University of BRISTOL

Dealing with missing data in physical activity models

Alex Griffiths
University of Bristol

How much data is missing?

- Three time points (~12, 14 and 16 years)
- Up to 7 days accelerometry at each age
  - Weekly average used in analysis
  - ≥3 days for reasonable reliability
- Numbers with valid data:

Age	Invited to clinic	≥1 day	≥3 days
12	11 952	6 112	5 741
14	11 267	4 423	3 885
16	10 692	2 246	2 099

- 6268 with valid data at one or more time points

Variables predicting drop-out

- Boys
- Lower birth weight
- Lower social class
- Younger mothers
- Mothers with fewer educational qualifications
- Higher physical activity at earlier ages
- Higher fat mass at earlier ages (girls only)

Complete case analysis

- PA and fat mass at 12 and 14
- Multilevel model allows subjects with data at only one age to be included
- BUT:
- Loss of power (n=4 614 with data at one or more time points plus confounders)
- Bias?

Prospective associations between objective measures of physical activity and fat mass in 12-14 year old children: the Avon Longitudinal Study of Parents and Children (ALSPAC)

Chris J Riddoch, professor of sport and exercise science,[1] Sam D Leary, lecturer in statistics,[2,3] Andy R Ness, professor of epidemiology and codirector of the Avon longitudinal study of parents and children,[2,3] Steven N Blair, professor of epidemiology,[4] Kevin Deere, research assistant,[3] Calum Mattocks, research fellow,[1] Alex Griffiths, research assistant in statistics,[2] George Davey Smith, professor of clinical epidemiology,[5] Kate Tilling, senior lecturer in medical statistics[3]

Types of missing data

- Missing Completely At Random (MCAR)
  - Probability that data is missing does not depend on true value of missing data itself or on other variables
  - Complete case analysis unbiased
- Missing At Random (MAR)
  - Probability that data is missing does not depend on true value of missing data but may depend on other (observed) variables
- Missing Not At Random (MNAR)
  - Probability that data is missing depends on true value of missing data

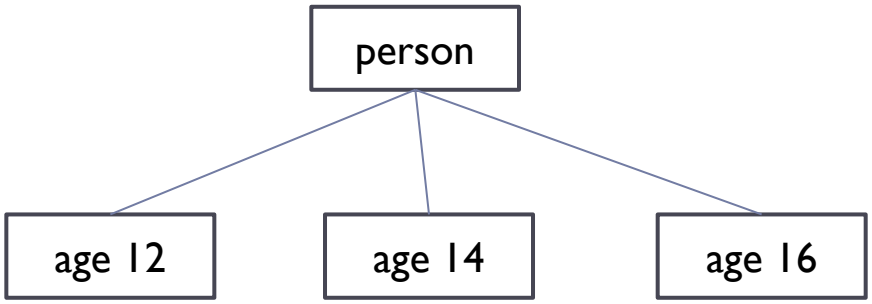
Multiple imputation

- Two stages:
  1. Create m (≥ 2) imputed datasets with each missing value filled in
  2. Analyse each imputed (complete) dataset using standard methods, and combine the results appropriately
- Multiple Imputation by Chained Equations (MICE)
  - For each variable with missing data, regress observed values on all other variables
  - Draw from predictive distribution to create imputed values
  - Repeat...

Variables to include

- Outcome and exposure of interest
- Possible confounders
- Variables predictive of missingness

Imputing missing PA data: 2-level model



- Multilevel model for PA – sum of residuals for each person at each age (MLwiN)
- Impute missing residuals and covariates (Stata `ice`)
- Regress outcome on PA residuals (Stata `mim`)

