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, 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.

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 Background

Our group have 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]. 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 previously used are statistical regressions and simple data modelling. These analyses consist of confidence intervals and other tests to indicate how valid the analysis is. Our new ideas consist of efficient computational approaches to find relationships between fetal and maternal data to non-communicable diseases (NCD) in children and adults in a greater scale as well as with greater confidence in any findings.

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 nation wide or even internationally.

Research within this field is being carried out throughout the world. Generation R projects have included asthma origins [5] as well as their symptoms in early childhood [6]. Other researchers from Italy and Russia are also looking at fetal growth trajectories in order to understand any relationship between fetal growth and NCD [7, 8, 9, 10].

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 NCD.

1.2 Research Question

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

1.2.1 Methodological

Our first question relates to the analysis of partially complete data and routinely acquired data which are inevitably incomplete, for example in the SEATON study [11, 12] 50% or first trimester and 25% of second trimester scans were missing. When there are large amounts of missing values in a dataset, can we still perform efficient data analysis, we consider this using imputation techniques. Thus, our hypothesis is: Data analysis is possible and efficient using appropriate imputation techniques on datasets with partially complete records.

How accurate is gestational assessment. To address this we will compare fetal size for pregnancies conceived by donor insemination (DI, where the date of conception is certain) to pregnancies conceived without assistance. Our hypothesis is that first, second and third trimester fetal size

is no different for pregnancies conceived by DI and without assistance.

1.2.2 Epidemiological

Using cluster analysis, one can gather complex relationships between fields in datasets. In this manner we are going to find the relationships between fetal and maternal characteristics to NCD in children and adults. Thus, our question is: What are the relationships between fetal measurements (including maternal characteristics) and NCD. One early hypothesis is that steady growth will be associated with a decreased risk for postnatal NCD, another hypothesis would then be that change of growth will increase the risk for postnatal NCD.

Chapter 2: Methodology

Our group has acquired a number of sample datasets, these datasets represent the actual data we will be analysing once we are granted permission. 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 Imputation

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 lack of reliability or confidence in any analytic results. For example, 30% of population would be enough to confidently state something about the population as a result of analysing 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[13]. There exist a full range of imputation techniques from simple default value substitution (ie replacing all missing values with some values)[14], slightly more complex ways such as mean values substitution [15](similar to default value substitution except here the values may change according to the dataset), to complicated imputation algorithms which use probabilistic equations to replace missing values depending on the know values[16, 17].

Default value imputation techniques are not appropriate as 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 considers 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) [16] considers both the relationship between the fields of each record and the behaviour of all the other records in the dataset. We have chosen to use MICE to impute the data, given that all data behaves differently, a method for evaluating the efficiency of MICE on our datasets will have to be found.

2.2 Clustering

We have been looking at ways to incorporate clustering techniques for real world data analysis. Clustering is a way of grouping data into different sections, called clusters, these clustering will contain data points which are most similar to each other [18]. The way a computer considers data

points to be similar is problem dependant. By grouping the data into different clusters, we are able to analyse patterns, trends or relationships in datasets.

We hope to find that by clustering maternity data, antenatal behaviours and trends will be made clear. For example, we might find that a certain cluster who's records all have similar growth rates also have a some NCD, implying relationships. This approach eliminates any bias statistical regression modelling might induce by specifically looking for a certain relationship. Here, the algorithm looks for all relationships without focusing on any specific one.

By letting the algorithm analyse the data without focusing on specific relationships between trends and outcomes, we hope to find interesting and possibly unexpected results, moreover, by not specifically looking for such results, the results will be more reliable.

Cluster validation is used to evaluate cluster outcomes[19]. This is useful in order to assess the validity of a clustering, it can then be used to compare clustering algorithms or even different datasets against each other.

Using clustering and cluster validation techniques, we will be able to assess the efficiency and other effects imputation will cause when you create artificially complete datasets from partially missing ones. We will also be able to evaluate the efficiency as well as the correctness or confidence in any outcomes we get by clustering data.

