THE USE OF AUXILIARY INFORMATION TO DEAL WITH INFORMATIVELY MISSING OR OBSERVED DATA

Vern Farewell
MRC Biostatistics Unit
Cambridge, UK

Major Collaborators: Andrew Copas, Michael Sweeting

'MISSING' OBSERVATIONS

- **Response variable of interest Y:**
  - **Observed:** Y^o
  - **Missing:** Y^m
- **Explanatory variable(s) or covariate(s):** X
- **Response or observation indicator:** R

'MISSING' OBSERVATIONS

- **Missing Completely at Random (MCAR)**
  - $f(R \mid Y^o, Y^m, X) = f(R)$
- **Covariate Dependent-MCAR (CD-MCAR)**
  - $f(R \mid Y^o, Y^m, X) = f(R \mid X)$
- **Covariate Dependent Missing at Random (CD-MAR)**
  - $f(R \mid Y^o, Y^m, X) = f(R \mid X, Y^o)$
- **Missing Not at Random (MNAR)**
  - $f(R \mid Y^o, Y^m, X) \neq f(R \mid X, Y^o)$

Two Modelling Approaches for MNAR Data

- **Selection Model:**

  $f(Y, R) = f(Y)f(R \mid Y)$

  or

  $f(Y, R \mid X) = f(Y \mid X)f(R \mid Y, X)$ **conditioning on X**

- **Pattern Mixture Model:**

  $f(Y, R) = f(Y \mid R)f(R)$

  or

  $f(Y, R \mid X) = f(Y \mid R, X)f(R \mid X)$ **conditioning on X**

National Survey of Sexual Attitudes and Lifestyles (NATSAL)

Involved face-to-face questioning and a self-completion booklet with more sensitive questions.

- **Responders:** provided answers to all questions
- **Item non-responders:** refused to answer some questions
- **Unit non-responders:** refused to answer any questions

Mock Example: Level of Virginity

	Responders	Item Non-responders	Unit Non-responders
Total	600	100	300

Estimate of level of virginity

Responders only: 12.5%

Mock Example: Level of Virginity

	Responders	Item Non-responders	Unit Non-responders
Embarrassed	150 20%	75	
Not Embarrassed	450 10%	25	
Total	600	100	300

Estimate of level of virginity

Responders only: 12.5%

Mock Example: Level of Virginity

	Responders	Item Non-responders	Unit Non-responders
Embarrassed	150 20%	75 20%	
Not Embarrassed	450 10%	25 10%	
Total	600	100	300

Estimate of level of virginity

Responders only: 12.5%

Responders + Item-nonresponders: 13.2%

Mock Example: Level of Virginity

	Responders	Item Non-responders	Unit Non-responders
Embarrassed	150 20%	75 20%	225
Not Embarrassed	450 10%	25 10%	75
Total	600	100	300

Estimate of level of virginity

Responders only: 12.5%

Responders + Item-nonresponders: 13.2%

Mock Example: Level of Virginity

	Responders	Item Non-responders	Unit Non-responders
Embarrassed	150 20%	75 20%	225 20%
Not Embarrassed	450 10%	25 10%	75 10%
Total	600	100	300

Estimate of level of virginity

Responders only:	12.5%
Responders + Item-nonresponders:	13.2%
Responders + Item-nonresponders + Unit-nonresponders	14.5%

Hepatitis C disease progression

- The Trent hepatitis C cohort follows patients sporadically through visits to hospital clinic
- For patients who attend clinic and are not lost to follow-up
  - Liver function tests (LFTs) are blood tests collected regularly.
  - Liver biopsies (invasive procedure) are infrequent and irregular. Each biopsy scored for stage of disease, *e.g.* 1 = Mild, 2 = Moderate, 3 = Cirrhosis
  - Other data collected at clinic visits (alcohol use, treatment regimes, BMI, end-stage liver diseases)

