# Linking Antenatal Factors to Non-Communicable Diseases in Children and Adults

*A Fight Against Missing Data*

**Anthony Chapman**
University of Aberdeen

**Abstract**

There is a large body of evidence linking reduced birth weight and increased risk for non-communicable diseases (NCD) such as type II diabetes and asthma, which implicates factors driving fetal growth in NCD aetiology. We are exploring the potential for a computing approaches to relating repeated measurements of fetal size during a pregnancy to post natal outcomes using routinely acquired data for the population of Grampian.

## Introduction

Our group has related fetal measurements to postnatal outcomes in childhood which include asthma and eczema in a local population [1, 2, 3] and also from Saudi Arabia [4].

We are exploring the possibility for a computational approach to relate repeated measurements of fetal size during a pregnancy to post natal outcomes using routinely acquired data for the population of Grampian. If successful, our approaches have the potential to be used nationwide or even internationally. One of the biggest problems we face are missing values in the routinely acquired data. For our project to be valid, missingness is one of the first problems we will have to remedy.

## Current Objective

Our current efforts are towards the creation of artificially complete data, that is to generate a dataset which is complete (i.e. no missing values) out of a dataset with only partial completeness. As you can see in Table 1, missingness ranges from 0% to 65% in each individual column. By analysing the dataset wee noticed that 100% of the records have at least 1 value present. On the other hand, we also noticed that only 15% of the records are complete in all fields. In order for any analytical result to reliable, we will need to use more than 15% of of dataset, otherwise have can we justify any results on such a small sample of any population. Thus we would like to find a way to complete the missing data and then evaluate how good (i.e. close to the truth) our new data is.

**Table 1:** Summary of the sample data, N is out of 2000

| Statistic | N | Mean | St. Dev. | Min | Max |
|---|---|---|---|---|---|
| matsm | 1,999 | 0.297 | 0.457 | 0 | 1 |
| Quintile_SIMD_2006 | 1,986 | 3.602 | 1.454 | 1 | 5 |
| Z_CRL_our | 852 | −0.000 | 0.999 | −4.930 | 4.799 |
| z_BPD_our | 1,480 | 0.000 | 1.000 | −4.040 | 3.737 |
| Z_BWT_ICH | 1,716 | 0.208 | 0.989 | −2.300 | 11.020 |
| Z_BMI_5years | 754 | 0.518 | 0.940 | −2.550 | 4.320 |

## Methods

### Benchmark databases

In order to evaluate any imputation method, we will need to have a benchmark to compare any results. For this benchmark to be relevant to our project, it needs to behave similarly to our own database. Thus we are going to use a subset of our database as our benchmark and then imitate the original database with an artificial one. Illustrated in Figure 1, we be able to replicate the distribution of missingness from a benchmark database which itself was obtained from the original database. Thus we can test what impact imputation has on any database and have something to compare it to.

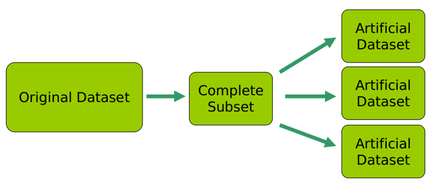


**Figure 1:** Creating a benchmark dataset

### Imputation

Imputation is the process of replacing missing data with some values[5]. We have chosen MICE [6], an R [7] package, for the creating of artificially complete data. MICE works by approximating a missing value by looking at not only the known fields in the same record but also by analysing the distribution of all the other records in a dataset.

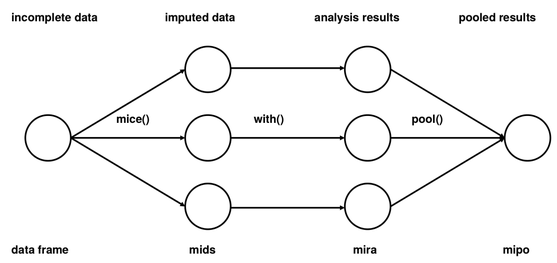


**Figure 2:** Main steps in MICE

### Evaluating Imputation

In order to evaluate whether MICE creates a good artificially complete dataset, we will use the already mentioned benchmarks together with a regression model for comparison. We start with an incomplete dataset, we extract a complete subset (our benchmark) from it, we then create several artificially incomplete datasets to mimic the original. Once we have these artificially incomplete datasets, we are able to apply MICE to them and compare the results to the benchmark database by creating the same model on all the datasets and decide whether imputation worked or not.

