maximization (EM) algorithm.<sup>18</sup> For a permutation-specific PM class,  $\hat{f}^\Delta(\ell) = n^{-1} \sum_i \delta_i I(L_i = \ell) / \prod_{k=1}^K \text{expit}[\hat{\gamma}'_k V_{ki}]$ , where, now, in estimating  $\gamma_k$  the linear logistic model (7) is made completely saturated, as described in Remark 1 of Section 4 and again  $\delta_i$  is the indicator for complete data on subject  $i$ .

Results are given in Table IV. In Table IV, we restrict attention to estimating  $\text{pr}[Y = 1|X_1, X_2, X_3]$  because the marginal distribution of  $(X_1, X_2, X_3)$  is of less interest. (Of course, in a case-control study,  $\text{pr}[Y = 1|X_1, X_2, X_3]$  computed from the case-control data does not reflect the same population proportions. However, because the calculations here are for illustrative purposes, we have chosen to analyse the data in this Section as if it were cohort data. In contrast, the estimates of  $\alpha_1, \alpha_2$ , and  $\alpha_3$  reported in rows 3 and 4 of Table II represent estimates of log odds ratios and thus are valid even though the data are from a case-control study.) Further, in Table IV, to prevent clutter, we report only  $\text{pr}[Y = 1|X_1 = 0, X_2 = 1, X_3 = 0]$  where now, to simplify our computations, we have redefined  $X_3$  to be a dichotomous indicator variable representing whether the  $10^4 \times \text{height/weight}^2$  ratio exceeds 25. We simplified the computation of the permutation-specific PM class estimates by noting the non-parametric PM estimate of  $\text{pr}[Y = 1|X_1 = 0, X_2 = 1, X_3 = 0]$  is equal to  $\{\sum_{\{i: X_{1i}=0, X_{2i}=1, X_{3i}=0\}} \delta_i Y_i / \hat{\pi}(\mathbf{1}, L_i)\} / \sum_{\{i: X_{1i}=0, X_{2i}=1, X_{3i}=0\}} \{\hat{\pi}_i / \hat{\pi}(\mathbf{1}, L_i)\}$ .

The six permutation-specific PM classes and the MAR class have a non-empty intersection. In particular, they all contain the missing completely at random (MCAR) processes, i.e., processes for which  $f(r|\ell)$  does not depend on  $\ell$ .

If the true missing data process  $f^*(r|\ell)$  were MCAR, then the complete case estimate reported in row 8 of Table IV (which is the average value of  $Y$  among subjects with  $X_1 = 0, X_2 = 1, X_3 = 0$ ) would be consistent. Ignoring sampling variability, the fact that the estimates in rows 1–7 differ from the complete case estimate in row 8 and each other shows that the missing data processes picked by non-parametrically fitting the 7 NPS classes are not MCAR processes.

Studies of missing data are commonly analysed solely under the assumption that the missing data mechanism is ignorable. However, this assumption, as we have noted, is untestable, and, based on subject matter considerations, is often suspected not to be true. Therefore it is important to do a sensitivity analysis in which one analyses the data under non-ignorable missing data mechanisms as well. Little and Rubin<sup>9</sup> as well as others have previously stressed the importance of investigating sensitivity of one's inferences to one's (non-identifiable) assumptions about the missing data process.

By using classes of NPS missing data processes, we have been able to carry out, in Table IV, such a sensitivity analysis non-parametrically, that is, without placing any restrictions on the joint distribution of the observed data. With rare exceptions,<sup>3,10</sup> previous researchers have only carried out sensitivity analyses using unsaturated non-ignorable missing data models that place *a priori* restrictions on the joint distribution of the data. When one carries out a sensitivity analysis using unsaturated models, it is difficult to know to what degree one's estimates are being influenced by the *a priori* restrictions on the distribution of the observed data implicit in the model.

