Multiple imputation: an overview & historical perspective

John Carlin

Clinical Epidemiology & Biostatistics Unit, Murdoch Childrens Research Institute
School of Population Health, University of Melbourne

26 April 2012

Outline

- What is multiple imputation (MI)?
- Illustrative example
- Foundations and basic theory
- Software implementation; MI in practice
- A case study (Victorian Adolescent Health Cohort Study)
- Where to from here?

What is MI?

Begin with simple "algorithmic" description of the method, requiring some notation:

- Focus of interest: rectangular dataset n 'units' by p 'variables'
- i.e. matrix Y of n exchangeable rows $\mathbf{Y_i'}$ each length p

Data analyst wishes to perform analysis using all of the variables in Y: often a linear model or regression analysis relating one of the variables to the others. . .

- For current purposes no need to distinguish between "response" or "outcome" and "predictors" or "covariates"
- For now assume that any of values in Y may be missing, while desired analysis requires all values to be available
- Standard (packaged) statistical procedure will proceed by including only those units that contain no missing values: so-called "complete-case analysis" or "listwise deletion"

Complete-case analysis may use only a small subset of data rows, raising concerns about both potential biases and loss of precision in inferences for the target parameters.

MI in brief

1. Create m copies of the incomplete dataset, and use appropriate procedure to *impute* (fill in) the missing values in each of these copies
2. For each completed copy of dataset, perform standard analysis (as would have been in absence of missing values), and store the parameter estimates of interest, along with their estimated variances (SEs)
3. Use formulas widely known as "Rubin's Rules," firstly to create a combined estimate of the parameter (as average of the m separate estimates) and then to obtain a standard error for this estimate...

Basic MI formulas (Rubin's rules of combination)

- Assume single (univariate) parameter of interest, β.
- From $k^{th}(k = 1, \ldots, m)$ imputed dataset obtain $\hat{\beta}^{(k)}$ with (estimated) variance $V^{(k)}$
- Combined estimate of β:

$$\hat{\beta}^{\mathrm{MI}} = \frac{1}{m} \sum_{1}^{m} \hat{\beta}^{(k)} \tag{1}$$

Basic MI formulas (Rubin's rules of combination)

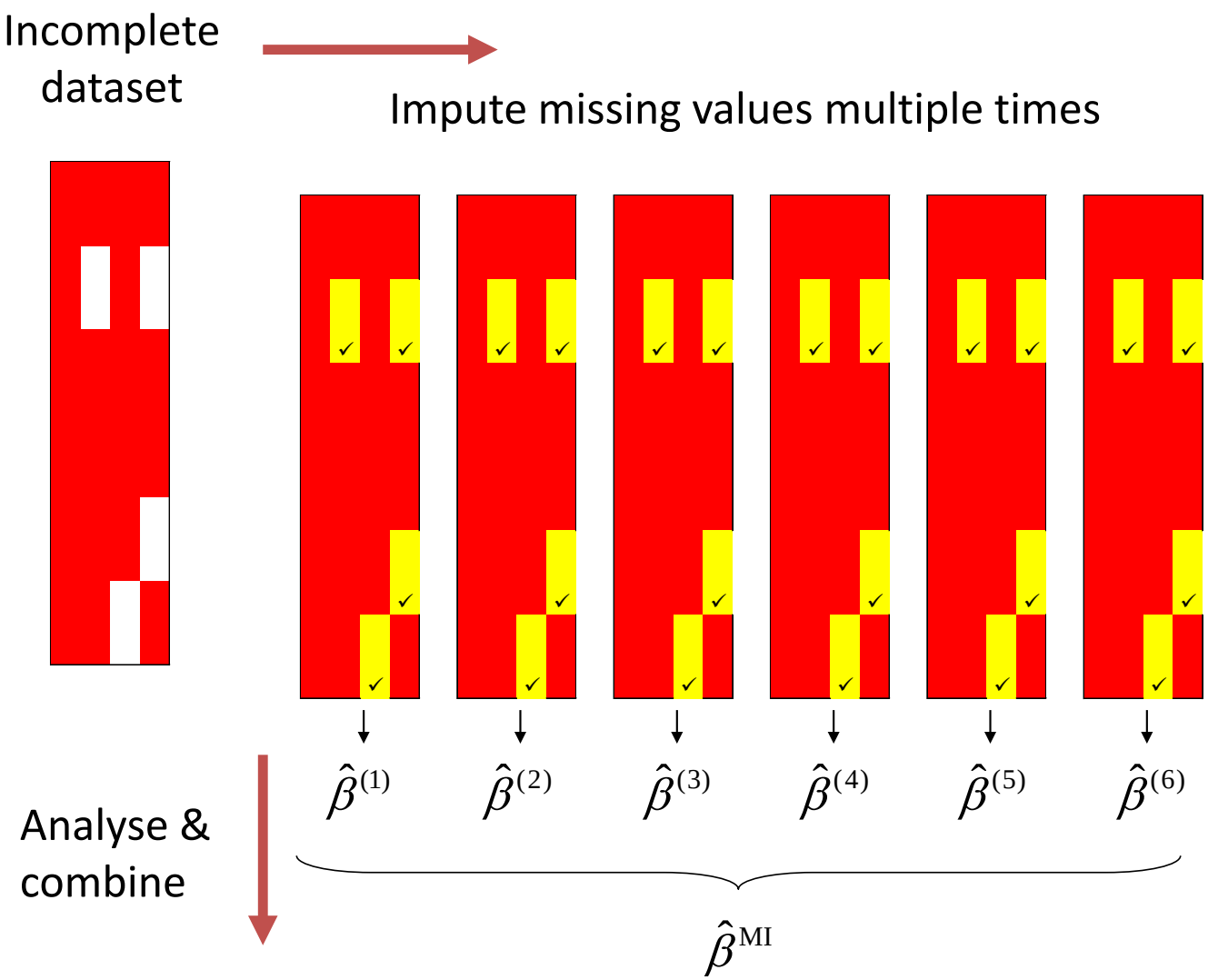
- Combined variance of β:

$$V^{\mathrm{MI}} = \bar{V} + \left(1 + \frac{1}{m}\right) B, \tag{2}$$

where $\bar{V} = \sum_1^m V^{(k)}/m$, and $B = \sum_1^m (\hat{\beta}^{(k)} - \hat{\beta}^{\mathrm{MI}})^2/m$, which estimates the *between-imputation* variance of the parameter of interest

- Form test statistics and confidence intervals in usual way, assuming that $(\hat{\beta}^{\mathrm{MI}} - \beta)/\sqrt{V^{\mathrm{MI}}}$ follows either standard normal or t distribution