Daily PA data

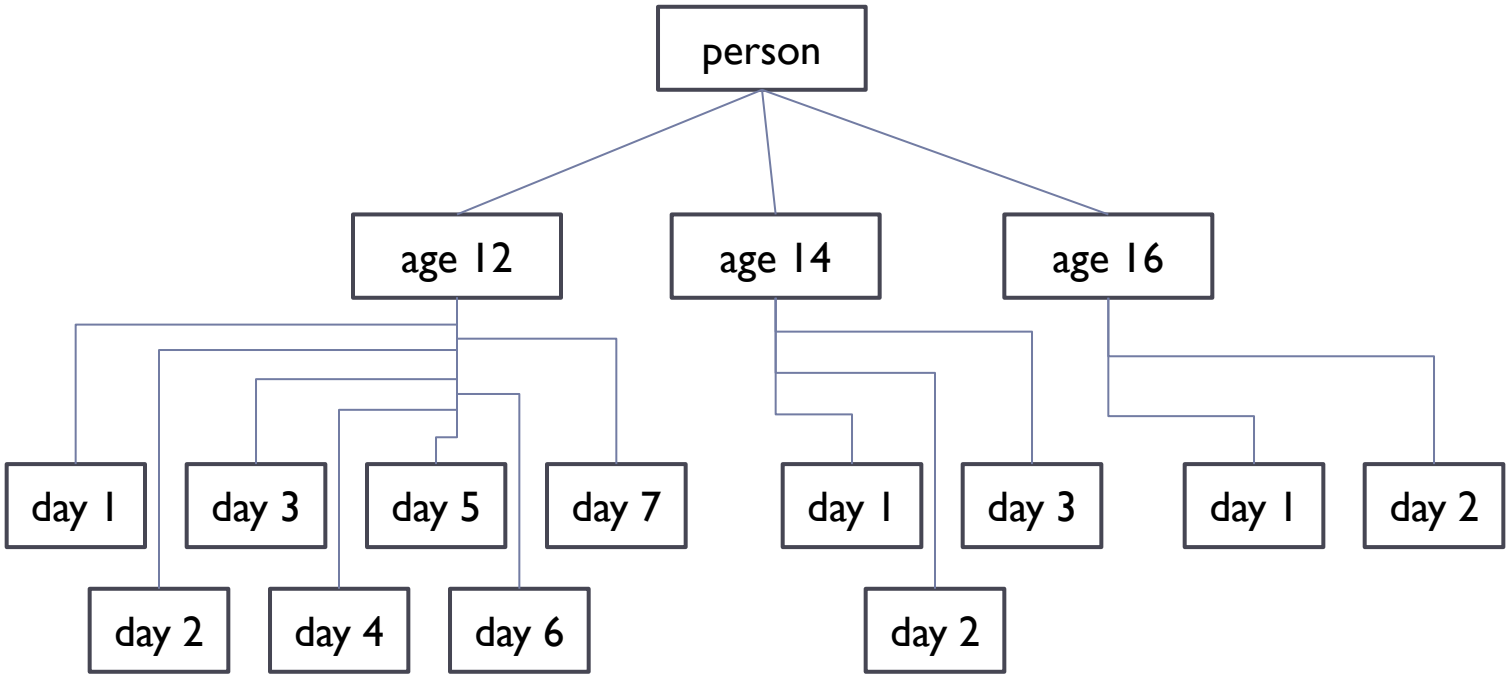
Age	Invited to clinic	≥1 day	≥3 days
12	11 952	6 112	5 741
14	11 267	4 423	3 885
16	10 692	2 246	2 099

- PA data only included if ≥3 valid days
- Is there a way to use the partial data (1 or 2 days) in the imputation model?

Plan A: impute individual days

- Impute missing days if ≥1 valid day – calculate weekly average (passive imputation)
- Impute weekly average if no valid days
- Then fit multilevel model in Stata (`xtmixed`)
  - Too computationally intensive in practice

Plan B: 3-level multilevel model



- Fit multilevel model to day-by-day PA data
- Calculate sum of residuals from top two levels only
- Impute missing residuals and covariates as before

Results

PA at age 12		% change in fat mass at age 16	
		Boys (n=2988)	Girls (n=3280)
+100 counts/min total activity	Unimputed	-8.5% (-10.3%, -6.6%)	-4.2% (-5.6%, -2.8%)
	Imputed	-7.5% (-9.0%, -6.0%)	-3.5% (-4.7%, -2.3%)
+ 15 mins/day MVPA	Unimputed	-14.7% (-17.6%, -11.7%)	-8.5% (-11.2%, -5.8%)
	Imputed	-14.3% (-16.7%, -11.7%)	-7.9% (-10.2%, -5.6%)

95% confidence intervals in brackets

- Associations between PA and fat mass weaker after imputation
- Gain in precision

Hourly PA data?

- Valid day = 10 hours
- "Green sheets"
  - Record activities if accelerometer was taken off
- Possible to incorporate into imputation model?