Process of Interest

Progress through biopsy states

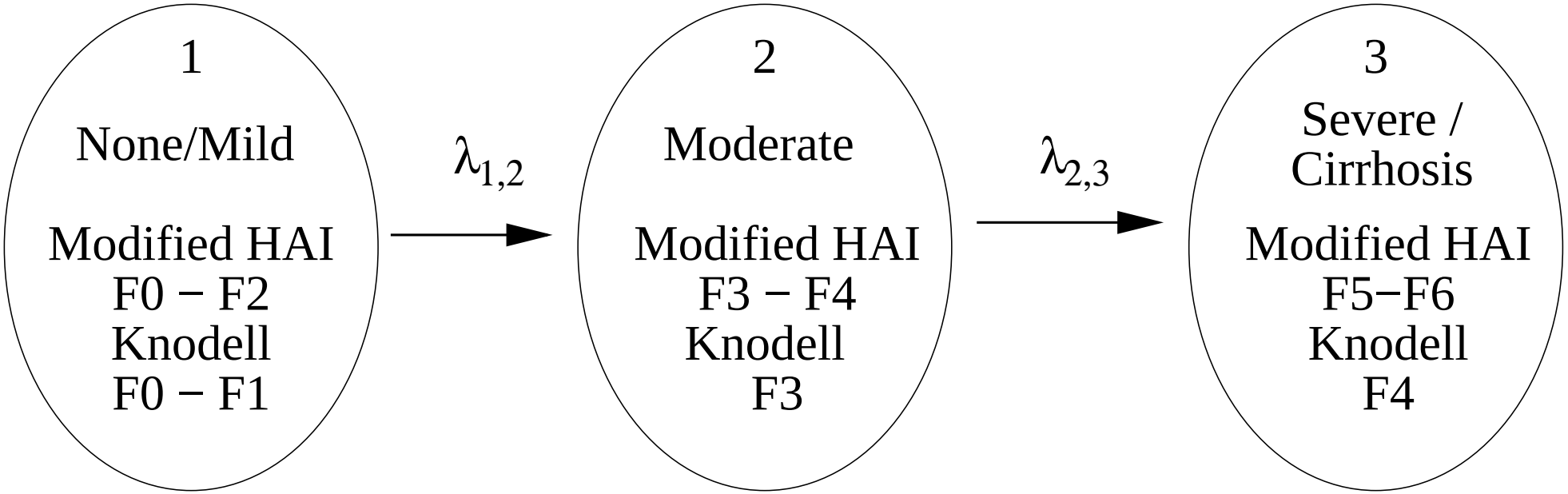


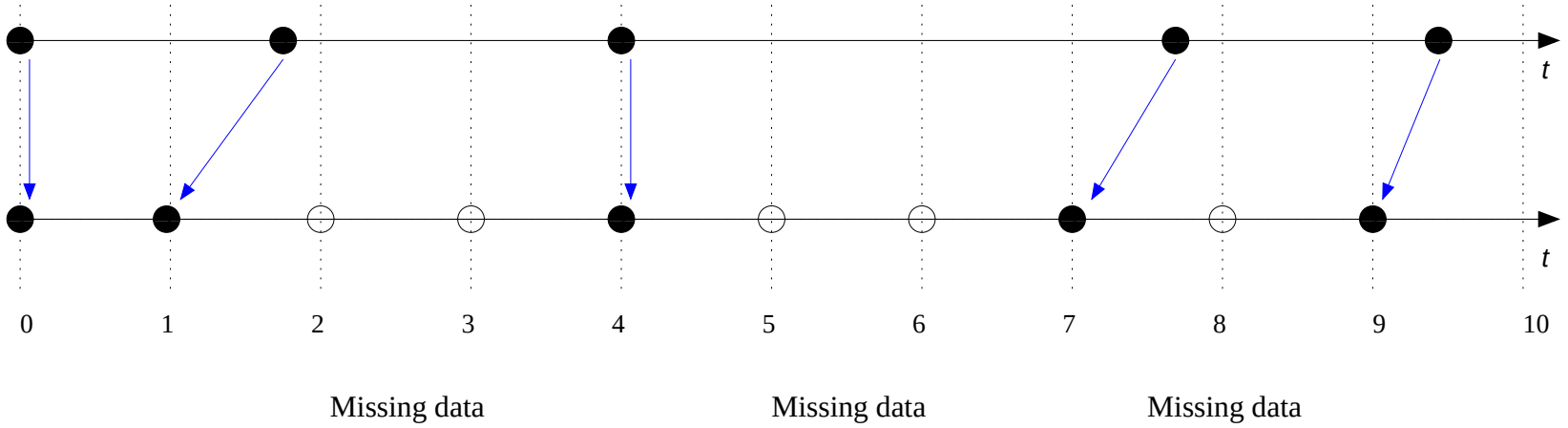
Figure 1: Fibrosis Model

The problem

- **Liver biopsies are the gold standard in assessing disease stage**
- **The occurrence of liver biopsies may be informative**
- **We need to jointly model the examination (liver biopsy) process and outcome process to obtain correct inferences.**
- **Can the much more frequently recorded LFTs help?**

Informative examination scheme as a missing data problem

- Consider whether a biopsy has occurred in each six-month period.
- Associate a single LFT value with each six-month period.



Notation

Observations denoted by $i = 1, \ldots, n$
Y_i **- categorical outcome at time** t_j (***e.g.*** **Stage of HCV disease)**

Notation

Observations denoted by $i = 1, \ldots, n$

Y_i - categorical outcome at time t_i (*e.g.* Stage of HCV disease)

R_i - missing data indicator variable equalling 1 if Y_i recorded at t_i, 0 otherwise