Schematic illustration



Attractive heuristic properties

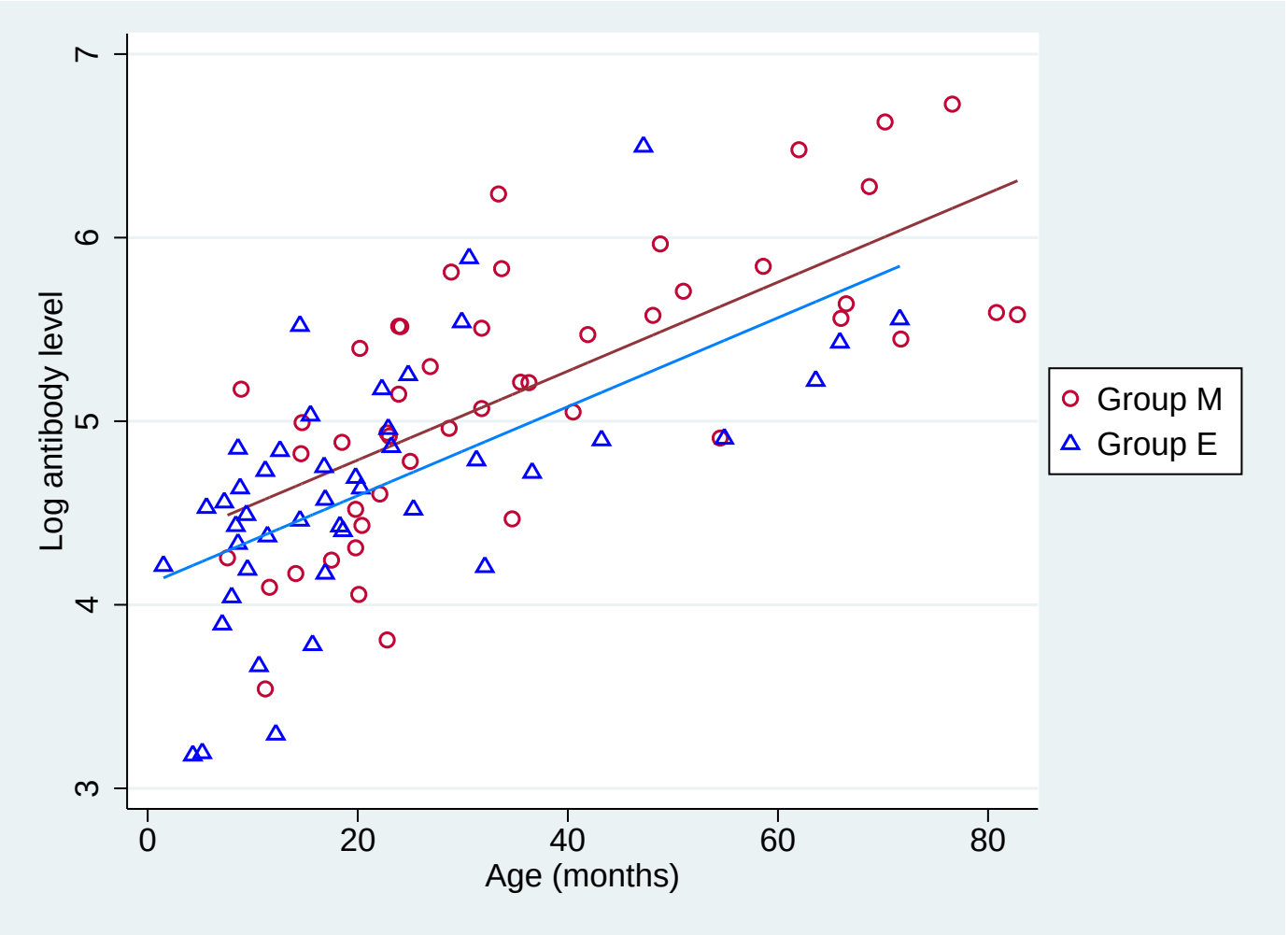
- Process of filling in or imputing has intuitive appeal of "restoring" the dataset that we wanted to have, while having multiple different versions reminds there is no way to recover the *actual* unknown missing values
- Core work of performing the analysis of interest (in each completed dataset) follows exactly the approach that would have been used in absence of missing data
- Emphasize, however, that imputed datasets should NOT be taken to represent true substitutes or true "completions" of the actual dataset

Attractive heuristic properties

- Variance formula has two components:
  1. First = average of *within-imputation* variances from each of completed datasets
  2. Second adds an amount that reflects *between-imputation* variance of parameter estimates—makes sense if imputation process validly reflects the uncertainty due to the missing data

So how is imputation done "properly," to ensure validity? We return to this question after an illustrative example and review of basic theory

Illustrative example: analysis of covariance



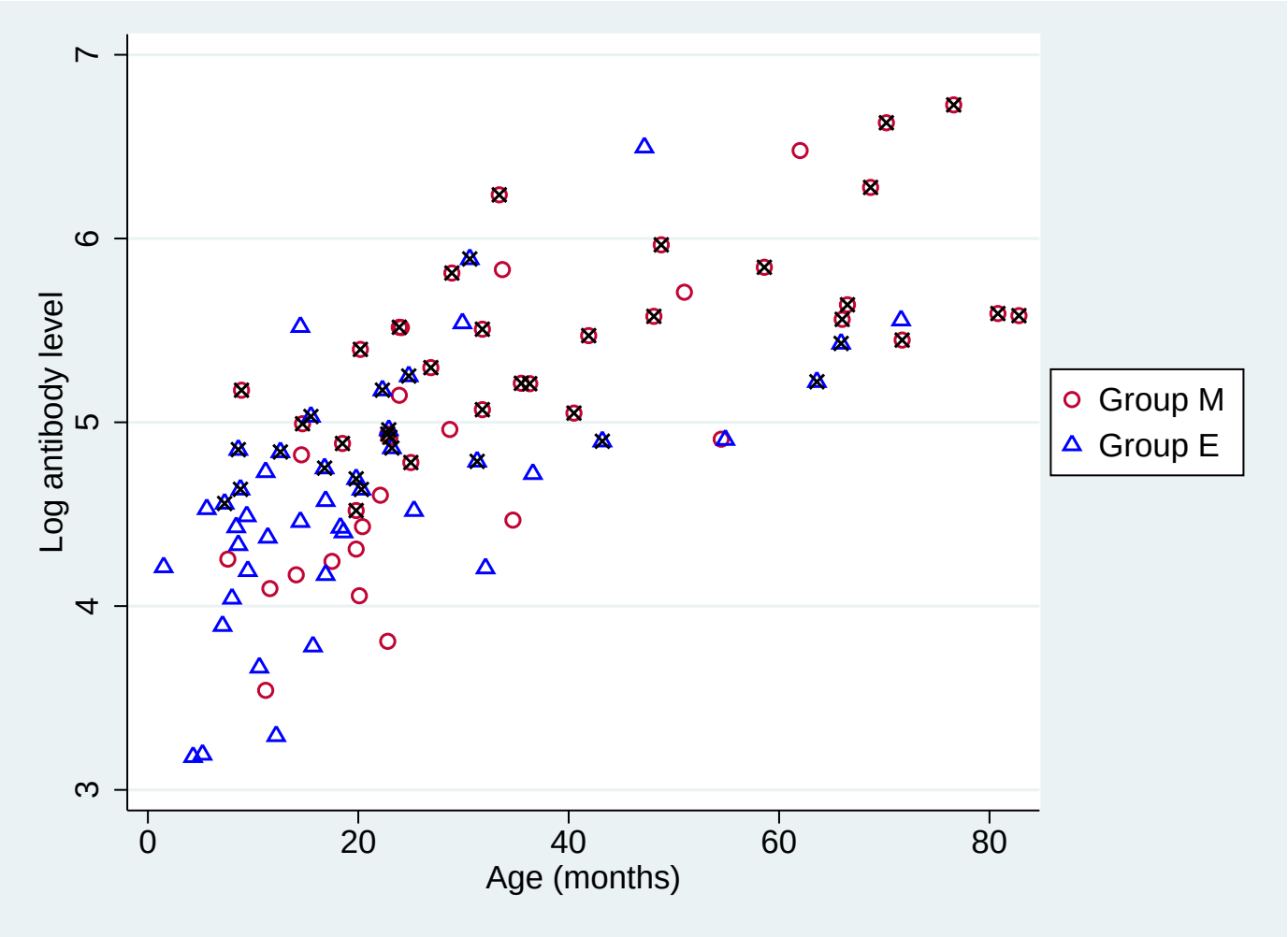
Group comparison by ANCOVA adjusting for age ($n = 93$)

