

# Dealing with missing data: Prediction model for developmental outcome at 2 years of age for babies born very preterm

Karan Vadher<sup>1</sup> (karan.vadher@ox.ac.uk) and Brad Manktelow<sup>2</sup> (brad.manktelow@le.ac.uk)

<sup>1</sup>Oxford Clinical Trials Research Unit, Centre for Statistics in Medicine, NDORMS, University of Oxford; <sup>2</sup>Department of Health Sciences, University of Leicester

## Aim

Comparing how well complete case analysis, inverse probability weighting (IPW) and multiple imputation methods deal with a dataset missing 50% of outcomes comparing their multinomial logistic regression models.

## Background

Children born preterm are at increased risk of developmental problems. The Preterm and After (PANDA) study is investigating the long-term outcomes of children born very preterm (<33 gestational weeks) admitted for acute neonatal care in the east of England. PANDA uses information from The Neonatal Survey (TNS), an ongoing study of neonatal intensive care activity in the same geographic area that collects clinical information on the child, their neonatal care and the child's developmental outcome:

1. alive with no developmental delay (DD)
2. alive with DD
3. death before 2 years of age

Parents in PANDA also filled in the PARCA-R questionnaire to measure their child's cognitive and language development at 2 years old. Questionnaires were not sent to parents whose child had died. Many parents did not complete PARCA-R, giving us 50% missing outcomes.

## Methods

- We used a dataset of 2028 participants (a subset of TNS), including babies born very preterm and admitted to neonatal care in 2009-2010 and their mothers.
- The three nominal outcomes (alive with no DD, alive with DD and death before 2 years of age) were modelled using multinomial logistic regression.
- We compared complete case analysis, IPW and multiple imputation for dealing with missingness.
- The same covariates were adjusted for in the final multinomial logistic regression outcome models of all three methods.
- The optimal model predicting developmental outcome contained gestational age, baby's sex, CRIB II score (risk adjustment tool that is used to predict mortality in preterm babies), and a quadratic term for gestational age.
- We compared the three approaches using probabilities, odds ratios, log odds and standard errors.

## Results

Table 1: Summary of mother and baby characteristics

Factors/Baby Characteristics	Outcome Response			
	Alive with no DD	Alive with DD	Dead	Missing
n (%)	100 (17.5)	215 (37.5)	204 (35.4)	103 (18.0)
Gestational age (weeks)	30.2 (3.8)	27.7 (3.2)	25.7 (2.3)	28.1 (3.8)
Sex of the baby - Male - n (%)	100 (50.0)	108 (50.3)	100 (50.0)	50 (50.0)
Sex of the baby - Female - n (%)	100 (50.0)	107 (49.7)	104 (50.0)	53 (50.0)
Mother's ethnicity - n (%)				
European	107 (96.7)	108 (49.7)	100 (50.0)	10 (10.0)
South Asian	10 (9.3)	10 (4.5)	10 (5.0)	0
African/Black British	10 (9.3)	10 (4.5)	10 (5.0)	0
Mixed Race	10 (9.3)	10 (4.5)	10 (5.0)	0
Other	10 (9.3)	10 (4.5)	10 (5.0)	0
Mother's Age (years)	30.1 (3.3)	29.4 (3.3)	28.1 (3.3)	27.2 (3.3)
Missing n (%)	10 (10.0)	10 (4.5)	10 (5.0)	10 (10.0)

The IPW missingness model (Table 2) showed that deprivation area, mother's age and mother's ethnicity all affected whether the PARCA-R survey was completed. Missing completely at random was discounted.

Table 2: A logistic model with categorized deprivation score, categorized mother's age and mother's ethnicity predicting the response of the PARCA-R (final IPW model)

Covariate	OR (95% CI)	p-value
Deprivation score		
Low deprivation (0-10)	0.27 (0.14, 0.50)	<0.001
Medium deprivation (11-20)	0.33 (0.18, 0.70)	<0.001
High deprivation (21-30)	0.47 (0.21, 1.00)	<0.001
Mother's age		
15-20 years	2.17 (1.36, 3.37)	<0.001
21-25 years	1.86 (1.16, 2.97)	<0.001
26-30 years	0.87 (0.51, 1.47)	0.001
31-35 years	0.40 (0.23, 0.69)	<0.001
36-40 years	0.28 (0.16, 0.48)	<0.001
Missing	0.17 (0.01, 3.75)	0.18
Mother's ethnicity		
South Asian	0.33 (0.14, 0.80)	0.01
African/Black British	0.33 (0.14, 0.80)	0.01
Mixed Race	0.33 (0.14, 0.80)	0.01
Other	0.17 (0.01, 3.75)	0.18

The imputation model also produced strong evidence of covariates predicting non-responders. Once the investigation of missingness had been conducted the same multinomial logistic regression was produced.

Unsurprisingly, complete case analysis yielded very different results to models that used IPW and multiple imputation. Odds ratios and probs of each outcome were broadly similar with multiple imputation yielding smaller standard errors of the odds ratios in the multinomial logistic regression.

Table 3: Covariate coefficient estimates from all three methods comparing alive with DD and alive with no DD

Covariate	Complete case analysis		IPW		Multiple imputation	
	Estimate (95% CI)	Standard Error	Estimate (95% CI)	Standard Error	Estimate (95% CI)	Standard Error
Gestational age	1.32	1.03	1.32	1.03	1.32	1.03
Sex of baby	0.52	0.52	0.52	0.52	0.52	0.52
CRIB II score	0.0001 (0.0001, 0.0002)	0.0001	0.0001 (0.0001, 0.0002)	0.0001	0.0001 (0.0001, 0.0002)	0.0001
Constant	11.48	10.42	11.48	10.42	11.48	10.42

Fig 1B and C show the models produced by IPW and multiple imputation respectively, have very similar results. In comparison, the discounted complete case analysis in Fig 1A shows an overestimation of the probability of dying.

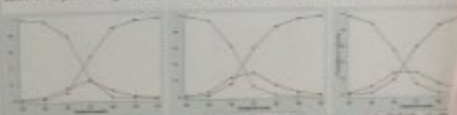


Figure 1: Predicted probability of being alive with no DD, alive with DD or dying before the age of 2 years for babies with a CRIB II score of 6, over gestational age, with missing data dealt with by (A) complete case analysis, (B) IPW, or (C) multiple imputation.

Table 4: Limitations to IPW and multiple imputation

IPW limitations	Multiple imputation limitations
The weights of the observations increase as the proportion of missing data increases, which can lead to unstable estimates.	Multiple imputation models are based on unobserved data, which can lead to biased estimates if the model is misspecified.
IPW requires the assumption that missingness is independent of the outcome and covariates.	Multiple imputation requires the assumption that missingness is independent of the outcome and covariates.

## Conclusion

- Complete case analysis was deemed inappropriate as missing data was not missing completely at random.
- IPW and multiple imputation produced very similar results and both effectively used the data available to "predict" the missing outcome.
- As IPW is conceptually simpler than multiple imputation, it is worth considering if the missing outcome is binary (as in this project).
- As multiple imputation is more flexible than IPW, it is worth considering if the dataset contains a number of observations with missing covariates.



The Neonatal Survey