Notation

Observations denoted by $i = 1, \ldots, n$

Y_i - categorical outcome at time t_i (*e.g.* Stage of HCV disease)

R_i - missing data indicator variable equalling 1 if Y_i recorded at t_i, 0 otherwise

Z_i - explanatory variable(s) for outcome Y_i at time t_i

Notation

Observations denoted by $i = 1, \ldots, n$

Y_i - categorical outcome at time t_i (*e.g.* Stage of HCV disease)

R_i - missing data indicator variable equalling 1 if Y_i recorded at t_i, 0 otherwise

Z_i - explanatory variable(s) for outcome Y_i at time t_i

X_i - surrogate variable for outcome Y_i at time t_i

Notation

Observations denoted by $i = 1, \ldots, n$

Y_i - categorical outcome at time t_i (*e.g.* Stage of HCV disease)

R_i - missing data indicator variable equalling 1 if Y_i recorded at t_i, 0 otherwise

Z_i - explanatory variable(s) for outcome Y_i at time t_i

X_i - surrogate variable for outcome Y_i at time t_i

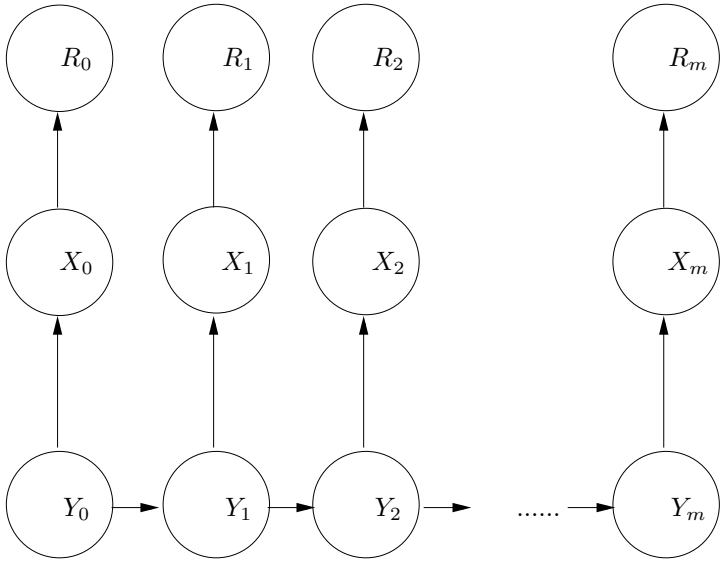
$\boldsymbol{Y}^o$ - vector of observed outcomes