Illustrative example: estimation results

Original data

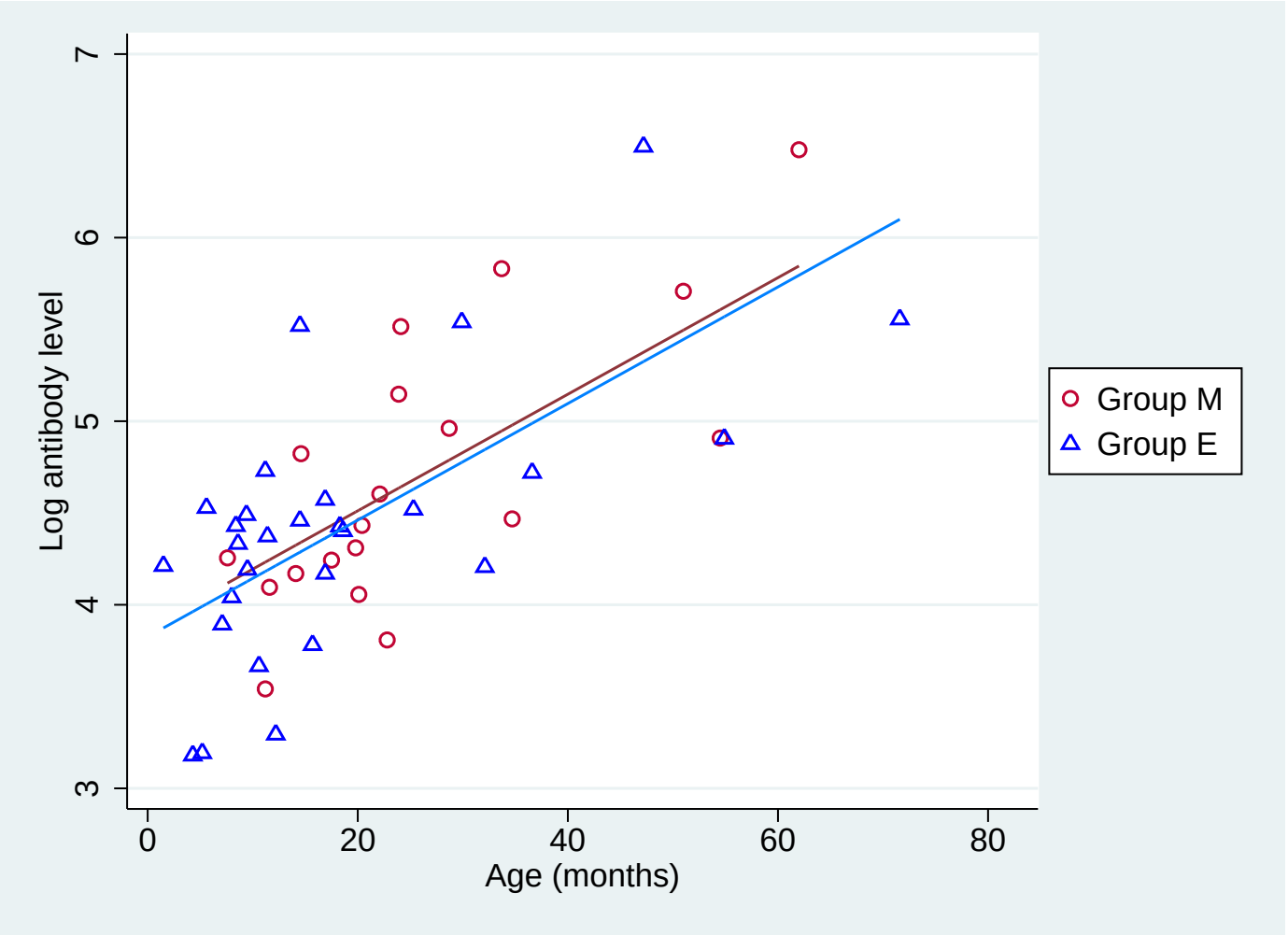
	full sample (n=93)		missing 50% age values (n = 47)
	Crude	ANCOVA	
group diff	-0.542	-0.194	
(SE)	(0.145)	(0.117)	
age effect	-	0.0242	
(SE)	-	(0.0029)	

Illustrative example: analysis of covariance



. . . now suppose 50% of age values go missing

Illustrative example: analysis of covariance



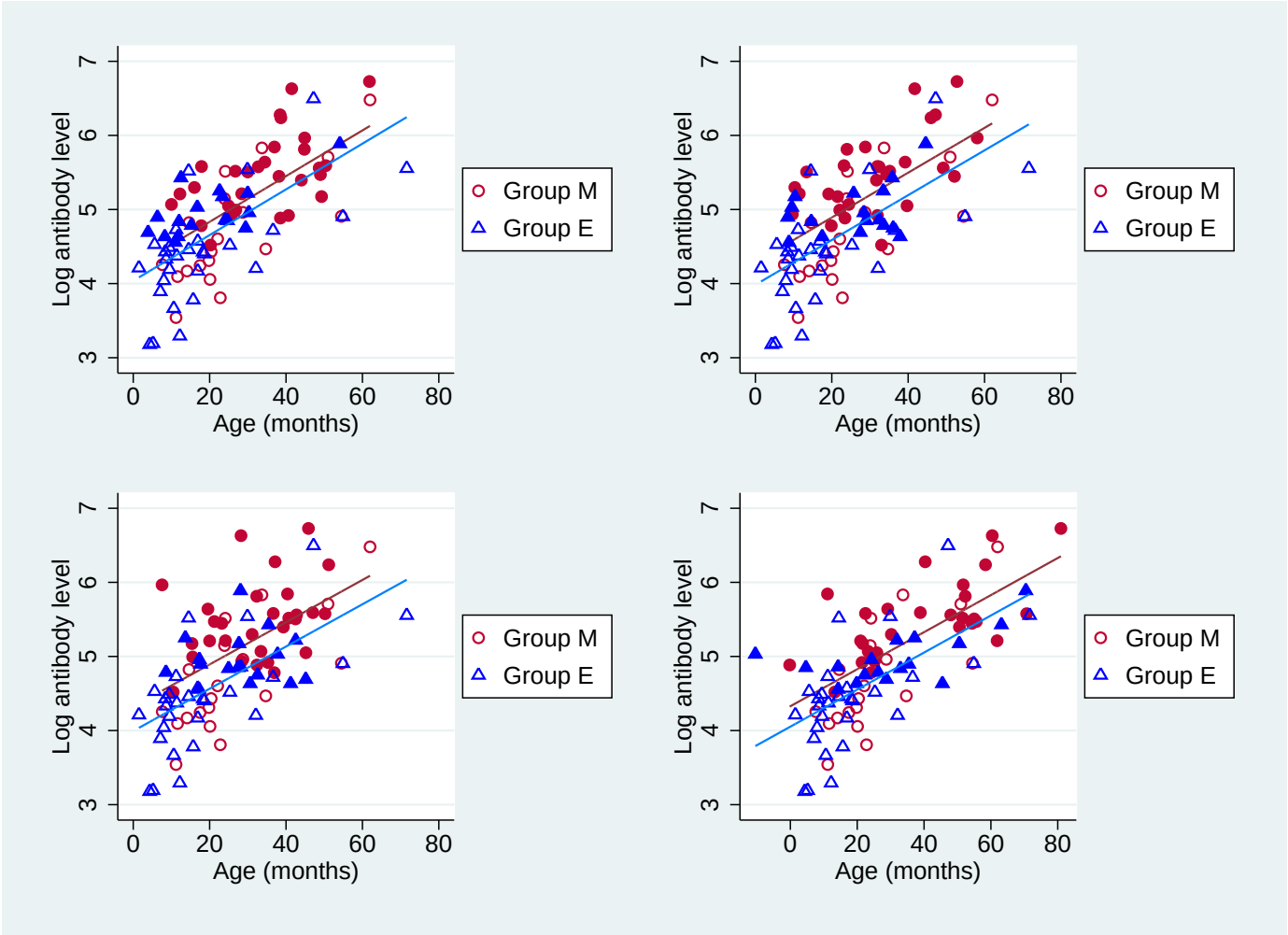
Complete cases analysis ($n = 47$)

Illustrative example: estimation results

Complete-case analysis

	full sample (n=93)		missing 50% age values (n = 47)
	Crude	ANCOVA	Complete Cases
group diff	-0.542	-0.194	-0.051
(SE)	(0.145)	(0.117)	(0.167)
age effect	-	0.0242	0.0318
(SE)	-	(0.0029)	(0.0051)

