Using pattern mixture modelling to reduce bias due to informative attrition in the Whitehall II study: a simulation study

Catherine Welch[1] Martin Shipley[1] Séverine Sabia[2]
Eric Brunner[1] Mika Kivimäki[1]

[1] Research Department of Epidemiology and Public Health, University College London
[2] INSERM U1018, Centre for Research in Epidemiology and Population Health, Villejuif, France

September 7, 2016

Outline

Introduction

- Informative attrition can bias longitudinal studies
  - reason for attrition associated with missing outcome values
- Multiple imputation (MI) assumes missing at random - not appropriate
- Clinical trials use pattern mixture modelling (PMM), monotone data simplifies analysis
- Observational studies non-monotone, more complex

Whitehall II cohort study

- 10,308 London civil servants, began 1985
- Health and lifestyle questionnaire completed every 2-3 years (phase), clinic at odd phases
- Epidemiological investigation:
  - Smoking status at baseline (Phase 5) is associated with 10-year cognitive decline
  - Attrition maybe informative, participants with reduced cognitive function withdraw
  - Replaced missing values with last observed value

Objectives

- Simulation study to investigate using pattern mixture modelling to reduce bias caused by informative attrition in longitudinal observational data
- Using Stata, create 1,000 datasets (10,000 participants) replicating the smoking-cognitive function analysis
- Make values missing using missing not at random (MNAR) missingness mechanisms
- Compare bias in intercept and slope
  - Simulated data (no missing values)
  - Complete case analysis
  - Analyse data imputed using MI
  - PMM sensitivity analysis

Outline

Substantive model

- Memory score (y_{ij}) for participant j at time i [1]
- Standardised using mean and standard deviation from baseline
- Stratified by sex - this analysis includes just men

Mixed effects model with random intercept and slope with interactions between coefficients and time

$$y_{ij} = \beta_0 + \beta_1 smoke_{5j} + \beta_1 smoke_{5j} time_{ij} + U_{0j} + U_{1j} time_{ij} + \varepsilon_i$$

Model also included participant characteristics at baseline (age, occupation grade and education) and their interactions with time

Generating missing values

- Participation status
  - Responder - participated at a given phase, may have item non-response
  - Non-responder - unit non-response
  - Confirmed death
- MAR - conditional on age, education and occupational grade at baseline
- If responders with item non-response, non-responder or died, replace y_{ij} with missing value

Withdrawn

- Informed Whitehall II they no longer wish to participate
- Participants withdraw at Phases 7, 9 and 11
- Informative (missing not at random)
  - Participants j and phase i assign withdrawal probability p_{ij} conditional on memory score at the same phase Y_{ij}

$$logit(p_{ij}) \quad = \quad \lambda_0 + \lambda_1 Y_{ij}$$

  - Selected λ_0 and λ_1 to achieve similar percentage withdrawn as Whitehall II study
  - Lower memory scores more likely to withdraw

Summary of multiple imputation

- Specify imputation model, which generates plausible values to replace missing values
- Generate M imputations for each missing value, creating M completed datasets
- Analyse each imputed dataset separately
- Pool estimates and standard errors - Rubins rules [2]
- Validity relies on plausible assumptions [3]
  - MAR missingness mechanism
  - Substantive model and imputation model are congenial

Stata command `twofold`

- The two-fold fully conditional specification algorithm [4]
- Suitable for longitudinal data [5]
- Imputes each time point in turn conditional on observations at adjacent time points (time window)
  - Within-time iteration - imputes missing values in time window
  - Among-time iteration - time window imputes at each time point
- No interactions with time because phases imputed separately
- Available from SSC repository [6]

`twofold` syntax

(data in wide form)

```
.
gen start = 3
gen end = 11 (or phase participant died)
gen base = 5
.
twofold, timein(start) timeout(end) base(base)
depmis(mem exsmoke) indobs(agec5 grade academ nonsmoke)
conditionon(nonsmoke) condval(0) condvar(exsmoke)
indmis(smkstop5) clear cat(nonsmoke exsmoke grade academ)
m(20) ba(20) bw(5) seed(100)
.
mi reshape long ...
.
mi estimate:  mixed mem b4.smokebase##c.time c.agec5##c.time
i.grade##c.time i.academ##c.time || stno:  time
```

Pattern mixture modelling

- Specify separate distributions for the observed and missing data [7]
- Distribution of observed outcomes - substantive model

$$y_{ij} = \beta_0 + \beta_1 smoke_{5j} + \beta_1 smoke_{5j} time_{ij} + U_{0j} + U_{1j} time_{ij} + \varepsilon_i$$

- Withdrawn indicator R_{ij}
- Distribution of missing outcomes - for withdrawn, use substantive model and change by k in the imputed outcome

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 smoke_{5j} + \beta_1 smoke_{5j} time_{ij} + \\ &\quad U_{0j} + U_{1j} time_{ij} + \varepsilon_i + kR_{ij} \end{aligned}$$