## 6. ADDITIONAL NPS CLASSES OF MISSING PROCESSES

Because of their usefulness for sensitivity analyses, it is interesting to determine whether there are NPS classes other than the permutation-specific PM classes and the MAR class. In this section, we show that there are other NPS classes. Indeed, both the PM classes and the MAR class are particular elements of the following more general set of NPS classes. Again, let  $X = (X_1, \dots, X_K)$ , with  $X_K$  univariate, be the variables for which there is some missing data, and let  $Y$  denote all variables that are always observed. The following algorithm produces a NPS class over the class  $\mathbf{F}$  described in Appendix II.

Divide  $(X_1, \dots, X_K)$  into  $M$ ,  $1 \leq M \leq K$ , disjoint and mutually exclusive non-empty groups of variables  $(Z_1, \dots, Z_M)$ . Let  $Z^{(1)}, \dots, Z^{(M)}$  represent a particular one of the  $M!$  permutations of the  $M$  groups. (Note that if  $M = K$ ,  $Z^{(1)}, \dots, Z^{(M)}$  equals  $X^{(1)}, \dots, X^{(K)}$  of Section 4 and each group  $Z^{(m)}$  has exactly one member.) Let  $R^m$  with realizations  $r^m$  be the vector of missing data indicators corresponding to group  $Z^{(m)}$ . Let  $Z_{(R^m)}^{(m)}$  be the observed components of  $Z^{(m)}$ .

*Examples:* Suppose  $Z^{(m)} = (X_1, X_3, X_7)$  is composed of the first, third, and seventh of the original variables. Then,  $R^m = (1, 0, 1)$  implies that  $X_1$  and  $X_7$  were observed but  $X_3$  was unobserved. If  $R^m = (1, 0, 1)$ ,  $Z_{(R^m)}^{(m)} = (X_1, X_7)$ .

*Definition:* A grouping-permutation-specific grouped permutation missing (GPM) process satisfies, for  $m = 1, \dots, M$ ,

$$\begin{aligned} \text{pr}(R^m = r^m | R^{m+1}, \dots, R^M, L = (X, Y)) \\ = \text{pr}(R^m = r^m | R^{m+1}, \dots, R^M, Y, Z^{(1)}, \dots, Z^{(m-1)}, Z_{(r^m)}^{(m)}, Z_{(R^{m+1})}^{(m+1)}, \dots, Z_{(R^M)}^{(M)}). \end{aligned} \quad (13)$$

That is, given  $R^{m+1}, \dots, R^M$ , the probability that the components  $Z_{(r^m)}^{(m)}$  of the  $m$ th class  $Z^{(m)}$  will be observed depends on and only on these components  $Z_{(r^m)}^{(m)}$ , the 'observed future,' and the entire past.

*Definition:* A grouping-permutation-specific GPM class is the set of all missing data processes satisfying (13).

In the Appendix we prove the following theorem.

*Theorem (2):* A grouping-permutation-specific GPM class is a NPS class over the set  $\mathbf{F}$  defined in Appendix II.

*Remark 2:* If  $K = M$ , equation (13) is equivalent to equation (5). Hence PM processes are the subset of GPM processes with each group  $Z_m$  containing exactly one variable. If  $M = 1$  so that all variables  $(X_1, \dots, X_K)$  are in a single group, (13) reduces to the missing at random assumption (12).

As shown in Table IV, we estimated  $\text{pr}[Y = 1|X_1 = 0, X_2 = 1, X_3 = 0]$  under seven different grouping-permutation-specific NPS classes of GPM processes. (Recall the MAR class is the GPM class with  $M = 1$ .) There are exactly six additional grouping-permutation-specific NPS classes of GPM processes derived from the two possible permutations of each of the following three groupings:  $(X_1, X_2) = Z_1$  and  $X_3 = Z_2$ ;  $(X_2, X_3) = Z_1$  and  $X_1 = Z_2$ ;  $(X_1, X_3) = Z_1$  and  $X_2 = Z_2$ . We have not carried out sensitivity analyses using these six additional classes.