Illustrative example: analysis of covariance



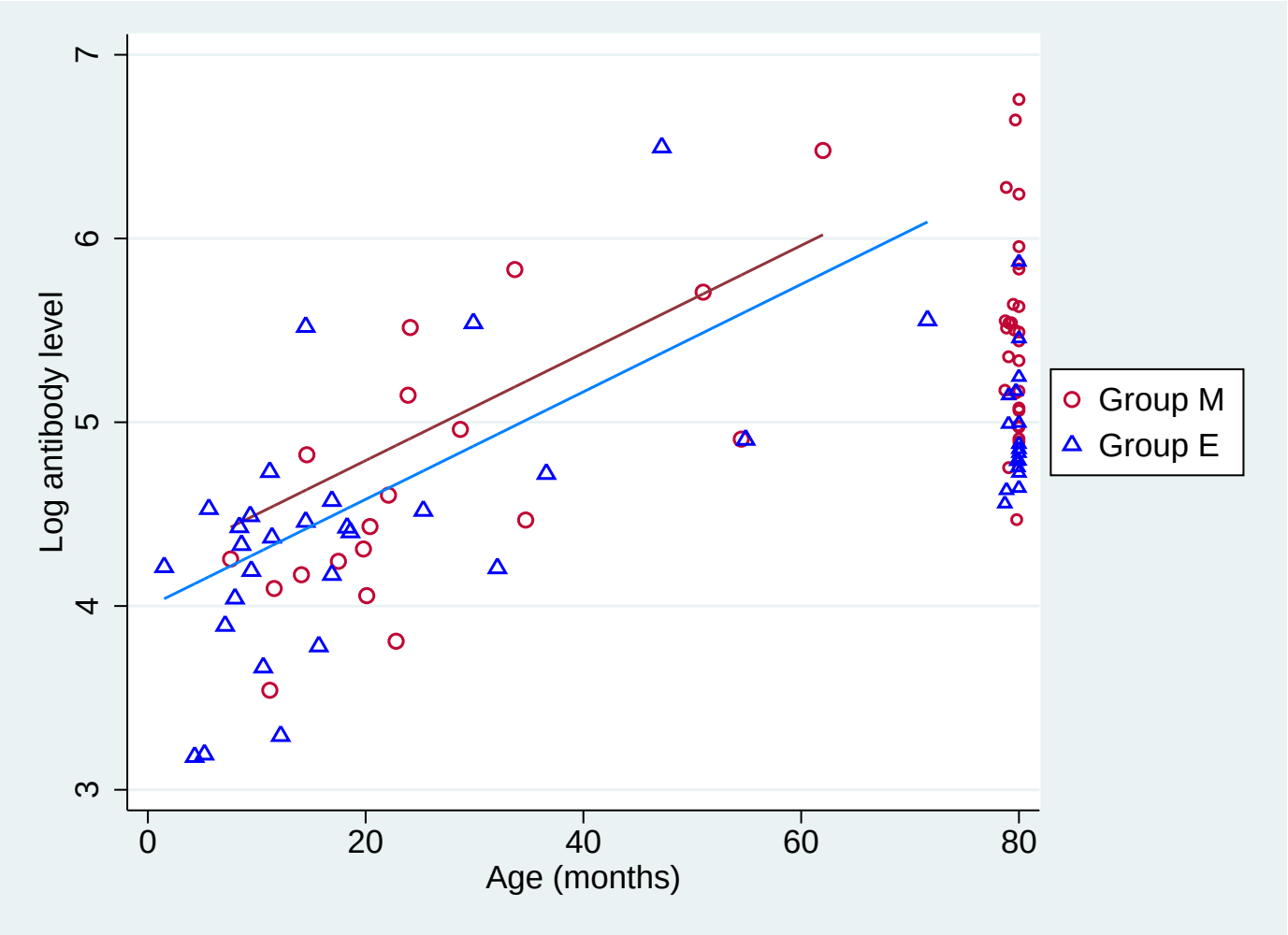
Analysis from four separate (proper) imputations

Illustrative example: estimation results

Single imputations

	full sample (n=93)		missing 50% age values (n = 47)				
	Crude	ANCOVA	Complete Cases	Imp-1	Imp-2	Imp-3	Imp-43
group diff	-0.542	-0.194	-0.051	-0.179	-0.303	-0.328	-0.275
(SE)	(0.145)	(0.117)	(0.167)	(0.121)	(0.117)	(0.123)	(0.113)
age effect	-	0.0242	0.0318	0.0308	0.0303	0.0285	0.0250
(SE)	-	(0.0029)	(0.0051)	(0.0039)	(0.0039)	(0.0042)	(0.0029)

Illustrative example: analysis of covariance



Multiple imputation estimates

Illustrative example: estimation results

Multiple imputation

	full sample (n=93)		missing 50% age values (n = 47)					
	Crude	ANCOVA	Complete Cases	Imp-1	Imp-2	Imp-3	Imp-43	MI (m=100)
group diff	-0.542	-0.194	-0.051	-0.179	-0.303	-0.328	-0.275	-0.211
(SE)	(0.145)	(0.117)	(0.167)	(0.121)	(0.117)	(0.123)	(0.113)	(0.147)
age effect	-	0.0242	0.0318	0.0308	0.0303	0.0285	0.0250	0.0293
(SE)	-	(0.0029)	(0.0051)	(0.0039)	(0.0039)	(0.0042)	(0.0029)	(0.0045)

Origins of MI

Invented by Donald Rubin in 1970s.

- First account: unpublished report for U.S. Social Security Administration (reprinted *American Statistician*, 2004)
- Rubin involved in survey sampling problems at U.S. Bureau of the Census
- According to Scheuren (*Amer Stat*, 2005), official surveys of 1940's and 1950's largely untroubled by nonresponse as "trust in government was very high"
- Latter part of 20^{th} century saw increasing levels of nonresponse both at unit and item levels
- Emphasis on "public use" datasets led survey statisticians to develop methods of filling in or imputing missing values, in particular *hot-deck* imputation was popular

Variance estimation

- As imputation by hot-deck became more popular, concerns grew about variance estimation
- Clear intuitively that applying standard inference tools to single imputed dataset produces variance (SE) estimates that are too small
  - Wrongly assumes that imputed value was actually observed
  - A problem even if the process causing missingness is completely understood

Rubin's key insight: Bayesian logic could solve this problem with MI as the tool arising . . .

Notation

- Y_{obs} represent (irregular) array of *observed* data values
- so Y_{obs} = concatenation of individual rows of observed values $\mathbf{Y_i^{o\prime}}$
- Y_{mis} = complementary array of missing values

When considered as random variables for modeling purposes, some care needed:

Y_{obs} is function of the *joint* random variables Y and R, where R is an array of response or missing data indicators.

MI as approximate Bayesian inference

Focus is on parameter β, assumed to be a single scalar for now, from a parametric model for Y, $P(Y|\beta)$. To fix ideas, think of β as a regression coefficient...

MI originates from Bayesian statistical thinking:

- Bayesian paradigm gives coherent guide for deriving methods of estimation that allow for multiple sources of uncertainty
- In this case, uncertainty due to some of desired data being missing
- Since the only data we have are Y_{obs} and R, Bayesian analysis for β involves calculating the posterior distribution: $P(\beta|Y_{obs}, R)$

MI as approximate Bayesian inference

Key representation from conditional probability:

$$P(\beta|Y_{obs},R) = \int P(\beta|Y_{obs},Y_{mis})P(Y_{mis}|Y_{obs},R)dY_{mis} \quad (3)$$

(where integral is a sum in the case of discrete Y_{mis})

- First term in integral = posterior distribution for β given a complete data set:
  Bayesian version of the standard complete-data analysis

Key insight: if generate or impute a sample of m values $Y_{mis}^{(k)}$ from the predictive distribution for the missing data, i.e. from $P(Y_{mis}|Y_{obs},R)$, then can approximate (3) by average over the complete-data posterior distributions:

$$P(\beta|Y_{obs},R) \approx \frac{1}{m}\sum_{1}^{m} P(\beta|Y_{obs},Y_{mis}^{(k)}) \quad (4)$$

MI as approximate Bayesian inference

Note: general formulas retain dependence on R in predictive distribution for Y_{mis}