$\boldsymbol{Y}^m$ - vector of missing outcomes

Approaches to joint modelling of Y and R

- **There are identifiability problems in estimating relationship between Y and R, since Y is unobserved when $R = 0$**
- **Assumptions MUST be made before carrying out any missing data analysis:**
  1. **A covariate dependent missing at random (CD-MAR) assumption $f(\boldsymbol{R}|\boldsymbol{Y}^o, \boldsymbol{Y}^m, \boldsymbol{Z}) = f(\boldsymbol{R}|\boldsymbol{Y}^o, \boldsymbol{Z})$.**
     If truly CD-MAR, then unbiased inferences can obtained using the observed data, and ignoring the missingness mechanism.
  2. **If not willing to assume CD-MAR given $\boldsymbol{Y}^o$ and $\boldsymbol{Z}$, must seek some extra information, X so that $f(\boldsymbol{R}|\boldsymbol{Y}^o, \boldsymbol{Y}^m, \boldsymbol{Z}, \boldsymbol{X}) = f(\boldsymbol{R}|\boldsymbol{Y}^o, \boldsymbol{Z}, \boldsymbol{X})$**

The Partially Hidden Markov model (PHMM)



The Partially Hidden Markov model Likelihood

The likelihood under this model is of the following form:

$$\propto \prod_{i=1}^{n} f(\boldsymbol{r}_i|\boldsymbol{x}_i, \boldsymbol{y}_i^o, \boldsymbol{z}_i, \psi) \sum_{\boldsymbol{y}_i^m} f(\{\boldsymbol{y}_i^m, \boldsymbol{y}_i^o\}|\boldsymbol{z}_i, \theta) f(\boldsymbol{x}_i|\{\boldsymbol{y}_i^m, \boldsymbol{y}_i^o\}, \boldsymbol{z}_i, \phi)$$

where ψ and ϕ denote the parameters defining the probability density functions $f(\boldsymbol{r}_i|\boldsymbol{x}_i, \boldsymbol{y}_i^o, \boldsymbol{z}_i)$ and $f(\boldsymbol{x}_i|\{\boldsymbol{y}_i^m, \boldsymbol{y}_i^o\}, \boldsymbol{z}_i)$, respectively.

- **Generalization (slightly) of CD-MAR since X cannot be regarded as a covariate for the Y process.**
- **Might be termed *Surrogate-Dependent MAR***

Simulation Study

- Simulation study of samples of 300 individuals observed at 5 time points.
- Exponential two-stage model, normally distributed auxiliary variable.
- Negative biases of 8% to 26% in estimation of baseline hazard if MCAR assumption is made incorrectly.
- Negative biases of 3% to 8% for a binary covariate (50% at each level) coefficient.
- PHMM eliminates these biases and gives appropriate coverage etc if observation depends on the auxiliary variable X.
- PHMM offers significant improvement even if MNAR model is correct.

Simulation Study

- **Data are generated for** 300 **individuals at five equally spaced examination times,** $(t_0, t_1, t_2, t_3, t_4)$**=(0,2,4,6,8)**
- $Z = -1$ **for 50% of individuals and** 1 **otherwise.**
- **Transition time out of state 1 is exponential**
  - $T|z \sim$ **Exponential**$(\lambda_0 e^{\beta z}), \lambda_0 = 0.2, \beta = 0.5$**.**
- **The binary response** $Y(t_j|T) = \mathbf{I}[T \leq t_j]$ **indicator for transition by** t_j**.**
- **The auxiliary variables,** X**, are normally distributed**
  - $(X(t_j)|y(t_j)) \sim$ **Normal**$(\mu_{y(t_j)}, \sigma^2)$
  - $\mu_{y(t_j)} = \phi_0 + \phi_1 y(t_j)$**,** σ**=1 (independent of** Y**)**
- **Missing data process is Bernoulli**
  - $\mathbf{Pr}\left(R(t_j) = 1|y(t_j), x(t_j)\right) = \mathbf{logit}^{-1}\left\{\psi_0 + \psi_1 y(t_j) + \psi_2 x(t_j)\right\}.$