## 7. EFFICIENCY, HISTORY OF PM MODELS, AND THE COX MODEL

### 7.1. Efficiency

The estimator recorded in row 3 of Table II is not efficient in the model characterized by the logistic regression model (1) and the PM missing data models (5–7). When the regressors  $X$  are discrete, this model is a fully parametric model in which the model for the distribution of the regressors  $X$  is saturated; hence, an efficient estimator of  $\alpha$  could be obtained by maximum likelihood. On the other hand, when  $X$  has continuous components, this model is a semi-parametric model since the distribution of the regressors is left unspecified; in this case, application of (non-parametric) maximum likelihood techniques might lead to inconsistent estimators. In this setting, an approach based on inverse probability weighted estimating equations can be extended to obtain a semi-parametric efficient estimator. Although the explicit construction of an efficient estimator is beyond the scope of this paper, it is useful to understand the outlines of the algorithm on which such a construction would be based.

The first step in the algorithm is to characterize (i) the set of all influence functions of regular asymptotically linear (RAL) estimators of  $\alpha$ , and (ii) the efficient influence function (EIF) in the complete data model (1) without missing data (that is, when  $R_1 = R_2 = R_3 = 1$  with probability one). An estimator  $\hat{\alpha}$  is asymptotically linear if  $\hat{\alpha} - \alpha_0$  is asymptotically equivalent to the average of independent and identically distributed mean zero random variables  $B$ , that is,  $n^{1/2}(\hat{\alpha} - \alpha_0) = n^{1/2} \sum_i B_i/n + o_p(1)$  where  $E[B] = 0$ ,  $\text{var}(B)$  is finite, and  $o_p(1)$  refers to a random variable converging to zero in probability.  $B$  is called the influence function of  $\hat{\alpha}$ . By the central limit theorem and Slutsky's theorem, an asymptotically linear estimator  $\hat{\alpha}$  is asymptotically normal with mean zero and variance equal to  $\text{var}(B)$ . Estimators, including parametric maximum likelihood estimators, that solve unbiased estimating equations are asymptotically linear. An estimator is regular if its convergence to its limiting distribution is locally uniform. Regularity is a necessary condition for 95 per cent Wald intervals based on  $\hat{\alpha} \pm 1.96$  estimated asymptotic standard errors to be a valid large sample confidence interval. The EIF is the influence function of the semi-parametric efficient estimator. In our logistic regression model (1) without missing data, the set of influence functions of RAL estimators is  $\{E[d(X)X' \text{expit}(\alpha'X) \{1 - \text{expit}(\alpha'X)\}]^{-1} \times d(X)(Y - \text{expit}(\alpha'X))\}$  where  $d(X)$  is any function of  $X$  of the same dimension as  $\alpha$ . The efficient influence function has  $d(X) = X$ . It is the influence function of the logistic regression maximum likelihood estimator.

The next step of the algorithm uses the representation theorem of Robins, Rotnitzky, and Zhao<sup>5</sup> (RRZ). Their representation theorem shows how to obtain the set of all influence functions and the EIF in an arbitrary semi-parametric model with the data missing at random from the set of all influence functions and the EIF in the corresponding semi-parametric model without missing data. Further, once the EIF is known for the missing data model, RRZ show how to construct locally efficient semi-parametric estimators. Thus, RRZ's representation theorem can be used to obtain locally efficient estimators in a semi-parametric model with ignorable missing data, provided that we know the set of influence functions and the EIF for the model without

missing data. Now, as non-ignorable models, the PM models are not MAR models. However, we will now show that any PM model can be represented as a nested sequence of MAR models. (As we shall see, a model that can be represented as a nested sequence of MAR models need not itself be an MAR model.) Using this representation of a PM model as a nested sequence of MAR models, we will be able to find the efficient influence function and thus a locally semi-parametric efficient estimator in our PM model by the repeated sequential use of the RRZ representation theorem.