- In general, predictive distribution used for imputation needs to take account not only of values Y_{obs} themselves but also of where they appear in the dataset and why those values were observed but others not
- In almost all practical applications, assumption of ignorability based on 'missing at random' (MAR) is invoked (see below)—issue is assumed away and imputation performed without modelling the process that led to the missing data

MI as approximate Bayesian inference

Generally sufficient to work with just the first two moments of the posterior distribution, the mean and variance of β, which may be obtained by the rules of iterated expectations:

$$\mathrm{E}(\beta|Y_{obs},R)=\mathrm{E}[\mathrm{E}(\beta|Y_{obs},Y_{mis})|Y_{obs},R], \tag{5}$$

and

$$\begin{aligned}\mathrm{Var}(\beta|Y_{obs},R)= \quad & \mathrm{E}[\mathrm{Var}(\beta|Y_{obs},Y_{mis})|Y_{obs},R]+\\ & \mathrm{Var}[\mathrm{E}(\beta|Y_{obs},Y_{mis})|Y_{obs},R]\end{aligned} \tag{6}$$

We estimate these from the imputed data by approximating the integrals (5) and (6) by sums over the sample of drawn values of $Y_{mis}^{(k)}\ldots$

MI as approximate Bayesian inference

Approximate posterior mean:

$$\mathrm{E}(\beta|Y_{obs},R) \approx \frac{1}{m}\sum_{1}^{m}\hat{\beta}^{(k)}, \tag{7}$$

which we denote $\hat{\beta}^{\mathrm{MI}}$ (as above) For the variance, have two

terms (second one obtained as standard unbiased estimate of the variance of $\hat{\beta}^{(k)}$ across the m completed datasets):

$$\mathrm{Var}(\beta|Y_{obs},R) \approx \frac{1}{m}\sum_{1}^{m}V^{(k)} + \frac{1}{m-1}\sum_{1}^{m}(\hat{\beta}^{(k)}-\hat{\beta}^{\mathrm{MI}})^2 \tag{8}$$

MI as approximate Bayesian inference

Approximate posterior variance:

$\bar{V} + B$, where $\bar{V} = \sum_1^m V^K/m$, and

$B = \sum_1^m (\hat{\beta}^{(k)} - \hat{\beta}^{\mathrm{MI}})^2/m$, which estimates the *between-imputation* variance of the parameter.

Gives valid approximation for large m, but for typical small m in practice, variance needs to include additional term B/m to reflect uncertainty in $\hat{\beta}^{\mathrm{MI}}$ as estimate of the true posterior mean. Thus (Rubin's rules):

$$V^{\mathrm{MI}} = \bar{V} + \left(1 + \frac{1}{m}\right) B, \tag{9}$$

MI as approximate Bayesian inference

With (approx) posterior moments, create inferences in usual way...

- If posterior distribution were normal, would use $\hat{\beta}^{\mathrm{MI}} \pm z_{(1-\alpha)}\sqrt{V^{\mathrm{MI}}}$ as $100(1-\alpha)\%$ credible interval for β
- Since variance parameters estimated with error, preferable to use t reference distribution
- Rubin et al provide formulas for appropriate degrees of freedom for t distribution

MI as approximate Bayesian inference

- Difficult to carry through formal arguments in presence of nuisance parameters.
- However, simple Bayesian view of MI remains if can define a statistic $\hat{\beta}$ that provides a valid estimate of posterior mean of β (irrespective of nuisance parameters) in (hypothetical) complete data
  - Also a key assumption for interpreting multiple imputation from a sampling theory (frequentist) point of view: desired property of $\hat{\beta}$ is consistency for estimating β in repeated samples
- If also assume a valid estimate V of the posterior variance of β is available from the complete data, then the results above flow through exactly as before

How to perform proper imputation

Bayesian development: need to generate "a sample of m values $Y_{mis}^{(k)}$ from the predictive distribution for the missing data, i.e. from the distribution $P(Y_{mis}|Y_{obs}, R)$"

- In Bayesian paradigm imputation is computational problem: how to sample from the posterior distribution of the missing data
- Software solutions have been developed...

Problem: the Bayesian paradigm is incomplete because predicated on the assumption that the proposed models are correct...

How to perform proper imputation

Implications:

- critical perspective re modeling assumptions is central (emphasis on model checking)
- should avoid using methods that can be expected to perform poorly in repeated sampling

Rubin coined the term "proper" imputation to address latter issue in MI context

How to perform proper imputation

Essential goal: inferences obtained by MI should have repeated-sampling validity

- Point estimates should be consistent for the target parameters
- Interval estimates should achieve at least the nominal coverage

Rubin showed that if the imputation method accounts appropriately for all sources of uncertainty (guaranteed if performed by full Bayesian inference under a "correct" model), then it will be proper.

Example of imputation approach that is not proper: classical hot-deck method (can be modified to be proper, in the form of the "approximate Bayesian bootstrap")

Further theoretical aspects

- df for t-distribution: small-sample modifications
- Indicators of impact of missing data on inference:
  - Relative Variance Increase: ratio of $\bar{V}$ to B
  - Fraction of Missing Information (FMI)
- Multi-parameter estimands
- How to choose m? Efficiency of finite m (justified historical emphasis on small m) vs Monte Carlo error in practice.

MAR and ignorability: packaged MI

General theory requires imputation from $P(Y_{mis}|Y_{obs}, R)$

'Missing at random' (MAR) assumption:
$P(R|Y_{obs}, Y_{mis}) = P(R|Y_{obs})$

If this holds, then it can be shown that

$$P(Y_{mis}|Y_{obs}, R) = P(Y_{mis}|Y_{obs}) \quad (10)$$

(suppressing subleties such as parametrisations!)

MAR implies *ignorability* of R in imputation model

MAR and ignorability: packaged MI

Under ignorability, imputation model is $P(Y_{mis}|Y_{obs})$

- Computation is feasible under certain joint models for $Y_{mis}|Y_{obs}$, i.e. joint model for Y
- Dominant approach has been by assuming *multivariate normal* for Y
- Requires estimation of the parameters (mean and variance-covariance) and drawing from the predictive distribution of the Y_{mis} values
- Can be done with "data augmentation" (MCMC) algorithm (Schafer, 1997)—approach labelled MVNI
- Other joint models for Y are available but these are limited and not widely used

MAR and ignorability: packaged MI

Alternative approach: MI using chained equations (MICE) or "Fully Conditional Specification"

- A joint model $P(Y)$ is not specified
- Instead:
  - a conditional (regression) model is specified for each variable (column of Y) that contains missing values
  - imputation is performed sequentially for each variable conditionally on the others
  - (N.B. each time using proper imputation, i.e. integrating to include parameter uncertainty)
- This method is popular because of its flexibility but...

MAR and ignorability: packaged MI

Practical tools available (historically):

- Schafer's "NORM" (MVNI) available for Windows (1998)
- Van Buuren's MICE in S-PLUS (2000)
- Raghunathan's IVEWARE (MICE in SAS) (2001)
- SAS Proc MI (MVNI): mid-2000s
- Royston's 'ice' for Stata (2004-)
- SPSS something? late 2000s
- Stata 11 'mi impute mvn', etc (2009)
- Stata 12 'mi impute chained', etc (2011)