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.

R, the statistical program [20], has some very good clustering packages available [21] for the public to use. The most widely used ones are "cluster" [22], "cclust" [23], "fclust" [24] and "mclust" [25]. They can all perform similar clustering techniques, they differ in terms of efficiency and the type of data they can cluster efficiently.

The package "mclust" 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 as other packages would perform the same clustering without over-complicating 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 ambiguous. It can cluster data with levels of certainty where other clustering techniques have a binary approach, it is either in a cluster or not. 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 and evaluations will have to be carried out to see which should be used. My prediction for the most efficient way to do this would be; use "cclust" for finding the optimal number of clusters and then, using this number, we perform a clustering using the package "cluster". It will need to be tested but by using both their strengths, I might have the best clustering available

2.3 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 [26]), a mixed effects modelling modelling package (lme4 [27]) and a latent growth modelling package (lavaan [28]).

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 variable.

2.4 Growth Trajectories

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

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 data have to be taken into account.

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. This is no to be confused with mixed effects modelling which uses fixed and random effect equations for statistical modelling.

2.5 Databases

A summary of the dataset can been seen in Figure 2.1, here, N stands of number of observations (there are 2000 records in the dataset), the rest of the columns are self-explanatory. Notice that categorical values such as matsm and Quintile_SIMD_2006 are represented my numerical values which can be analysed using numeical methods. For example we can see that the mean for matsm is 0.297, meaning that most mothers do not smoke. The first five records are shown in Figure 2.2

Table 2.1: Summary of the sample data

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

The sample dataset consists of the following fields:

- bsexmf24 : sex of the fetus (Male, Female)

- matsm : does the mother smoke (0,1 - no,yes)

- matasev : has the mother has asthma (yes,no)

- Quintile_SIMD_2006 : economic status of mother (1 to 5 - 5 being wealthy)

- Z_CRL_our : first trimester scan (continuous number)

- z_BPD_our : second trimester scan (continuous number)

- Z_BWT_IC : third trimester, birth weight (continuous number)

- Z_BMI_5years : BMI score at ages 5 (continuous number)

Table 2.2: First 5 records from sample dataset

	bsexmf24	matsm	matasev	Quintile_SIMD_2006	Z_CRL_our
1	Female	0	No	4	1.920
2	Female	1	No	4	
3	Female	0	No	4	
4		1	Yes	4	
5	Male	1	Yes	1	

	z_BPD_our	Z_BWT_ICH	Z_BMI_5years	whzyr_5
1	1.011	1.960		No
2				
3	-0.557	0.850	0.100	Yes
4				
5	-1.021	-0.940	1.350	Yes

Chapter 3: Training and Approvals

3.1 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.

3.2 Statistics

Coursera [29], an education platform that partners with top universities, provides many online courses which anyone can take. They have courses in applied statistics which have been very helpful in letting me catch on and learn all the statistical knowledge required for this PhD.

3.3 Good Clinical Practise

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 (GCP) or registration has been completed and we are waiting to do the courses (information governance online course).

3.4 PAC Application

We are in the process of submitting an application to the Privacy Advisory Committee (PAC) [30]. This application needs to be approved in order to gain access to NHS Scottish maternity data.

