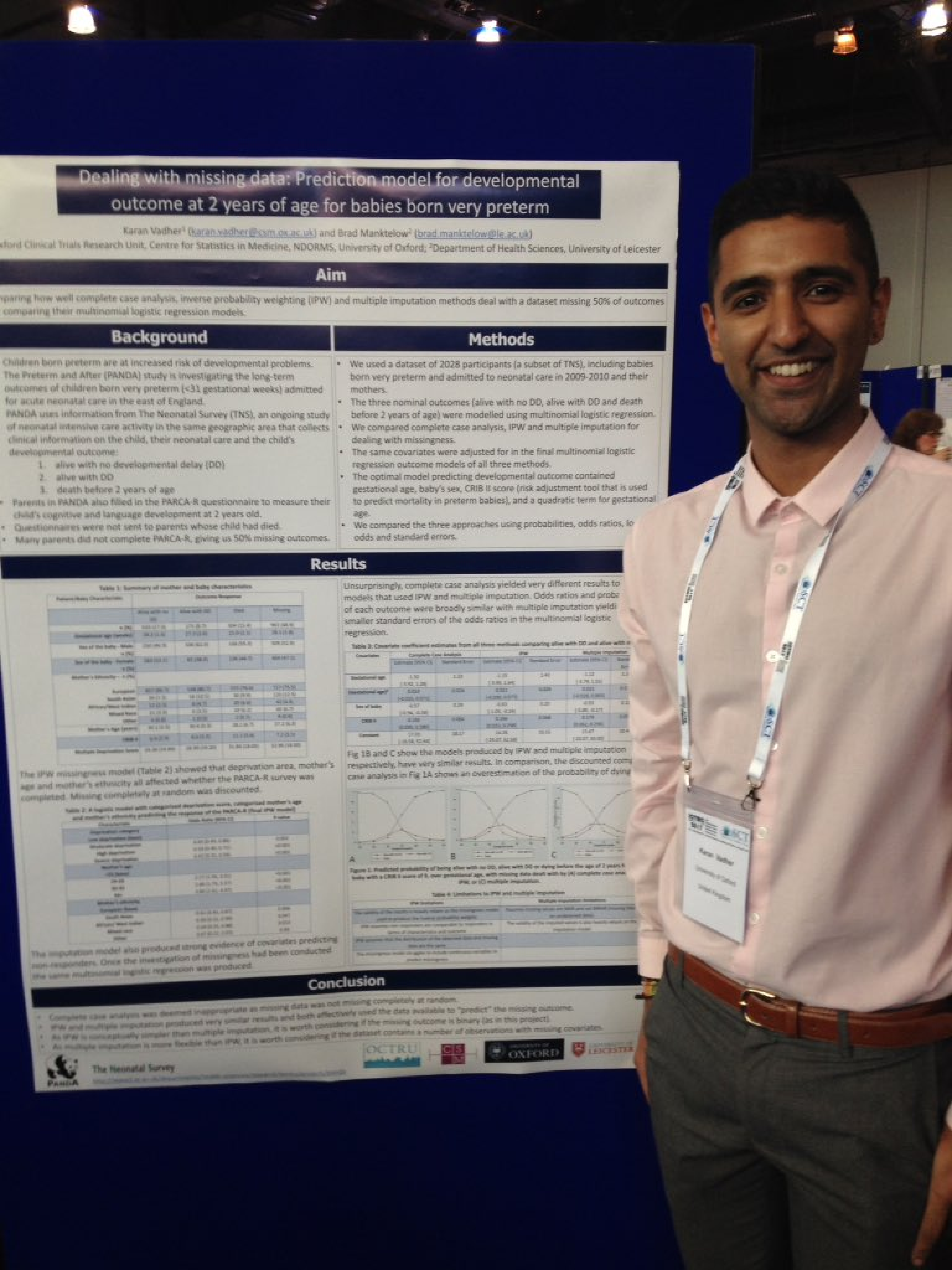
Dealing with missing data: Prediction model for developmental outcome at 2 years of age for babies born very preterm
Karan Vadher (karan.vadher@csm.ox.ac.uk) and Brad Manktelow (brad.manktelow@le.ac.uk)
xford Clinical Trials Research Unit, Centre for Statistics in Medicine, NDORMS, University of Oxford; Department of Health Sciences, University of Leicester
Aim
mparing 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 (<31 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, lo odds and standard errors.
Results
Table 1: Summary of mother and baby characteristics
Unsurprisingly, complete case analysis yielded very different results to models that used IPW and multiple imputation. Odds ratios and proba of each outcome were broadly similar with multiple imputation yieldi smaller standard errors of the odds ratios in the multinomial logistic regression.
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.
Fig 1B and C show the models produced by IPW and multiple imputation respectively, have very similar results. In comparison, the discounted comp case analysis in Fig 1A shows an overestimation of the probability of dying
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.
Table 4: Limitations to IPW and multiple imputation
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.
OCTRU
OXFORD
LEICESTER
PANDA
The Neonatal Survey