Proliferation of multiple imputation in practice

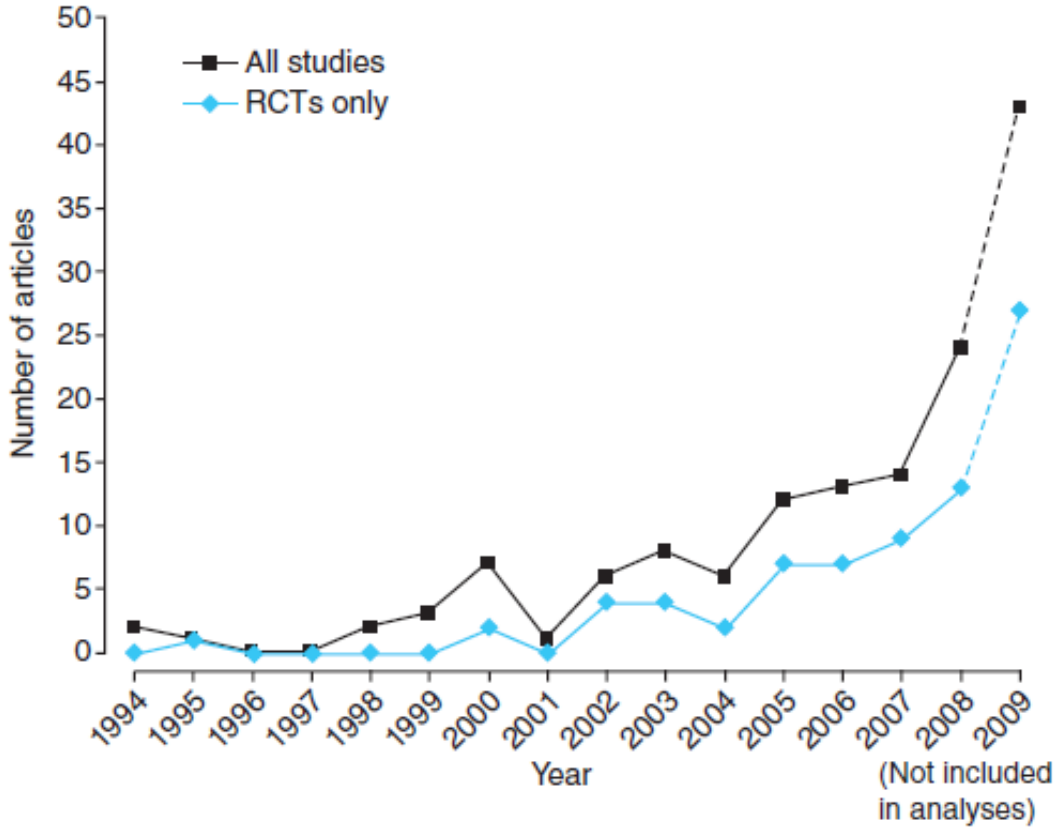


Fig. 1 *Research articles in BJM, JAMA, Lancet and NEJM using multiple imputation.*

Mackinnon A, "The use and reporting of multiple imputation in medical research – a review" *J Internal Med*, 2010

Proliferation of multiple imputation in practice

- A valuable tool but not a panacea
- Guidelines for sensible application are still incomplete, many questions remain
  - Is MI worth considering?
  - How should MI be carried out?
  - How should MI be checked once performed?

Challenge of research on MI in practice

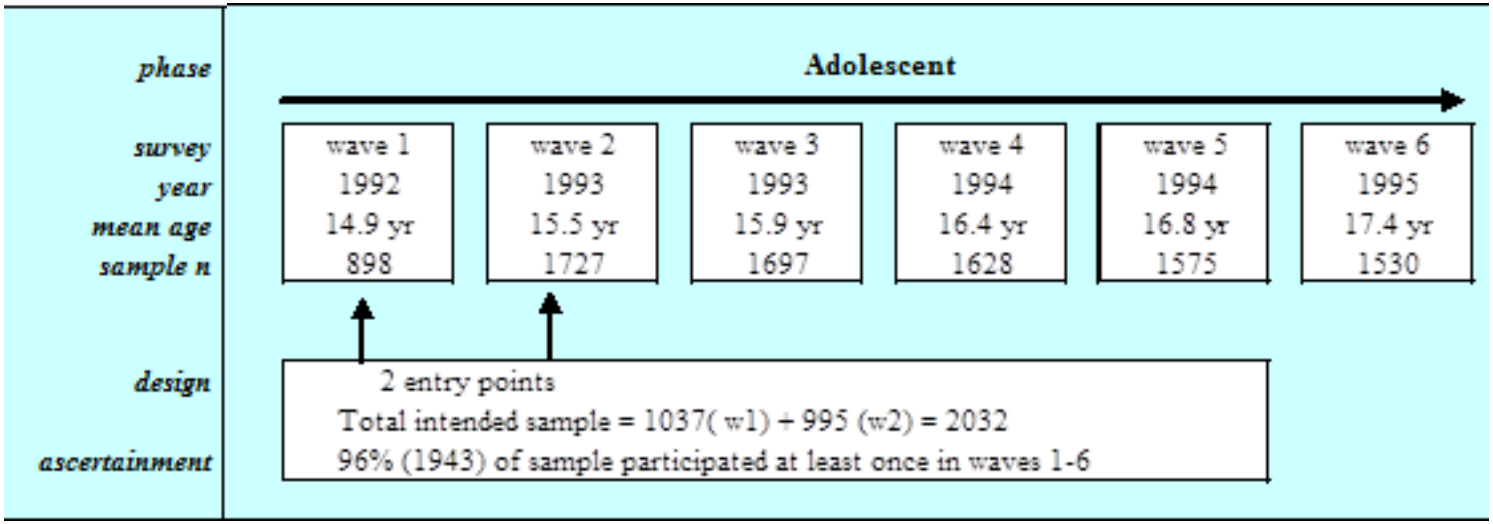
- As with most complex methods, theoretical results are limited
- May evaluate repeated sampling behaviour via *simulation*
  - Always limited to particular "simulation world" that is constructed
  - May be worth thinking about the question in more limited sampling frame: how close to desired complete-data results can we get?
- *Case studies:* examine in depth how results vary for a particular applied problem, across range of MI approaches

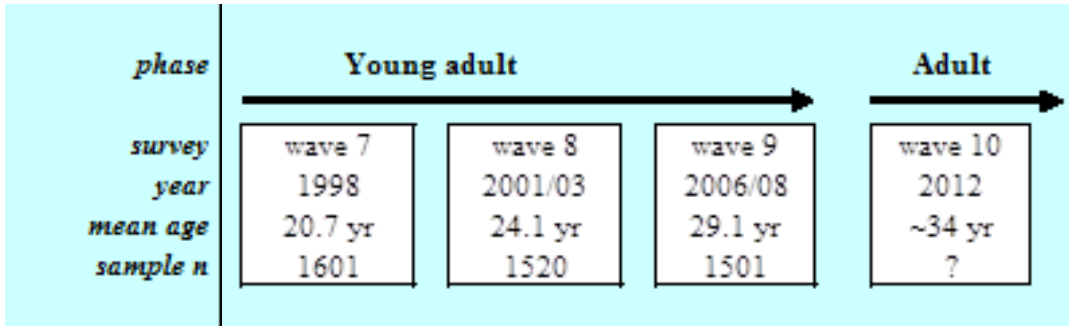
2000 Stories: the Victorian Adolescent Health Cohort Study

Longitudinal study of adolescent behaviours & mental health and interrelationship between them

- Substance use (tobacco, alcohol, cannabis, others)
- Depression/anxiety, etc
- Extended to adult phase: "continuity of risk"
- Representative school-based sample at inception
- Adolescent phase: 6 waves of frequent (6-monthly) follow-up
- Adult phase: 3 (now 4) waves at 3-4 year intervals

VAHCS: adolescent and young adult follow-up





VAHCS: missing data patterns

- Missing data generally by wave, not item
- Lots of gaps/ missing waves despite overall strong retention (*non-monotone* missing)
  - participated wave 9: $n = 1501$, 75%
  - participated wave 8: $n = 1520$, 78%
  - complete waves 7 & 8: 72%
  - complete waves 6, 7 & 8: 64%
  - *only 30% of cohort had complete data for waves 1-6...*

VAHCS: analysis themes and missing data issues

- Longitudinal dataset with large sample size & many variables
- Large numbers of analyses performed
  - . . . using many variables
  - . . . conducted by a range of analysts
- Missing data in outcomes (later waves) and in covariates (earlier waves)
- Non-monotone patterns of missingness
- Data missing for many reasons (mostly unknown!)
- Large numbers of cases lost if analysis limited to complete cases

Conclusion: perfect setting for use (and abuse?!) of MI!

VAHCS: integration of MI into analysis strategy

Early paper: Patton et al "Cannabis use and mental health in young people: a cohort study" (*BMJ*, 2002).