Simulation results for the baseline log hazard

Scenario	Relative bias (%)			95% coverage (%)			MSE		
	IG	PHMM	MNAR	IG	PHMM	MNAR	IG	PHMM	MNAR
MCAR, $\psi_1 = \psi_2 = 0$									
1) X **independent of** Y, $\phi_1 = 0$	**0.1**	**0.1**	**0.1**	**95.9**	**96.0**	**94.5**	**0.009**	**0.009**	**0.013**
2) $\phi_1 = 0.5$	**0.1**	**0.1**	**0.1**	**95.9**	**95.4**	**94.5**	**0.009**	**0.008**	**0.013**
3) $\phi_1 = 1$	**0.1**	**0.1**	**0.1**	**95.9**	**94.7**	**94.5**	**0.009**	**0.008**	**0.013**
MAR, $\psi_1 = 0$, $\psi_2 = 1$									
1) X **independent of** Y, $\phi_1 = 0$	**0.2**	**0.2**	**0.2**	**96.1**	**95.7**	**95.5**	**0.008**	**0.008**	**0.011**
2) $\phi_1 = 0.5$	**-8.4**	**0.2**	**0.2**	**64.2**	**96.0**	**95.2**	**0.026**	**0.008**	**0.011**
3) $\phi_1 = 1$	**-16.8**	**0.2**	**0**	**12.0**	**95.0**	**95.1**	**0.081**	**0.007**	**0.011**
MNAR, $\psi_1 = \psi_2 = 1$									
1) X **independent of** Y, $\phi_1 = 0$	**-13.5**	**-14.0**	**0.0**	**21.8**	**19.1**	**96.0**	**0.053**	**0.057**	**0.008**
2) $\phi_1 = 0.5$	**-19.9**	**-13.1**	**0.1**	**1.1**	**25.0**	**95.4**	**0.109**	**0.051**	**0.008**
3) $\phi_1 = 1$	**-26.4**	**-9.9**	**0.0**	**0.0**	**48.7**	**94.1**	**0.188**	**0.032**	**0.008**

Simulation results for the binary covariate coefficient

Scenario	Relative bias (%)			95% coverage (%)			MSE		
	IG	PHMM	MNAR	IG	PHMM	MNAR	IG	PHMM	MNAR
MCAR, $\psi_1 = \psi_2 = 0$									
1) X **independent of** Y, $\phi_1 = 0$	**0.4**	**0.3**	**0.4**	**94.8**	**94.8**	**94.9**	**0.009**	**0.009**	**0.009**
2) $\phi_1 = 0.5$	**0.4**	**0.2**	**0.4**	**94.8**	**95.4**	**94.9**	**0.009**	**0.008**	**0.009**
3) $\phi_1 = 1$	**0.4**	**0.1**	**0.4**	**94.8**	**95.0**	**94.9**	**0.009**	**0.007**	**0.009**
MAR, $\psi_1 = 0$, $\psi_2 = 1$									
1) X **independent of** Y, $\phi_1 = 0$	**0.4**	**0.3**	**0.3**	**94.5**	**94.4**	**94.3**	**0.008**	**0.008**	**0.008**
2) $\phi_1 = 0.5$	**-3.0**	**0.2**	**0.3**	**93.4**	**94.5**	**94.4**	**0.008**	**0.008**	**0.008**
3) $\phi_1 = 1$	**-5.8**	**0.4**	**0.6**	**91.9**	**94.7**	**94.3**	**0.009**	**0.007**	**0.008**
MNAR, $\psi_1 = \psi_2 = 1$									
1) X **independent of** Y, $\phi_1 = 0$	**-4.2**	**-4.5**	**0.2**	**92.4**	**92.3**	**94.8**	**0.007**	**0.007**	**0.007**
2) $\phi_1 = 0.5$	**-6.1**	**-1.8**	**0.3**	**90.6**	**93.6**	**94.4**	**0.008**	**0.007**	**0.007**
3) $\phi_1 = 1$	**-7.8**	**0.6**	**0.4**	**90.0**	**94.9**	**95.4**	**0.008**	**0.006**	**0.006**

Transitions in Trent Cohort Database

From state	To state 'None/Mild'	'Moderate'	'Severe/Cirrhosis'	Unknown
'None/Mild'	326	20	6	403
'Moderate'	0	8	6	109
'Severe/Cirrhosis'	0	0	2	100

Observed disease state transitions in the Trent hepatitis C cohort.