Chapter 4: Results

4.1 Missingness

At first glance, we noticed that that the sample dataset had a large proportion of data missing. This is illustrated in Figure 4.1. We can see that the first column has a missing percentage of around 16%, the second has a percentage of 24% and so on.
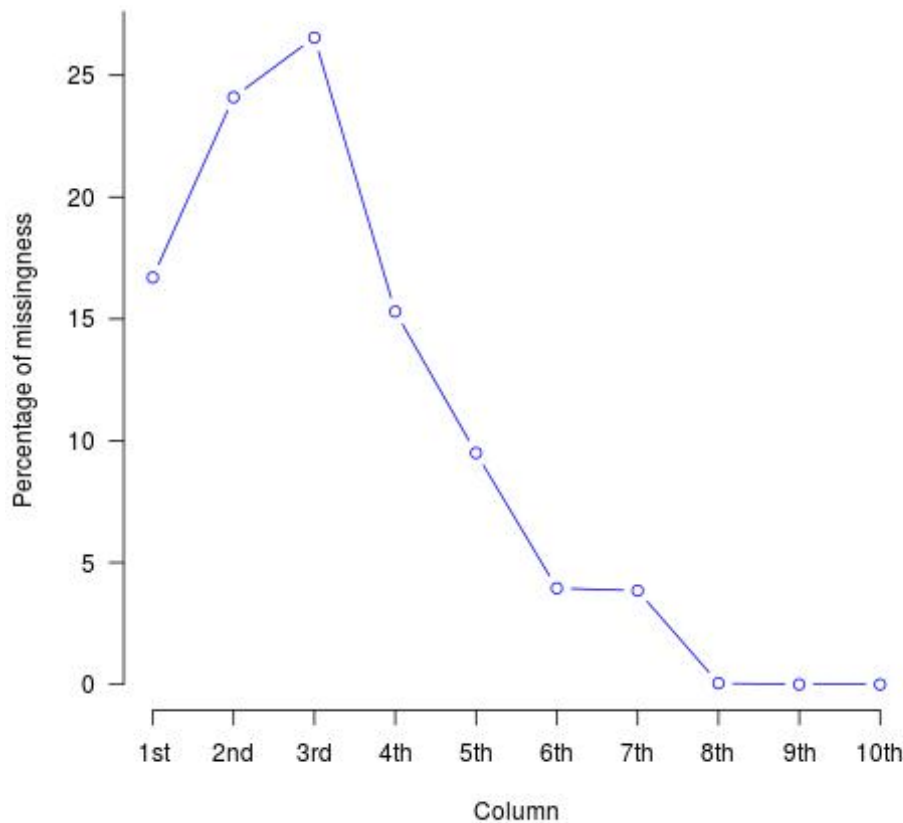

Figure 4.1: Summary of missing data from sample dataset.

In order to test whether imputation creates a realistic artificially complete dataset, we will compare the results after fitting regressions models to several chosen datasets. Firstly we will start with a complete dataset, using the information from Figure 4.1, we will create several new datasets by copying the complete dataset and induce missingness into it sever times. Once we have several artificially incomplete datasets, we will apply the MICE algorithm to create several artificially complete datasets. Finally, we will apply the same regression model on each dataset in order to compare the results. The idea being that if imputation works, the resulting models will be the same.

From Table 4.1, column (1) represents the complete dataset and columns (2) and (3) represent two artificially complete datasets. 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 you can see that that they values are very close to each other. This is enough to indicate that imputation works on this dataset.

Table 4.1: (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_CRL_our	0.080	0.031	0.023
	(0.053)	(0.059)	(0.059)
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: $^{*}p<0.1; ^{**}p<0.05; ^{***}p<0.01$

4.2 ACERO

The Aberdeen Centre for Energy Regulation and Obesity (ACERO) held their annual symposium at the Rowett Institute of Nutrition and Health in Aberdeen[31]. I gave 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 things 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.

4.3 Farr PhD Symposium

The Farr Institute PhD Symposium will be held on 9th June 2015 in Manchester[32] 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.

4.4 FARR International

In August 2015 the Farr Institute shall be hosting their first International Conference in St Andrews[33]. 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.

4.5 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 [7, 8, 9].

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.

Chapter 5: Future Plans

I will do stuffs.

5.1   The next steps:

- FARR PhD symposium academic poster
  - On the 10th June 2015, I will be attending the annual FARR PhD symposium where I have been asked to create an academic poster and present it at the symposium. The poster will be based on our efforts to create artificially complete data and some results showing that the new data is a good representation of the truth.

- Finish and submit PAC application for approval
  - I am in the process of completing this lengthy application in order to send it soon. Once this is send, we can undertake other courses required for the approval to access data.

