9-Month Assessment Report

Anthony Sergio Chapman

SUPERVISORS:
Dr Steve Turner & Dr Wei Pang

Doctor of Philosophy
at the
University of Aberdeen.

Department of Applied Health Sciences

2015

Declaration

I declare that this document and the accompanying code has been composed by myself, and describes my own work, unless otherwise acknowledged in the text. It has not been accepted in any previous application for a degree. All verbatim extracts have been distinguished by quotation marks, and all sources of information have been specifically acknowledged.

Signed:

Date: 2015

Abstract

What is this shizz about?!?!

Acknowledgements

Thank god for tea!

Contents

Chapter 1: Introduction

This chapter introduces the 9 month report, presents the research questions and gives a quick overview of some of the background and motivations for this PhD.

1.1 Early Assessment

I would like to mention that I started this PhD in mid-December but wish to take part in the 9-month assessment at the same time as everyone else.

One of the main reasons for this is that I wish to go to the Summer Symposium and present my work to and with everyone else.

I believe that I have worked hard enough to justify an early assessment. My progress up to date has been efficient and I have identified key issues, problems and possible solutions to my research questions very early on. I will talk about these in more details throughout this report. I will also mention our communications with other universities whom are interested in similar research and the possibility of collaborative work.

1.2 Background

Dr Steve Turner, from the Applied Health department, is the person motivating this project. Dr Turner's background with this field is very broad and he wishes to introduce computational approaches to classical health questions. Together with Dr Lorna Aucott, they inspired the project and created interest with the FARR institute, whom fund the project and focus on health informatics research.

Dr Turner has already worked on projects which focus on the relationship between antenatal measurement and postnatal outcomes and has shown that certain types of growth inside the womb lead to an increased chance of the baby having asthma when it grows up [12].

Research within this field is being carried out throughout the world. Generation R projects have included asthma origins [10] as well as their symptoms in early childhood [11]. Other researchers from Italy and Russia are also looking at fetal growth trajectories [2, 1, 13, 14].

Given such research, Dr Turner would like to find any type of link between antenatal factors and postnatal diseases or disorders like ADHD, diabetes, epilepsy and adult asthma.

1.3 Research Questions

The main research question that needs to be answered is this:

What is the relationship between fetal and maternal characteristics to non-communicable diseases in children and adults?

Sub questions:

Are IVF babies small or do IVF mums produce small babies? IVF vs Spontaneous from same mother

If they are born small, do they catch up? IVF +StonesâĂŃ

If they are born small, at what point do they become small? All datasets

How accurate is gestational assessment?

Chapter 2: Key Issues

Early in the PhD, I have collected a number of sample datasets. These datasets, I have been told, are good indicators on what to expect when receiving the actual data. By experimenting with these datasets, key issues and ideas have been encountered.

2.1 Missingness

The first thing I noticed when looking at the sample datasets was the sheer amount of missing data. If the sample data I have truly represents the data I shall receive, missingness is a problem that has to be resolved.

One of the biggest problems missingness induces, is that of reliability or confidence in any analysis results. For example, 30% of population would be enough to confidently state anything about the population as a result of analysis the 30%, but what if only 15% of that 30% is complete? That leaves us with only 4.5% of the population, which would not be enough to justify any statement about the population.

We can not, however, disregard the data with partial missingness. Information, important or not, can still be gathered from missing or partially missing data.

Imputation is the process of replacing missing data with some values. There exist a full range of imputation techniques from simple default value substitution (ie replacing all missing values with some values), slightly more clever ways such as mean values substitution (similar to default value substitution except here the values may change according to the dataset), to very complicated imputation which works by calculating probabilities of values according to the know ones.

2.2 Clustering and Cluster Validation

Dr Wei Pang and I have been discussing ways for data analysis using clustering techniques. Clustering is a way of separating the data into sections, called clusters, these clusters will have the data points which are closest to each other. It works by separating points which are not similar to each other and thus telling us the characteristics of a dataset.

Our general idea is that similar antenatal behaviours will lead to certain outcomes. Thus by clustering the dataset, we will hope to find that some clusters have certain tenancies and other have different ones. What will we actually find? I am trying to not look for any type of results, I have a believe that statistical analysis is slightly biased by the fact that they are specifically looking for certain outcomes.

By just analysing the data without looking into the relationships between trends and outcomes I hope to find interesting results, moreover, by not specifically looking for such results, I believe the results will be more reliable.

Cluster validation is used for evaluating cluster outcomes. This is useful in order to assess the validity of a clustering, it can be used to compare clustering algorithms or even different datasets against each other.

If we are going to use clustering to discover information from the data, cluster validation will be used to test the efficiency as well as the correctness of the outcomes we discover. It will also be useful when handling missing data, we will be able to used cluster validation to check the effects on running any imputation technique to datasets.

2.3 Growth Trajectories

From the sample data, we can see that the measurements are consist of some volume/size measurements (trimesters and weight at 5 years) and some categorical measurements (maternal data, smoking, previous asthma economics). In order to analyse the growth characteristics and determine any relationships to diseases and disorder, a growth trajectory needs to be defined.

What we have are growth measurements, which alone are not enough to describe a growth trajectories which might represent the whole data. What we need is some sort of formula which takes into account the growth measurements and produces a growth curve or formula.