- For withdrawn participants, change already imputed y_{ij} values by k
- Sensitivity analysis: k=-0.2, -0.4, -0.6, -0.8 and -1.0

Outline

Simulated participation status

- 6,210 male participants from Whitehall II study

Whitehall II study

Participation Status	**5**	**7**	**9**	**11**
Participated,%	88.1	78.8	76.6	71.8
Died, %	N/A	2.6	5.9	10.1
Non-response, %	11.9	14.6	12.2	11.8
Withdraw, %	N/A	4.0	5.3	6.3

Simulated data

Participation Status	**5**	**7**	**9**	**11**
Participated,%	89.6	80.3	78.1	73.3
Died, %	N/A	2.4	5.5	9.0
Non-response, %	10.4	13.6	11.2	11.0
Withdraw, %	N/A	3.8	5.3	6.6

Analysing simulated data, mean

Simulated data, complete case and imputed data estimates averaged over 1,000 datasets

Smoking status at baseline		**WII study**	**Simulated data**	**Complete Case**	**Multiple imputation**
Intercept	Current smoker	-0.080	-0.079	-0.140	-0.051
	Recent ex-smoker	-0.081	-0.079	-0.138	-0.016
	Long-term ex-smoker	0.071	0.073	0.004	0.098
	Never smoker	0.026	0.027	-0.039	0.057
Slope	Current smoker	-0.412	-0.414	-0.354	-0.338
(per 10 years)	Recent ex-smoker	-0.313	-0.316	-0.264	-0.282
	Long-term ex-smoker	-0.409	-0.410	-0.366	-0.368
	Never smoker	-0.354	-0.355	-0.311	-0.311

Also adjusted for age, education and employment grade and interactions with time

Pattern mixture modelling results, mean

Simulated data, imputed and pattern mixture modelling estimates averaged over 1,000 datasets

Smoking status at baseline		WII study	Imputed data	Pattern mixture modelling (k) -0.2	-0.4	-0.6	-0.8	-1.0
Intercept	Current	-0.079	-0.051	-0.051	-0.054	-0.056	-0.057	-0.059
	Recent ex	-0.079	-0.016	-0.016	-0.019	-0.021	-0.022	-0.024
	Long-term ex	0.073	0.098	0.096	0.094	0.093	0.091	0.090
	Never	0.027	0.057	0.056	0.055	0.054	0.053	0.051
Slope (per 10 years)	Current	-0.414	-0.338	-0.360	-0.383	-0.406	-0.429	-0.452
	Recent ex	-0.316	-0.282	-0.304	-0.324	-0.346	-0.367	-0.388
	Long-term ex	-0.410	-0.368	-0.388	-0.407	-0.427	-0.448	-0.468
	Never	-0.355	-0.311	-0.328	-0.345	-0.362	-0.378	-0.395

Also adjusted for age, education and employment grade and interactions with time

Outline

Conclusions

- Results suggest pattern mixture modelling and the two-fold fully conditional specification algorithm may reduce bias due to informative attrition in longitudinal, observational data
- In this example, PMM reduced bias in the slope due to participants withdrawing after baseline
- Reduced bias in main effect for time and interaction with time
- Recommend considering an appropriate approach as sensitivity analysis if suspect attrition is informative
- Next: apply these methods to impute missing values for withdrawn participants in Whitehall II study

Whitehall II Data Sharing

The Whitehall II research data are available to *bona fide* researchers for research purposes and public benefit.
Please visit our website on:

http://www.ucl.ac.uk/whitehallII/data-sharing

References I

S. Sabia, A. Elbaz, A. Dugravot, J. Head, M. Shipley, G.H. Hagger-Johnson, M. Kivimaki, and A. Singh-Manoux.
Impact of smoking on congitive decline in early old age.
Arch Gen Psychiatry, 69(6):627–635, 2012.

D.B. Rubin.
Multiple imputation for nonresponse in surveys.
Wiley, New York, 1987.

J. Carpenter and M.G. Kenward.
Multiple Imputation and its Application.
Wiley, UK, 2013.

References II

J. Nevalainen, M.G. Kenward, and S.M. Virtanen.
Missing values in longitudinal dietary data: a multiple imputation approach based on a fully conditional specification.
Statistics in Medicine, 28(29):3657–3669, 2009.

C. Welch, Petersen I., J. Bartlett, I. White, L. Marston, R. Morris, I. Nazareth, K. Walters, and J. Carpenter.
Evaluation of two-fold fully conditonal specification multiple imputation for longitudinal electronic health record data.
Stat.Med., 33(21):3725–3737, 2014.

C. Welch, J. Bartlett, and Petersen I.
Application of multiple imputation using the two-fold fully conditional specification algorithm in longitudinal clinical data.
Stata Journal, 14(2):418–431, 2014.

References III

D. Hedeker and R.D. Gibbons.

Application of random-effects pattern-mixture models for missing data in longitudinal studies.

Psychological Methods, 2(1):64–78, 1997.