As the next step in our algorithm we suppose, for the moment, that in the semi-parametric model characterized by the (1) and the PM model (5–7),  $X^{(1)}$  and  $X^{(2)}$  are always observed but  $X^{(3)}$  may be missing (that is,  $R^1 = R^2 = 1$ ). Under such circumstances,  $X^{(3)}$  is missing at random since (5–7) imply  $\text{pr}[R^{(3)} = 1 | X, Y] = \text{expit}(\gamma'_3 V_3)$  and  $V_3$  does not depend on  $X^{(3)}$ . Since in the previous paragraph but one, we have exhibited the set of all influence functions of RAL estimators and the EIF in this semi-parametric model with complete data, it follows that we can use RRZ's representation theorem to find all influence functions of RAL estimators and the EIF in the missing data model in which  $R^1 = R^2 = 1$  but  $X^{(3)}$  may be missing.

The next step in the algorithm is to regard as the 'complete' data model the model that was the 'missing' data model in the previous step. That is, we regard as the 'complete' data model the semi-parametric model characterized by the logistic model (1) and the PM model (5–7) in which  $X^{(1)}$  and  $X^{(2)}$  are both always observed and  $X^{(3)}$  may be missing. We consider as the 'missing' data model the corresponding model in which  $X^{(1)}$  is always seen, so  $R^{(1)} = 1$  with probability 1, but  $X^{(2)}$  and  $X^{(3)}$  may be missing. Since (i) we have characterized the set of influence functions of RAL estimators and the EIF in this 'complete' data model in the previous paragraph and, (ii) with respect to this 'complete' data model, in our 'missing' data model variable  $X^{(2)}$  is missing-at-random (by (5–7)  $\text{pr}[R^2 = 1 | X^{(1)}, X^{(2)}, R^{(3)}, R^{(3)} X^{(3)}, Y] = \text{expit}(\gamma'_2 V_2)$  and  $V_2$  does not depend on  $X^{(2)}$ ); it follows that we can use RRZ's representation theorem to obtain all influence functions of RAL estimators and the EIF in the 'missing' data model characterized by (1), (5–7), and  $X^{(1)}$  always observed.

The last step of the algorithm is to apply our trick once again, and now consider the 'missing' data model of the last step as our new 'complete' data model. With respect to this 'complete' data model, our actual model of interest (in which all of the variables  $X^{(1)}$ ,  $X^{(2)}$ , and  $X^{(3)}$  may be missing) has, by (5–7),  $X^{(1)}$  missing at random. It follows that we can obtain the influence functions of all RAL estimators and the EIF in our actual model of interest (that is, the model characterized by the logistic model (1) and the PM model (5–7) in which any of the regressors may be missing) by applying the RRZ representation theorem to the current 'complete data model'.

Thus we have shown that any semi-parametric model with a PM missingness process can be viewed as a nested sequence of MAR models in the above sense. Exactly the same logic can be used to obtain the EIF and thus locally semi-parametric efficient estimators in the more general GPM models.

## 7.2. History of PM Models and the Cox Proportional Hazards Model

The concept of PM and GPM missing data processes can be extended to more general missing data models. In fact, as I now describe, I first recognized the possibility of PM models in the context of a Cox proportional hazard model with missing covariates. In a Cox model, we observe the time  $T$  to failure in addition to regressors  $X$  when we obtain complete data. The Cox model specifies that  $\lambda_T(t|X) = \lambda_0(t)\exp(\alpha'X)$  where  $\lambda_T(t|X)$  is the hazard of  $T$  at  $t$  given  $X$ , and  $\lambda_0(t)$  is an unspecified baseline hazard. In practice, the failure time variable  $T$  is often not observed due to censoring. In that case, we observe  $T^* = \min(T, C)$  and  $\delta = I(T < C)$  where  $C$  is the time to

censoring. In addition, data on the covariates  $X$  may be missing for some subjects. To simplify our discussion, let us suppose  $X$  is univariate. Then a Cox model with missing data due to both missing covariates and to censoring constitutes a non-monotone missing data problem since, for some subjects,  $T$  is observed but not  $X$ , for some  $X$  is observed but not  $T$ , for some neither is observed, and finally for some subjects both are observed.