Basline Hazards Estimates

Parameter	Model		
	Ignorable	CD-MAR (ALT)	MNAR
Baseline intensities			
$\lambda_{1,2}$	**0.0120**	**0.0119**	**0.0119**
	(0.0079, 0.0182)	**(0.0078, 0.0181)**	**(0.0078, 0.0181)**
$\lambda_{2,3}$	**0.0773**	**0.0769**	**0.0794**
	(0.0396, 0.1509)	**(0.0399, 0.1485)**	**(0.0386, 0.1634)**

A Different Sort of Example

Return to NATSAL (Survey of Sexual Attitudes and Lifestyle)

- Two surveys in 1990 and 2000
- Interested in changes between 1990 and 2000
- As seen before, bias is expected in each survey.
- Change in bias is relevant to any examination of change in results of the surveys

Comparison of NATSAL-1990 with NATSAL-2000

- Bias will depend on the question.
- Classify questions to be of high, medium and low sensitivity (effectively reflecting expected bias).
- Should any information be the same in the two surveys?
- Population cohort eligible for both surveys are those:
  - Aged 16-34 in 1990.
  - Aged 26-44 in 2000.
- Questions answered by this *common cohort* should be similar if they, e.g., refer to events before a fixed age.

Comparison of NATSAL-1990 with NATSAL-2000

- **Homosexual experience before 1990 [High sensitivity]:**
  - **Men: 5.0% (1990) vs 8.5% (2000)**
  - **Women: 3.5% (1990) vs 6.7% (2000)**
- **Heterosexual intercourse before 16 years [Medium sensitivity]:**
  - **Men: 24.7% (1990) vs 27.5% (2000)**
  - **Women: 12.9% (1990) vs 18.2% (2000)**

From these type of questions. estimate odds ratios (ORs) for change in bias

- **High sensitivity: Men 1.80(1.46,2.21); Women 1.99(1.62,2.46)**
- **Medium sensitivity: Men 1.11(1.01,1.21); Women 1.19(1.10,1.29)**

Comparison of NATSAL-1990 with NATSAL-2000

•

Homosexual partners, past 5 years

- **Men: 1.5% (1990) vs 2.6% (2000) → OR: 1.75(1.29,2.36)**
  - **:**
- **Women: 0.8% (1990) vs 2.6% (2000) → OR: 3.43(2.42,4.87)**
  - **:**

Comparison of NATSAL-1990 with NATSAL-2000

Change in bias results

- **High sensitivity OR: Men 1.80; Women 1.99**

Homosexual partners, past 5 years

- **Men: 1.5% (1990) vs 2.6% (2000) → OR: 1.75(1.29,2.36)**
  - **Minimum established change: 1.29/1.80 = 0.72**
- **Women: 0.8% (1990) vs 2.6% (2000) → OR: 3.43(2.42,4.87)**
  - **Minimum established change: 2.42/1.99 = 1.22**

Concluding Remarks

- **Classifications of missing data structures are useful.**

Concluding Remarks

- **Classifications of missing data structures are useful.**
- **Such structures can sometimes give the impression that the solution to missing data is then simply to model the structure.**

Concluding Remarks

- Classifications of missing data structures are useful.
- Such structures can sometimes give the impression that the solution to missing data is then simply to model the structure.
- The collection and use of auxiliary information which is directly linked to missing or informatively collected data should be sought in such modelling efforts.

Concluding Remarks

- Classifications of missing data structures are useful.
- Such structures can sometimes give the impression that the solution to missing data is then simply to model the structure.
- The collection and use of auxiliary information which is directly linked to missing or informatively collected data should be sought in such modelling efforts.
- The type of information and appropriate model is likely to be application specific.

Concluding Remarks

- Classifications of missing data structures are useful.
- Such structures can sometimes give the impression that the solution to missing data is then simply to model the structure.
- The collection and use of auxiliary information which is directly linked to missing or informatively collected data should be sought in such modelling efforts.
- The type of information and appropriate model is likely to be application specific.
- Caution is still strongly advised.