9-Month Assessment Report

Anthony Sergio Chapman

SUPERVISORS:

Dr Steve Turner

Dr Lorna Aucott
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

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, and this 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.

Routinely acquired data will be linked and approvals sought. The current approach is as follows: using multiple imputation, we will address missing data. After the data are ready for processing, clustering will be used to group the data into subsets with similar traits. Each subset will be statistically analysed and outcomes will be derived by their antenatal measurements. The outcomes of interest are NCD which can be determined in children and include obesity, asthma, eczema, epilepsy and type I diabetes.

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.

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

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[23, 25, 24, 21, 22], this implicates factors driving fetal growth in NCD aetiology. There is an emerging literature describing associations between fetal size and growth and NCD. The goal of this project is to find efficient computational approaches to solve such problems in a greater scale.

We are exploring the possibility for a computational approach to relating repeated measurements of fetal size during a pregnancy to post natal outcomes using routinely acquired data for the population of Grampian. If successful, our approach has the potential to be used nation wide or even internationally.

Research within this field is being carried out throughout the world. Generation R projects have included asthma origins [21] as well as their symptoms in early childhood [22]. Other researchers from Italy and Russia are also looking at fetal growth trajectories [4, 1, 26, 27].

1.3 Research Questions

Our current research can be split into two sections which are as follows:

1.3.1 Methodological

1.3.2 Epidemiological

The main question we would like to answer is: What is the relationship between fetal and maternal characteristics to non-communicable diseases in children and adults? An early hypothesis is that change of growth will be associated with increased risk for postnatal NCD compared to steady growth . Issues we will need to address include missingness of data, comparison between different anthropometric measurements for the same individual and the anticipation that different growth trajectories will be associate with the same increased risk for NCS.

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?

Methodological : Missing and gestetional ages

Epidemiological: Size vs NCS

Chapter 2: Key Issues

Our group has acquired a number of sample datasets, these datasets represent the actual data to be analysed. By analysing and experimenting with these datasets we have been able to identify key issues and ideas which are critical for the success of this project.

2.1 Missingness

At first glance, we noticed that the sample datasets have vast amounts of data missing. If the sample data we have truly represents the real data, missingness is an issue that will have 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[8]. There exist a full range of imputation techniques from simple default value substitution (ie replacing all missing values with some values)[11], slightly more clever ways such as mean values substitution [7](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[28, 9].

2.2 Clustering and Cluster Validation

Our group has been discussing ways for the analysis of data 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[2]. 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. We are not trying to look for any specific result, we have an idea 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, we hope to find interesting results, moreover, by not specifically looking for such results, the results will be more reliable.

Cluster validation is used for evaluating cluster outcomes[12]. 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 techniques 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), as shown in Figure 2.1. In order to analyse the growth characteristics and determine any relationships to diseases and disorder, a growth trajectory needs to be defined.

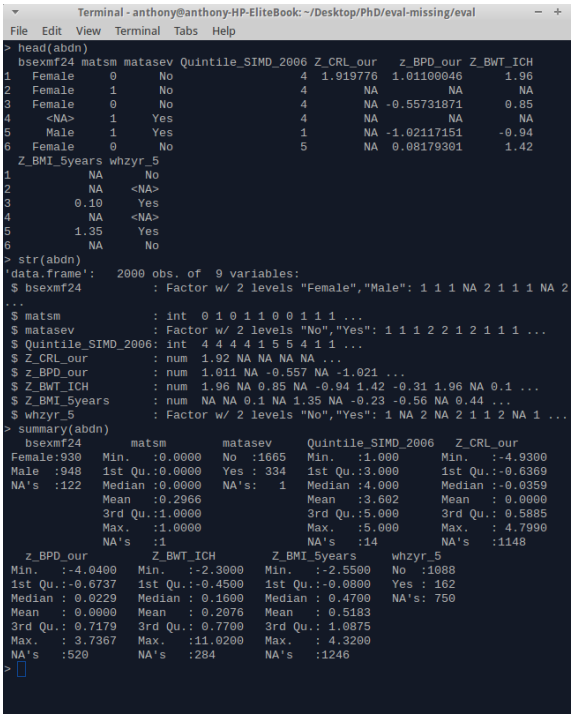

Figure 2.1: Summary of one of the sample datasets.

The datasets consist of growth measurements, which alone are not enough to describe a growth trajectories that could represent the whole data. What we need is a model to describe a growth trajectory in terms of the growth measurements and produces a growth curve or formula. 15‘ ’ 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 our current ideas, some topics related to the research questions and some of the methods we believe will help solve the questions.

3.1 Imputation

Imputation is the process of replacing missing fields with values[8]. There is a huge array of imputation techniques ranging from straight forward default value imputation [11], to mean-value imputation [7] or even imputation by equations [9, 28].

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 field 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) [28] 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 [19] has some very good clustering packages available [18] for the public to use. The most widely used ones are "cluster" [13], "cclust" [5], "fclust" [10] and "mclust" [6]. 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 used. 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

Our group have related fetal measurements to postnatal outcomes in childhood which include asthma and eczema in a local population [23, 25, 24] and also one from Saudi Arabia. Our work and a systematic review of the literature demonstrates that small absolute fetal size and either accelerating or faltering fetal growth are all associated with adverse outcomes.

The methods used are simple statistical regressions, with confidence intervals to indicate how valid the analysis is. Our new ideas consist of efficient computational approaches to solve such problems in a greater scale as well as with greater confidence in the findings.