Thus two PM classes should exist depending on the ordering of the variables  $T$  and  $X$ . The first PM class allows (i) censoring (that is, missingness in  $T$ ) to depend on  $X$  (even when  $X$  is not observed) but not on  $T$ , and (ii) missingness in  $X$  to depend on the observed part of  $T$ , that is,  $(T^*, \delta)$ . Assumption (i) is the usual assumption of independent censoring given the regressors. This PM process represents precisely the assumptions concerning the missing data process made by Pugh *et al.*,<sup>19</sup> and Robins *et al.*<sup>5</sup> in their treatment of missing covariates in the Cox model. It was when I realized that the missingness mechanism assumed in references 5 and 19 was not ignorable (since missingness on  $T$  (censoring) can depend on the possibly unobserved value of  $X$ ), that I recognized the existence of and the ease of fitting the general class of PM models. The second PM class allows: (a) missingness in  $X$  to depend on  $T$  (even if unobserved) but not on  $X$ ; and (b) missingness in  $T$  (that is, censoring) to depend on the observed part of  $X$  (that is, the hazard of censoring can depend on  $X$  only among subjects for whom  $X$  was observed).

From our discussion of GPM models, it follows that a third alternative is to assume that missingness in both  $T$  and  $X$  is ignorable given  $T$  and  $X$ . One could specify such an ignorable non-monotone missing data model by using the class of randomized monotone missingness models considered in Robins and Gill.<sup>14</sup> Analysis of the Cox model with missing covariate data using either the second non-ignorable PM class described above or the ignorable randomized monotone missingness class has yet to be implemented in practice.

## 8. DISCUSSION AND FURTHER CONSIDERATION

We have proposed a large number of classes of NPS missing data processes, the grouping-permutation-specific GPM classes. We showed how for particular GPM classes, the PM classes, canned logistic regression software can be used to construct semi-parametric estimators of the parameters  $\alpha$  indexing a logistic model for the conditional mean of a dichotomous response variable  $Y$  given a vector of (possibly continuous) regressors. As NPS classes, the GPM classes are useful for sensitivity analysis. In particular, with discrete data, they allow us to investigate, in a non-parametric fashion, the sensitivity of our estimates of a regression parameter  $\alpha$  to non-identifiable assumptions about the missing data process. However, the range of estimates obtained using the full set of grouping-permutation-specific GPM classes will still not be as wide as that obtained using the method of probabilistically bounding estimates of  $\alpha$  by filling in the missing data with values that make one's completed data estimates of  $\alpha$  as extreme as possible. This latter approach would constitute the most robust form of sensitivity analysis but would give the least informative bounds.

When the complete data  $L$  is discrete, Baker *et al.*<sup>3</sup> and Baker and Laird<sup>10</sup> have displayed NPS classes for which the probability that the variable  $L_k$  is missing can depend on the possibly missing value of that variable. However, in contrast to the GPM classes, the Baker *et al.*<sup>3</sup> and Baker and Laird<sup>10</sup> NPS classes cannot be directly extended to data with continuous components. An important open question is the question of the existence, when  $L$  has continuous components, of NPS classes for which the probability that the variable  $L_k$  is missing can depend on the possibly missing value of that variable.

A final question about GPM processes: how often do such processes appropriately model beliefs about the missing data process in important, substantive contexts?

As an example, suppose a reform school offered HIV testing. The HIV positive rate was 35 per cent among the 30 per cent accepting the offer, i.e.  $\widehat{pr}(X_1 = 1 | R_1 = 1) = 0.35$ . Because of concern over non-random non-response, data on risks for and fears about HIV were abstracted as variable  $X_2$  from a 20 per cent simple random sample of counseling files, so that, by design,  $pr(R_2 = 1 | X_1, X_2, R_1) = 0.2$ . Under the additional working hypothesis that being tested ( $R_1$ ) was independent of HIV status ( $X_1$ ) given  $X_2$ , the data follow a non-monotone non-ignorable PM process under the permutation  $(X^{(1)}, X^{(2)}) = (X_2, X_1)$ .