> *Although the response rate was high and attrition low, 70% of respondents missed at least one wave of data collection, which led to* potential bias in summary measures of exposure to cannabis and mental health problems calculated from the six waves of data collection among adolescents. *To circumvent this, we used multiple imputation . . .*

More than 10 papers since then have used MI.

Approach has evolved, from "mega-imputation" (150+ variables) to allow several analyses, to tailored imputation for specific analysis.

Many questions remain . . .

VAHCS case study: analysis of illicit drug use

(Work in progress with Helena Romaniuk)

Based on analysis for Swift et al. "Cannabis and progression to other substance use in young adults: findings from a 13-year prospective population-based study" (*JECH*, 2011)

Focus on subset of results:

- overall prevalence of cannabis (wave 7) and amphetamine use (wave 9)
- prevalence of amphetamine use stratified by concurrent level of cannabis use (wave 9)
- association between incidence of amphetamine use in adulthood and cannabis use at previous wave, controlling for potential confounders

VAHCS case study: analysis of illicit drug use

Question of interest: how sensitive are final results to the method of imputation?

Factors examined:

1. Choice between MVNI and MICE
2. Inclusion of auxiliary variables (big vs small model)
3. Inclusion of cases with large fraction of missing values
4. Truncation of extreme values (for alcohol consumption) in imputation

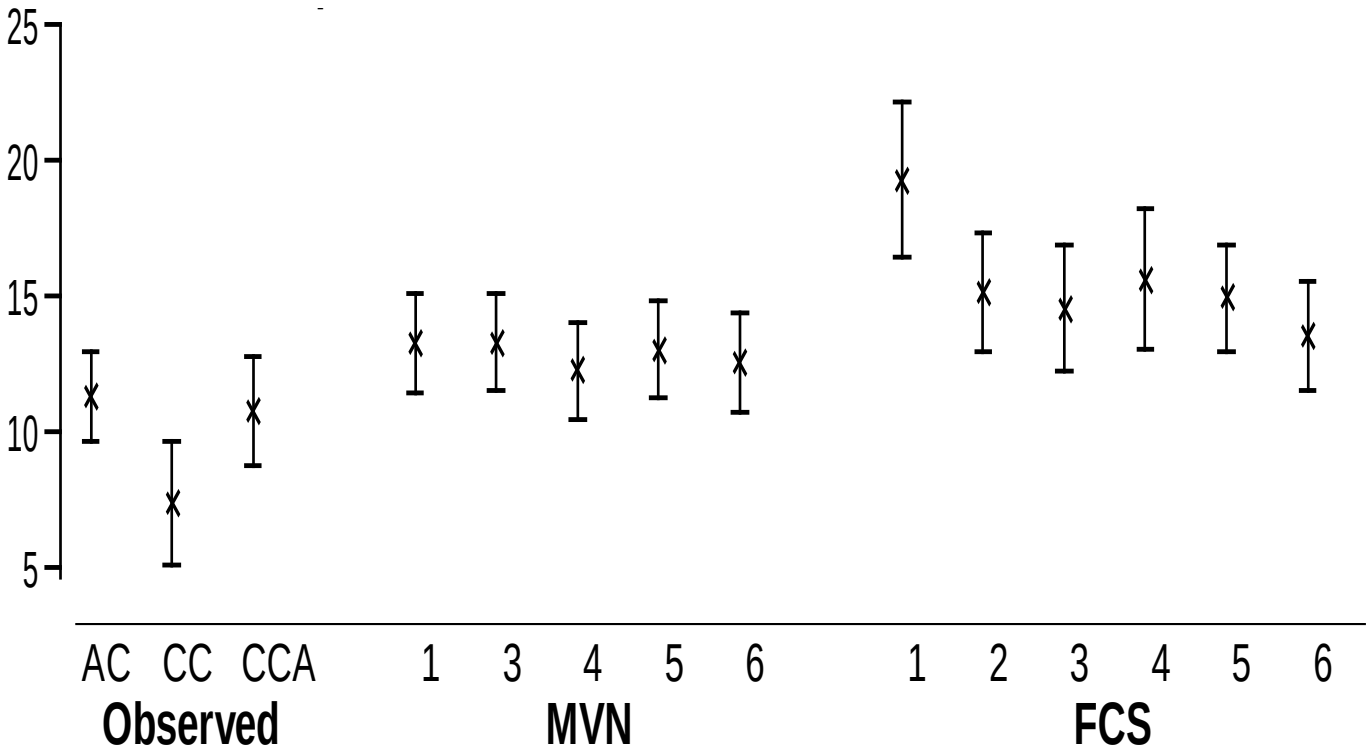
VAHCS case study: MI settings compared

Analysis Dataset	Details				Analysis Key Labels	Subjects included	
						Number	(%)
Observed	Available case – complete data for subset of vars used in analysis				AC	1384-1586	(71-82)
	Complete case – complete data for all variables used in all analyses				CC	516	(27)
	Complete case adult data with partial adolescent data				CCA	941	(49)
Imputed	**Method**	**Auxiliary vars**	**Subjects included**	**Adult alcohol vars**			
	MVN	yes	all	continuous	MVN_1	1934	(100)
	MVN	yes	all	binary	MVN_3	1934	(100)
	MVN	yes	≤50% missing	continuous	MVN_4	1679	(87)
	MVN	no	all	continuous	MVN_5	1934	(100)
	MVN	no	≤50% missing	continuous	MVN_6	1731	(90)
	ICE	yes	all	continuous	ICE_1	1934	(100)
	ICE	yes	all	cont – truncated	ICE_2	1934	(100)
	ICE	yes	all	binary	ICE_3	1934	(100)
	ICE	yes	≤50% missing	continuous	ICE_4	1679	(87)
	ICE	no	all	continuous	ICE_5	1934	(100)
	ICE	no	≤50% missing	continuous	ICE_6	1731	(90)

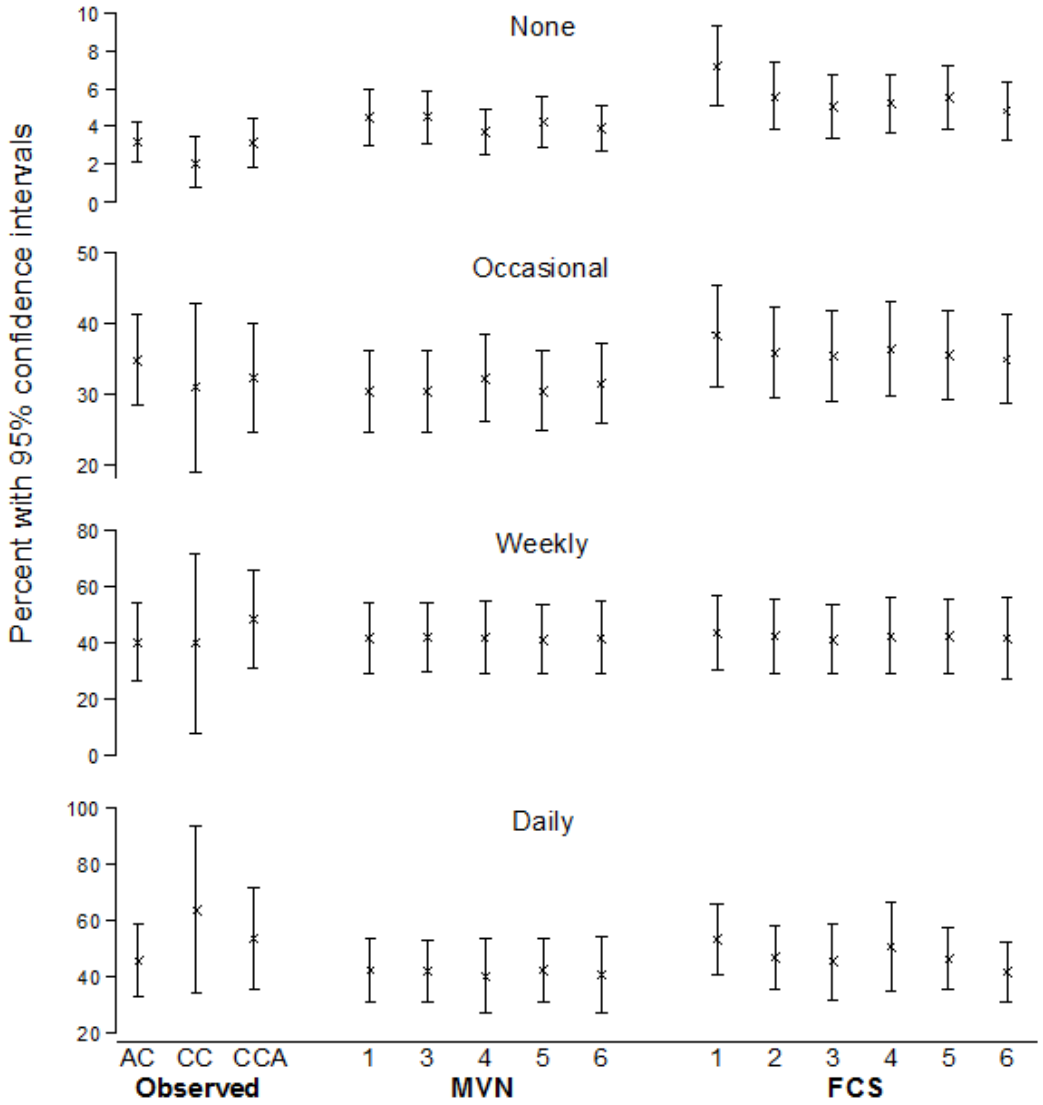
VAHCS case study: key variables of interest