Using only the growth measurements would not be enough. As already mentioned, we also have growth characteristics such as whether the mother smoked. These categorical datum have to be taken into account also.

Mixed modelling is a way of statistically modelling data with mixed (numerical and categorical) data. It will be able to take into account categorical data and use it to change any growth trajectory to make it more realistic.

Chapter 3: Literature Review

This chapter covers some of the current work which inspire my current ideas, some topics related to the research questions and some of the methods I believe will help solve the questions.

3.1 Imputation

Imputation is the process of replacing missing fields with values. There is a huge array of imputation techniques ranging from straight forward "default value imputation"[8], to "mean value imputation" [5] or even imputation by equations [6, 15].

Default value imputation techniques are not appropriate as I believe they are too biased. By choosing to replace all missing fields with one value, the data is shifted into a direction which might (with high probability) jeopardise any underlying relationships within the dataset. Similarly, mean value imputation does not consider enough of the dataset to produce reliable imputations. It only looks at one fields at a time and does not consider the relationship between different fields in each record. This could also negatively affect the results of any analysis carried out.

Multiple Imputation by Chained Equations (MICE) [15] considers both the relationship between the fields of each record and the behaviour of all the other records in the dataset. I have chosen to use MICE to impute my data, given that all data behaves differently, a method for evaluating the efficiency of MICE on my datasets will have to be created.

3.2 Clustering

R has some very good clustering packages available for the public to use. The most widely used ones are "cluster" [9], "cclust" [3], "fclust" [7] and "mclust" [4]. They can all perform similar clustering techniques, they differ in terms of efficiency and the type of data they can cluster efficiently.

I will begin with mclust, it is comfortable performing model-based clustering using mixture models. This is useful when the datasets follow a multi model tier structure. Although powerful, our data is not complicated enough to justify using this package, other packages would perform the same clustering without overcomplicating the process. This leads to a more efficient (in terms of speed in this case) process.

Similarly, fclust works well with fuzzy data were the clustering can be a bit ambiguous. It can cluster data with levels of certainty where other clustering techniques have a binary approach, it is either in a cluster or now. Again, our data is not so complex to need such techniques.

The main choice lies between cluster and cclust, cluster is the original package and has more online support, whereas cclust has a better indexing system which is better for finding the optimal number of clusters a dataset needs.

Tests will have to be carried out to see which one should be use. My prediction would be to use cclust for finding the optimal number of clusters and them using this number to perform a clustering using the package cluster. It will need to be tested but by using both their strengths, I will have the best clustering available

3.3 Growth vs Asthma

It has been statistically proven than reduced fetal size from the first trimester is associated with increased risk for asthma and obstructed lung function in childhood [12]. It was proven using a longitudinal study / statistical analysis on around 1k subjects.

The methods used are simple statistical analytics, with confidence intervals to indicate how valid the tests are.

The problem with statistical models is that regardless of the confidence level, they are wrong. We just need to find one that is the least wrong.

3.4 European Thesis

3.5 Regression Modelling

Chapter 4: Transferable Skills

4.1 Presentation Skills

4.2 Approvals and training

Chapter 5: Progress

5.1 Italian Partners

5.2 ACERO Symposium

5.3 FARR International

5.4 FARR PhD

Chapter 6: Future Plans

Chapter 7: Conclusion

Bibliography

[1] Lucia Vaira Antonio Malvasi Andrea Tinelli, Mario Alessandro Bochicchio. Ultrasonographic fetal growth charts: An informatic approach by quantitative analysis of the impact of ethnicity on diagnoses based on a preliminary report on salentinian population, 2014.

[2] Longo Antonella Malvasi Antonio Tinelli Andrea Bochicchio Mario, Vaira Lucia. Fpgt: An online system for customized fetal and pediatric growth tracking, 2014.

[3] Evgenia Dimitriadou and Kurt Hornik. Convex clustering methods and clustering indexes, 2015.

[4] Chris Fraley, Adrian E. Raftery, and Luca Scrucca. Normal mixture modelling for model-based clustering, 2015.

[5] Andrew Gelman and Jennifer Hill. Data analysis using regression and multilevel/hierarchical models.

[6] Andrew Gelman, Jennifer Hill, and Yu-Sung Su. Missing data imputation and model checking, 2015.

[7] Paolo Giordani and Maria Ferraro. Fuzzy clustering, 2015.

[8] Trevor Hastie, Robert Tibshirani, Balasubramanian Narasinhan, and Gilbert Chu. impute: Imputation for microarray data, 2015.

[9] Martin Maecchler and Peter Rousseeruw. Cluster analysis extended, 2015.

[10] PhD thesis Dr. Agnes Sonnenschein-van der Voort. Fetal and infant origins of childhood asthma, 2014.

[11] PhD thesis Dr. Esther Hafkamp-de Groen. Asthma symptoms in early childhood: a public health perspective, 2014.

[12] Steve Turner. First- and second-trimester fetal size and asthma outcomes at age 10 years, 2011.

[13] Luccia Vaira. Quantitative fetal growth curves comparison: a collaborative approach, 2014.

[14] Luccia Vaira and Mario Bochicchio. Are static fetal growth charts still suitable for diagnostic purposes?, 2014.

[15] Stef van Buuren, Karin Groothuis-Oudshoorn, and ALexander Robitzsch. Multivariate imputation by chain equations, 2015.