Whatever the answer to the above question, the study of GPM processes sheds light on the mathematical structure underlying non-ignorable missing data processes. Greater understanding of this structure can only help clarify the benefits and risks associated with fitting non-ignorable missing data models.

### APPENDIX I

To show that, under mild regularity conditions,  $(\hat{\gamma}_1, \dots, \hat{\gamma}_k)$  solving (8) and (9) are consistent and asymptotically normal for the parameters  $(\gamma_1, \dots, \gamma_k)$  under (5), (6), and (7), it is sufficient to prove that, for  $k = 1, \dots, K$ ,  $E[A_{k1}] = 0$  where, for  $j = 1, \dots, k-1$ ,  $A_{kj} \equiv R^j R^{j+1} \dots R^{k-1} \times \{[\prod_{m=j}^{k-1} \exp(\gamma'_m V_m)]\}^{-1} A_{kk}$  where  $A_{kk} \equiv [R^k - \text{expit}(\gamma'_k V_k)] V_k$ . Since by (5) and (7),  $E[A_{kk}] = 0$ , it is sufficient to show that  $E[A_{k(j-1)}] = E[A_{kj}]$ . Now

$$E[A_{k(j-1)}] = E[E\{R^{j-1}/\text{expit}(\gamma'_{j-1} V_{j-1}) | R^j, R^{j+1}, \dots, R^K, X, Y\} A_{kj}] = E[A_{kj}]$$

by (5), (6), and (7).

### APPENDIX II

In this Appendix, we (i) characterize the class **F** of observed data laws necessary for Theorems (1) and (2) to hold, and (ii) prove these theorems by explicitly constructing the laws  $f^\Delta(r|\ell)$  and  $f^\Delta(\ell)$  satisfying equation (10) for  $f(r, \ell_{(r)})$  in **F**.

Define for  $k = 1, \dots, M$ ,  $U^k = \{R^k, \dots, R^M, Z_{(R^k)}^{(k)}, \dots, Z_{(R^M)}^{(M)}\}$  and  $U^{M+1} \equiv 0$ . Define  $\bar{Z}^k = (Z^{(1)}, \dots, Z^{(k)})$  and  $Q_k \equiv (U^{k+1}, \bar{Z}^{k-1})$  with  $\bar{Z}^0 \equiv 0$ . For any random variable  $W$ , let **W** denote its support and  $w$  be a realization, so  $w \in \mathbf{W}$ .

Given a grouping-permutation-specific ordering  $Z^{(1)}, \dots, Z^{(M)}$ , for each  $f(r, \ell_{(r)})$  in the class **F** described below, we will recursively characterize for  $k = 1, \dots, M$  laws  $f_k(z^{(k)}, q_k)$  and  $f_k(r^k | z^{(k)}, q_k)$  as described below. Having done so, we then define

$$f^\Delta(r|\ell) = \prod_{k=1}^M f_k(r^k | z^{(k)}, q_k) \quad \text{and} \quad f^\Delta(\ell) = f_M(z^{(M)}, q_M) \equiv f_M(\bar{Z}^M)$$

since  $\bar{Z}^M = (Z^{(M)}, Q_M)$ .

We then prove Theorem (2) and its special case Theorem (1) by showing (i)  $f^\Delta(r|\ell)$  is a grouping-permutation-specific GPM process (that is, equation (13) is true), (ii) equation (10) is satisfied, and (iii)  $f^\Delta(r|\ell)$  and  $f^\Delta(\ell)$  are unique in the sense that they are the only laws for which (i) and (ii) both hold.