- Finish joint paper with Lucia and find appropriate place to publish
  - I have to carry on working with Lucia Vaira, an Italian PhD student currently working on way to model fetal growth. I will need to finish several case studies testing the efficiency of the proposed fetal growth regression model. Once the paper is complete (or whilst we are completing it) we will need to find a suitable conference, journal or symposium to publish and/or present the paper. We will then need to arrange funding, travel and accommodation for trip.

- FARR Institute International Conference 2015 academic poster
  - On the 27th August 2015, I will be attending the annual FARR International Conference where I may be asked to create an academic poster to present or an oral presentation to give at the conference. Either will be based on our efforts to create artificially complete data and some results showing that the new data is a good representation of the truth.

- Analyse IVF data after approvals are granted
  - Soon, I will be grated permission to access and analyse IVF maternity data. I will test the imputation, regression and clustering theories we have been studying on these data. We will be able to fine tune any necessary ideas from the results when testing on IVF data. Any results we find will also be important to the assisted reproduction community.

- Take information governance course with Steve
  - In order to get access to Aberdeen's neonatal records, we will need to complete and information governance online course. Without this, we do not have permission to access or perform any analysis on the data.

Bibliography

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

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

[7] L. A. M. A. T. A. Bochicchio Mario, Vaira Lucia, "Fpgt: An online system for customized fetal and pediatric growth tracking," 2014.

[8] L. V. A. M. 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.

[9] L. Vaira, "Quantitative fetal growth curves comparison: a collaborative approach," 2014.

[10] L. Vaira and M. Bochicchio, "Are static fetal growth charts still suitable for diagnostic purposes?," 2014.

[11] S. Willers, G. Devereux, L. Craig, G. McNeill, A. Wijga, W. Abou El-Magd, S. Turner, P. Helms, and A. Seaton, "Maternal food consumption during pregnancy and asthma, respiratory and atopic symptoms in 5-year-old children.," 2007.

[12] S. Turner, G. Devereux, D. Campbell, N. Smith, L. Craig, G. McNeill, and S. Forbes, "Associations between fetal size, maternal alpha-tocopherol and childhood asthma.," 2010.

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

[14] T. Hastie, R. Tibshirani, B. Narasinhan, and G. Chu, "impute: Imputation for microarray data," 2015.

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

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

[17] A. Gelman, J. Hill, and Y.-S. Su, "Missing data imputation and model checking," 2015.

[18] K. Bailey, "Numerical taxonomy and cluster analysis," 1994.

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

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

[21] R. D. C. Team, "The comprehensive r archive network." `http://cran.r-project.org/`, 1993. Accessed: 10 May 2015.

[22] M. Maecchler and P. Rousseeruw, "Cluster analysis extended," 2015.

[23] E. Dimitriadou and K. Hornik, "Convex clustering methods and clustering indexes," 2015.

[24] P. Giordani and M. Ferraro, "Fuzzy clustering," 2015.

[25] C. Fraley, A. E. Raftery, and L. Scrucca, "Normal mixture modelling for model-based clustering," 2015.

[26] R. D. C. Team, "The r project for statistical computing." , 1993. Accessed: 14 May 2015.

[27] D. Bates, M. Maechler, and B. Bolker, "Linear mixed-effects models using eigen and s4." `http://cran.r-project.org/web/packages/lme4/`, 2015. Accessed: 14 May 2015.

[28] Y. Rosseel, D. Oberski, and J. Byrnes, "Latent variable analysis." `http://cran.r-project.org/web/packages/lavaan/`, 2015. Accessed: 14 May 2015.

[29] C. team, "Coursera online courses." `https://www.coursera.org/`, 2015. Accessed: 18 May 2015.

[30] NHSNSS, "Pac application process." `http://www.nhsnss.org/pages/corporate/about_pac.php`, 2015. Accessed: 24 May 2015.

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

[32] F. R. 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.

[33] F. R. Team, "The farr institute of health informatics research." `http://farrinstandrews.org/`, 2015. Accessed: 13 May 2015.