Our proposed approach is differs as of the biggest problems in using statistical models for analysis, is that regardless of the confidence level, they are wrong. The analyst's job is to find the least wrong. Our approach will be less probabilistic and thus give the outcomes more confidence.

3.4 Regression Modelling

An important aspect of this project is to choose an appropriate way to model fetal growth from the maternity data. R has many regression modelling packages, the ones considered for this project are a linear modelling package (lm [20]), a mixed effects modelling modelling package (lme4 [3]) and a latent growth modelling package (lavaan [14]).

Each package treats the data slightly differently, currently we are focusing on addressing the issue of missingness and thus are using a simple linear model under the package "lm" to evaluate imputed datasets.

Mixed effects modelling techniques are not required at this stage of the project although they might be useful if we want to find complex inter-field relationships, ie relationships between fields, whilst simultaneously performing a regression model.

We also have an idea, its at early stages yet, for using latent growth modelling to create a fetal growth trajectory. Currently, the data contains only fetal size variables and maternal details, not any actual growth measurements. Thus we can not create a regression model on growth itself, we can only model size in terms of the other variables. Using latent growth modelling we can eliminate that as we can take growth to be our latent varibale.

Chapter 4: Progress

4.1 Presentation Skills

At the beginning of this PhD I attended two courses in presentation skills and one in good clinical practice (GCP). The presentation skills courses, offered by the university, were very educational. The first focused on key factors for giving a memorable, entertaining and educational presentation, it included some interesting techniques to keep the audience focused. The second concentrated on how to use ones voice when presenting work, it included basic voice training, warming techniques and some practise.

4.2 Approvals and training

Approvals for gaining access to the routinely acquired maternity data for analysis have been submitted. Coursed required to acquire the data have been identified and have been taken or registration has been done and we are waiting to do the courses.

4.3 Italian Partners

At early stages of this project, we became acquainted with a Lucia Vaira, a second year PhD student at the University of Salento, Italy. Vaira's research is in fetal growth curves and a unified method for global fetal growth analysis [4, 1, 26].

After brief talks on our relevant fields of research, we discussed the possibility of collaborating together. We are currently writing a joint research paper focusing on fetal growth regression modelling. We have not yet identified a conference to present the paper and the paper is near completion.

4.4 ACERO Symposium

The Aberdeen Centre for Energy Regulation and Obesity (ACERO) held their annual symposium at the Rowett Institute of Nutrition and Health in Aberdeen[15]. I volunteered to give a short presentation on how to evaluate artificially completing datasets.

It was a great experience, nerve-racking but good practise for future presentations. The questions at the end raised good points and it gave me a chase to experience public speaking. The courses taken on presentation skills earlier in the year certainly came to good use.

One this to work on from this experience is to finish presentations on harder notes. We need to reinforce to the audience what exactly they should take away from the presentation. Overall, it was a success.

4.5 FARR International

In August 2015 the Farr Institute shall be hosting their first International Conference in St Andrews[16]. I have submitted an abstract and await their decision on whether I shall be presenting there. It will be great experience regardless on whether I present or not as it will let me

know meet other researchers from around the world.

4.6 FARR PhD Symposium

The Farr Institute PhD Symposium will be held on 9th June in Manchester[17] and is open to all PhD students associated with the Farr Institute. All PhD students who wish to go to the symposium are required to submit an abstract, we will then be chosen to either give a talk or present a poster.

Again, this is a great chance to meet other PhD students from across the United Kingdom. If chosen for a presentation, I will talk about missingness and my plan on how I'm going to carry out the rest of this PhD.

Chapter 5: Future Plans

Chapter 6: 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] Kennedy Bailey. Numerical taxonomy and cluster analysis, 1994.

[3] Douglas Bates, Martin Maechler, and Ben Bolker. Linear mixed-effects models using eigen and s4. `http://cran.r-project.org/web/packages/lme4/`, 2015. Accessed: 14 May 2015.

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

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

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

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

[8] Andrew Gelman and Jennifer Hill. Data analysis using regression and multilevel/hierarchical models, 2006.

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

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

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

[12] A. Jain and R. Dubes. Algorithms for clustering data, 1988.

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

[14] Yves Rosseel, Dan Oberski, and Jarrett Byrnes. Latent variable analysis. `http://cran.r-project.org/web/packages/lavaan/`, 2015. Accessed: 14 May 2015.

[15] ACERO team. Aberdeen centre for energy regulation and obesity. http://www.abdn.ac.uk/acero/symposia/symp15/, 2015. Accessed: 13 May 2015.

[16] FARR Research Team. The farr institute of health informatics research. http://farrinstandrews.org/, 2015. Accessed: 13 May 2015.

[17] FARR Research Team. The farr institute of health informatics research. http://www.farrinstitute.org/events/82/2015-06-09/the-farr-institute-phd-symposium-2015.html, 2015. Accessed: 13 May 2015.

[18] R Development Core Team. The comprehensive r archive network. http://cran.r-project.org/, 1993. Accessed: 10 May 2015.

[19] R Development Core Team. The r project for statistical computing. http://www.r-project.org/, 1993. Accessed: 10 May 2015.

[20] R Development Core Team. The r project for statistical computing. , 1993. Accessed: 14 May 2015.

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

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

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

[24] Steve Turner and Graham Devereux. Fetal ultrasound: Shedding light or casting shadows on the fetal origins of airway disease, 2012.

[25] Steve Turner, Nanda Prabhu, Peter Danielian, and Geraldine McNeill. Perinatal programming of childhood asthma: Early fetal size, growth trajectory during infancy, and childhood asthma outcomes, 2011.

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

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

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