*Definitions:* Given an observed data law  $f(r, \ell_{(r)})$ , let  $f_0(r^1, z^{(1)} | q_1)$  be the law of  $(r^1, z^{(1)})$  given  $q_1$ . Then if

$$f_{k-1}(r^k = \mathbf{1}, z^{(k)} | q_k) \neq 0 \tag{14}$$

for all  $q_k \in \mathbf{Q}_k$ ,  $z^{(k)} \in \mathbf{Z}^{(k)}$ , we uniquely define  $f_k(z^{(k)}|q_k)$  and  $f_k(r^k|z^{(k)}, q_k)$  by the assumption that given  $q_k$ , the distribution of  $(R^k, Z^{(k)})$  under  $f_k(\cdot, \cdot)$  is MAR and marginalizes to  $f_{k-1}(\cdot, \cdot)$ . Formally, given  $f_{k-1}(\cdot, \cdot)$  and (14),  $f_k(r^k|z^{(k)}, q_k)$  and  $f_k(z^{(k)}|q_k)$  are the unique laws satisfying

$$\text{MAR: } f_k(r^k|z^{(k)}, q_k) = f_k(r^k|z_{(r^k)}^{(k)}, q_k). \quad (15)$$

$$\text{Marginalization: } f_{k-1}(r^k, z_{(r^k)}^{(k)}|q_k) = f_k(r^k|z_{(r^k)}^{(k)}, q_k) \sum_{\{z^{(k)*} \in \mathbf{Z}^{(k)}, z_{(r^k)}^{(k)}\}} f_k(z^{(k)*}|q_k). \quad (16)$$

*Remark 1:* Given (14), existence and uniqueness are proved for discrete and conjectured for continuous variables by Gill *et al.*<sup>16</sup> Equation (14) is needed for uniqueness but not existence.<sup>16</sup>

*Remark 2:* If, as in a PM process,  $z^{(k)}$  is one-dimensional, we obtain the following closed form expressions in terms of  $f_{k-1}(\cdot, \cdot)$ :

$$f_k(r^k|z^k, q_k) = f_{k-1}(r^k|q_k) \quad (17)$$

$$f_k(z^{(k)}|q_k) = f_{k-1}(z^{(k)}|r^k = 1, q_k). \quad (18)$$

In the general case,  $f_k(z^{(k)}|q_k)$  is obtained by maximizing the expected log-likelihood based on  $f_{k-1}(\cdot, \cdot)$  and  $f_k(r^k|z^{(k)}, q_k)$  is then obtained by division.<sup>16</sup>

*Definition of the set  $\mathbf{F}$ :* The set  $\mathbf{F}$  in Theorem (2) and Theorem (1) is the set of laws  $f(r, \ell_{(r)})$  for which (14) is true for  $k = 1, \dots, M$ .

*Proof of Theorem (2):* By construction,  $f^\Delta(r|\ell)$  satisfies equation (13) and thus is a GPM process. Further by construction, equation (10) is satisfied. As mentioned above, uniqueness follows from Gill *et al.*<sup>16</sup>

### APPENDIX III: PROOF THAT NPS MODELS IN TABLE IId-IIIf OF BAKER *ET AL.*<sup>3</sup> ARE NOT GPM MODELS

*Proof:* Since there are only two variables, any GPM model is either a MAR model or a PM model. The Baker *et al.*<sup>3</sup> models are not MAR. Further, all four of the Baker *et al.*<sup>3</sup> models in the above-cited tables impose the restriction that neither

$$\text{pr}[R_2 = 1|R_1 = 1, X_1, X_2] \quad (19)$$

nor

$$\text{pr}[R_1 = 1|R_2 = 1, X_1, X_2] \quad (20)$$

is dependent on both  $X_1$  and  $X_2$ , since, in Baker *et al.*'s<sup>3</sup> notation, neither  $a_{ij}$  nor  $b_{ij}$  are functions of both  $i$  and  $j$ . However, in a saturated PM model, either (19) or (20) will depend on both  $X_1$  and  $X_2$ , since, using Bayes theorem and the definition (5) of the PM model,

$$\begin{aligned} \text{pr}[R^2 = 1|R^1 = 1, X^{(1)}, X^{(2)}] &= \text{pr}[R^1 = 1|R^2 = 1, X^{(2)}] \text{pr}[R^2 = 1|X^{(1)}] / \\ &\quad \{\text{pr}[R^1 = 1|R^2 = 1, X^{(2)}] \text{pr}[R^2 = 1|X^{(1)}] + \text{pr}[R^1 = 1|R^2 = 0] \text{pr}[R^2 = 0|X^{(1)}]\}, \end{aligned}$$

which depends on both  $X^{(1)}$  and  $X^{(2)}$ , i.e. on  $(X_1, X_2)$ , which completes the proof.