## Results

In Table 2, column (1) represents the complete dataset and columns (2) and (3) represent two artificially complete datasets, each dataset has 400 records. The values in the columns are the regression coefficients for each values in the first column. Essentially we are regressing Z_BWT_ICH in terms of the other values. In the table one can see that that the values are very close to each other.

**Table 2:** (1)- regression model of complete data, (2) - regression model of first artificially complete data, (3) - regression model of second artificially complete data

| | *Dependent variable:* | | |
|---|---|---|---|
| | Z_BWT_ICH | | |
| | (1) | (2) | (3) |
| z_BPD_our | 0.148*** | 0.173*** | 0.201*** |
| | (0.054) | (0.055) | (0.054) |
| Quintile_SIMD_2006 | 0.056 | 0.050 | 0.059* |
| | (0.036) | (0.037) | (0.034) |
| matasevYes | −0.150 | −0.102 | −0.121 |
| | (0.126) | (0.127) | (0.121) |
| matsm | −0.182 | −0.218* | −0.223* |
| | (0.118) | (0.122) | (0.114) |
| bsexmf24Male | −0.247*** | −0.320*** | −0.329*** |
| | (0.093) | (0.094) | (0.090) |
| Constant | 0.217 | 0.266 | 0.223 |
| | (0.166) | (0.169) | (0.160) |
| Observations | 334 | 334 | 334 |
| $R^2$ | 0.091 | 0.090 | 0.113 |
| Adjusted $R^2$ | 0.075 | 0.073 | 0.097 |
| Residual Std. Error (df = 327) | 0.828 | 0.833 | 0.799 |
| F Statistic (df = 6; 327) | 5.482*** | 5.399*** | 6.942*** |
| *Note:* | | | *p<0.1; **p<0.05; ***p<0.01 |

## Conclusions

We mimicked the missingness distribution of a dataset onto a benchmark dataset so we could compare the effect of applying MICE to it. Our preliminary results, based on a sample dataset, suggest that imputation by MICE creates an efficient artificially complete dataset.

## Discussion

These methods worked on this dataset, but it does not mean it will work on any dataset. The framework proposed here will verify whether it is possible to impute any incomplete dataset.

## Forthcoming Research

Our next steps will be to apply these imputation methods to the real data (approvals pending acceptance), then we can analyse the dataset without any missing values. We will then be use clustering techniques to find relationships within the data.y

## References

[1] S. Turner, "First- and second-trimester fetal size and asthma outcomes at age 10 years," 2011.

[2] S. Turner, N. Prabhu, P. Danielian, and G. McNeill, "Perinatal programming of childhood asthma: Early fetal size, growth trajectory during infancy, and childhood asthma outcomes," 2011.

[3] S. Turner and G. Devereux, "Fetal ultrasound: Shedding light or casting shadows on the fetal origins of airway disease," 2012.

[4] A. AlMakoshi, A. Ellahi, B. Sallout, G. Devereux, and S. Turner, "Fetal growth trajectory and risk for eczema in a saudi population," 2015.

[5] A. Gelman and J. Hill, "Data analysis using regression and multilevel/hierarchical models," 2006.

[6] S. van Buuren, K. Groothuis, and A. Robitzsch, "Multivariate imputation by chain equations," 2015.

[7] R. D. C. Team, "The r project for statistical computing." `http://www.r-project.org/`, 1993. Accessed: 10 May 2015.

## Acknowledgements

I would like to thank my supervisors and the FARR Institue for this great chance to show some of my work.

**Contact Information:**

Department of Applied Health Sciences
Computing Science Department
University of Aberdeen
238 Meston Building, Old Aberdeen

Phone: 07708288138
Email: `r01ac14@abdn.ac.uk`