Variable	Waves	Variable Type (distribution)	Categories/Range	Percent missing in datasets for imputations					
				1, 2, 3 & 5		4		6	
				Male (N=943)	Female (N=1000)	Male (N=772)	Female (N=911)	Male (N=813)	Female (N=924)
KEY VARIABLES Cannabis use in last year	2 to 9	ordinal (positively skewed)	0 non-user 1 occasional 2 weekly 3 daily	13 to 29	10 to 21	10 to 18	7 to 13	11 to 19	8 to 15
Cigarette smoking in last month	2 to 9	ordinal (positively skewed)	0 non smoker 1 occasional 2 daily smoker	9 to 28	7 to 20	9 to 17	7 to 13	8 to 18	7 to 15
Alcohol use in the last week[&]	2 to 6	binary	0 not risky drinker 1 risky drinker	16 to 34	15 to 23	15 to 24	14 to 19	15 to 25	14 to 18
	7 to 9	continuous (positively skewed)	0 to 120 units	26 to 36	16 to 26	14 to 23	10 to 19	16 to 27	11 to 21
Illicit Drug use in last year	7 & 8	binary	0 no 1 yes	22 & 26	13 & 18	9 & 12	7 & 10	12 & 15	8 & 12
Amphetamine[+] Ecstasy Cocaine	9	ordinal (positively skewed)	0 none 1 <weekly 2 weekly+	33	24	19	17	23	18
Sex		binary	0 male 1 female	0	0	0	0	0	0
Age mean centred[^]	2	continuous (symmetrical)	-3.0 to 4.6 years	0	0	0	0	0	0

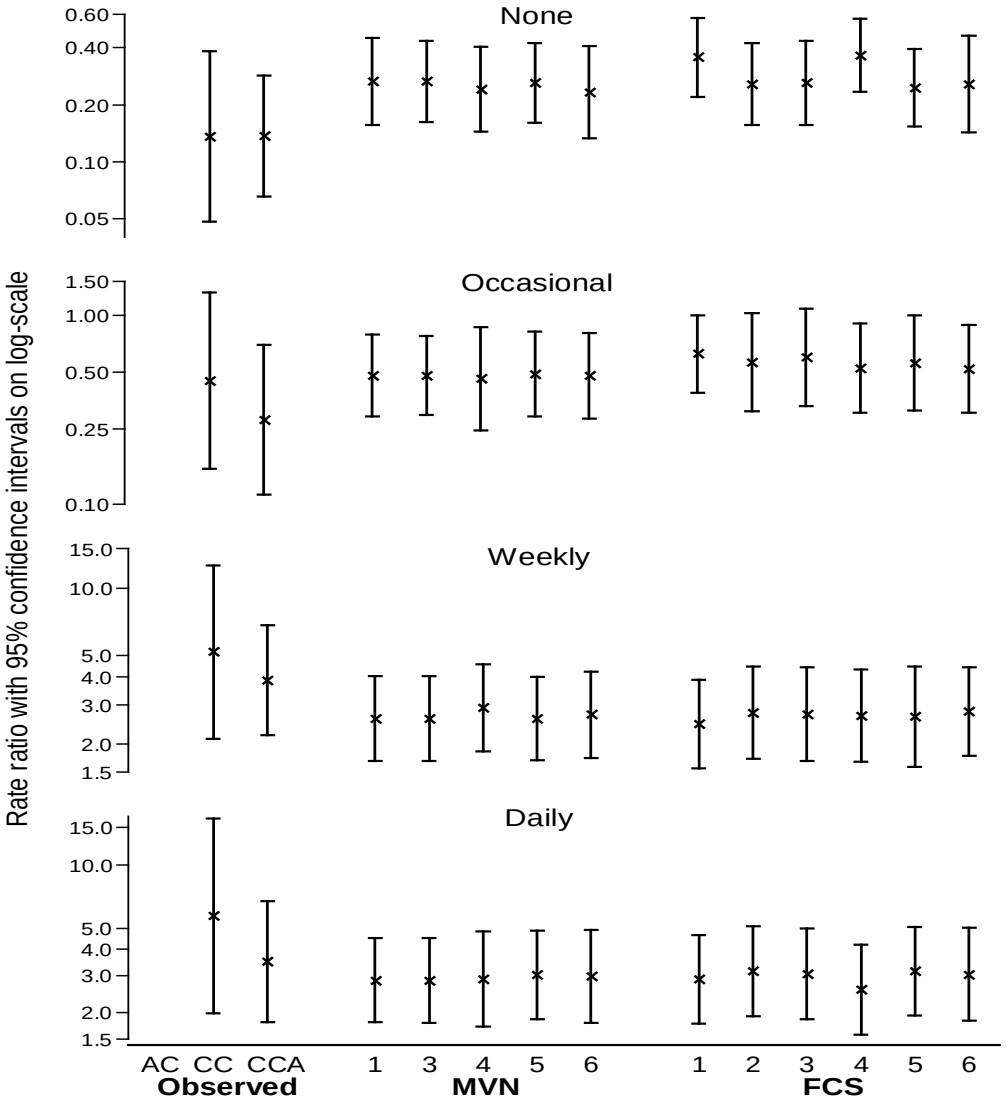
Results: amphetamine prevalence at wave 9



Results: amphetamine prevalence, by cannabis level



Results: association of amphet incidence with prior cannabis



VAHCS case study: conclusions

- Confidence intervals narrower with MI: prima facie evidence of information recovered
- Prevalence estimates vary considerably, between MVNI & MICE
  - Higher under MI (missing values generally associated with higher risk behaviours)
  - Greater variation, wider intervals with MICE cf MVNI
- Less variation for association estimates
- MVNI slightly more stable? (not necessarily less biased though)
- Auxiliary variables have little effect
- Inclusion of cases with substantial missing data makes little difference
- Extreme values may cause havoc with MICE

VAHCS case study: conclusions

For further investigation:

- Sources of potential instability in MICE
  - Collinearity, over-parametrisation?
- Clear and defensible strategies for imputing "difficult" distributions: low frequency categorical, extreme skewness, etc.
- Can similar conclusions be supported by other case studies or simulations?
- Better understanding of limits of MVNI (interactions, nonlinearities)
- Diagnostics: methods to identify important lack of fit of imputation models

Acknowledgements

- NHMRC Project Grant support
- NHMRC co-investigators: Kate Lee, Julie Simpson
- CEBU team: Helena Romaniuk, John Galati
- VAHCS collaborators: George Patton, Carolyn Coffey
- U.K. collaborators: Ian White, Patrick Royston, Shaun Seaman