#### ACKNOWLEDGEMENTS

Support for this research was provided in part by Grants 2 P30 ES00002, RO1-A132475, and RO1-ESO3405.

## REFERENCES

1. Storm, H. H., *et al.* 'Adjuvant radiotherapy and risk of contralateral breast cancer', *Journal of the National Cancer Institute*, **84**, 1245–1250 (1995).
2. Vach, W. and Blettner, M. 'Logistic regression with incompletely observed categorical covariates: Investigating the sensitivity against violation of the missing at random assumption', *Statistics in Medicine*, **14**, 1315–1329 (1995).
3. Baker, S. G., Rosenberger, W. F. and DerSimonian, R. 'Closed-form estimates for missing counts in two-way contingency tables', *Statistics in Medicine*, **11**, 643–657 (1992).
4. Brown, C. H. 'Protecting against nonrandomly missing data in longitudinal studies', *Biometrics*, **46**, 143–155 (1990).
5. Little, R. A. 'A class of pattern-mixture models for normal incomplete data', *Biometrika*, **81**, 471–483 (1994).
6. Robins, J. M., Rotnitzky, A. and Zhao, L. P. 'Estimation of regression coefficients when a regressor is not always observed', *Journal of the American Statistical Association*, **89**, 846–866 (1994).
7. Rotnitzky, A. and Robins, J. M. 'Analysis of semiparametric regression models with non-ignorable non-response', *Statistics in Medicine* **16**, 81–102 (1997).
8. Robins, J. M., Rotnitzky, A. and Zhao, L. P. 'Analysis of semiparametric regression models for repeated outcomes in the presence of missing data', *Journal of the American Statistical Association*, **90**, 106–121 (1995).
9. Little, R. J. and Rubin, D. B. *Statistical Analysis with Missing Data*, Wiley, New York, 1987.
10. Baker, S. G. and Laird, N. M. 'Regression analysis for categorical variables with outcome subject to non-ignorable non-response', *Journal of the American Statistical Association*, **83**, 62–69 (1988).
11. Conaway, M. R. 'The analysis of repeated categorical measurements subject to non-ignorable non-response', *Journal of the American Statistical Association*, **87**, 817–824 (1992).
12. Diggle, P. and Kenward, M. G. 'Informative drop-out in longitudinal data analysis (with discussion)', *Journal of Applied Statistics*, **43**, 49–93 (1994).
13. Fay, R. E. 'Causal models for patterns of nonresponse', *Journal of the American Statistical Association*, **81**, 354–365 (1986).
14. Robins, J. M. and Gill, R. 'Non-response models for the analysis of non-monotone ignorable missing data', *Statistics in Medicine*, **16**, 39–56 (1997).
15. Horvitz, D. G. and Thompson, D. J. 'A generalization of sampling without replacement from a finite universe', *Journal of the American Statistical Association*, **47**, 663–685 (1952).
16. Gill, R., van der Laan, M. and Robins, J. M. 'Coarsening at random: characterizations, conjectures and counter-examples', (to appear).
17. Rubin, D. B. 'Inference and missing data', *Biometrika*, **63**, 581–592 (1976).
18. Dempster, A. P., Laird, N. M. and Rubin, D. B. 'Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion)', *Journal of the Royal Statistical Society, Series B*, **39**, 1–38 (1977).
19. Pugh, M., Robins, J. M., Lipsitz, S. and Harrington, D. 'Inference in the Cox proportional hazards model with missing covariates', Technical Report, Harvard School of Public Health, Department of Biostatistics.