APPLICATION OF WHOLE GENOME AMPLIFICATION FOR THE INVESTIGATION OF GENOMIC MUTATIONS IN HODGKIN’S LYMPHOMA

by

PRADEEP RAMAGIRI

This project is submitted in partial fulfilment of the requirements for the award of the MRes in Biomedical Research

School of Cancer Sciences
College of Medical and Dental Sciences
University of Birmingham
August 2012
This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.
Abstract

Tumour cells of classical Hodgkin’s lymphoma (cHL) comprise only a fraction (1% or less) of the total cellular infiltrate. The low tumour cell numbers pose a great difficulty in the investigation of underlying genetic events in cHL development. This hurdle can be overcome by exome sequencing of a small number (~50) of isolated HL cells. This study was aimed at establishing ‘whole genome amplification (WGA) of the HL cell lines and patients’ samples, which can be potentially used to for the exome sequencing. This study used 6 HL cell lines (HDLM-2, KM-H2, L-428, L540, L-591, and L-1236) and 4 HL patients’ samples (1.5, 6.4, 7.28, and 7.8) to generate PCR products for exons 2 to 9 of the TNFAIP3 gene. We also aimed to demonstrate whether microdissection and whole genome amplification methods will introduce any changes to the sequences, by analysing the mutation spectra for the whole genome ‘amplified’, and ‘unamplified’ (KM-H2 and L-591) cell lines. We obtained good quality WGA product for all the microdissected cell line samples and patient samples. However, only patient sample 6.4 amplified all of these exons except the exon 6 and 7.1 in the following PCR experiments. The quality of the sequences obtained for the amplified cell line samples (76.5%) was as good as the quality for the unamplified cell lines (79.7%). We observed same nucleotide changes in the cell lines KM-H2 (deletion of intron 2– exon 6 region), and L-1236 (G491A) as reported in a previous study (16).
Acknowledgement

I sincerely thank Dr. Wenbin Wei, Prof. Paul Murray, and Dr. John Arrand for their kind support and guidance.
Table of contents

1) Introduction ..................................................................................................................... 6

  1.1) Molecular biology of Hodgkin’s lymphoma ............................................................. 6

  1.2) Deregulation of multiple signalling pathways in HRS cells ........................................ 7

    1.2.1) Anti-apoptotic mechanisms in HRS cells ......................................................... 7

    1.2.2) NF-κB pathway activation ............................................................................. 7

  1.3) TNFAIP3 ..................................................................................................................... 7

  1.4) Microdissection method ......................................................................................... 8

  1.5) Whole genome amplification (WGA) .................................................................... 8

  1.6) Aim ......................................................................................................................... 9

2) Materials and Methods ................................................................................................ 10

  2.1) COSMIC (Catalogue of Somatic Mutations in Cancer) search for the frequently mutated genes in Hodgkin’s lymphoma ............................................................................. 10

  2.2) Patient samples and cell lines .............................................................................. 10

  2.3) Laser microdissection and pressure catapulting of cells ........................................ 10

  2.4) DNA extraction of the HL cell lines ..................................................................... 11

  2.5) Whole genome amplification (WGA) of the HL cell lines ..................................... 12

  2.6) PCR ....................................................................................................................... 12

  2.7) PCR product purification .................................................................................... 13

  2.8) Sequencing ........................................................................................................... 13

  2.9) Sequence Analysis .............................................................................................. 13

3) Results ........................................................................................................................... 14

  3.1) COSMIC search of frequently mutated genes in Hodgkin’s lymphoma .................. 14

  3.2) PCR of TNFAIP3 exons using unamplified genomic DNA ..................................... 14

  3.3) Whole genome amplification .............................................................................. 15

  3.4) Whole genome amplification – PCR .................................................................. 16

  3.5) Quality of DNA sequences from unamplified cell lines and amplified cell lines and patient samples .......................................................................................................................... 17
3.6) Comparison of WGA amplified and unamplified DNA sequence .............................................. 19
3.7) Analysis of the TNFAIP3 mutations in HL cell lines and patient sample............................... 19
4) Discussion.......................................................................................................................... 22
6) SUPPLEMENTARY NOTES............................................................................................ 26

Table of figures

Figure 1. Microdissection of a single KM-H2 cell. ................................................................................. 11
Figure 2. Genes frequently mutated in Hodgkin’s lymphoma.............................................................. 14
Figure 3. Gel images of the PCR experiments...................................................................................... 15
Figure 4. Images of a single KM-H2 cell, and 50 L-591 cells WGA products is showed....................... 15
Figure 5. KM-H2, L-591 and patient samples amplification using WGA4 kit................................. 16
Figure 6. Test PCR on WGA4 products of KM-H2 and L-591 samples.............................................. 17
Figure 7. PCR on WGA4 products of KM-H2, L-591, and patient samples...................................... 17
Figure 8. Chromatograms showing nucleotide changes observed in the cell lines and patient samples ........................................................................................................................................... 20
1) Introduction

Hodgkin’s lymphoma (HL) is one of the most prevalent types of malignant lymphomas in the western countries, with an incidence rate of 3 in 100,000 persons each year. In the United Kingdom (UK), 1,852 people diagnosed with HL, and 319 HL related deaths were registered between the year 2009 and 2010. Although HL accounts for less than 1% of all cancers in the UK (3), about 4-5% of all childhood cancers in the UK (around 64 children per year) are registered as HL. HL is predominant in older children (between 10 to 14 years old) and no infant cases with HL are registered so far (12). HL can be successfully treated in the developed countries like the UK. However, the mortality rate is high in under-developed countries (one in two registered cases), and developing countries (one in four cases) (3).

This disease was first described in 1832 by Thomas Hodgkin (15). A majority of HL cases (95%) belong to the classical form of the disease and the remaining 5% are of nodular lymphocyte predominant HL (NLPHL). Based on the histology and cellular composition, classical HL is subdivided into nodular sclerosis (NS), mixed cellularity (MC), lymphocyte-depleted (LD), and lymphocyte-rich (LR) forms. The tumour cells in classical Hodgkin’s lymphoma (cHL) are known as the Hodgkin and Reed-Sternberg (HRS) cells, and in nodular lymphocyte predominant lymphoma (NLPHL) they are known as lymphocyte predominant (LP) cells. Irrespective of the subtype, the tumour cells account for only 1-10% of the total cellular infiltrate. This low tumour cell content along with poor chromosome morphology has made it very difficult for researchers to investigate the molecular events in the development of HL.

1.1) Molecular biology of Hodgkin’s lymphoma

Germinal centre (GC) B cells that escape apoptosis after acquiring fatal immunoglobulin V gene mutations are believed to give rise to HRS cells. In rare cases (~ 2%), HRS cells originate from T cells. On the other hand, LP cells derive from antigen-selected GC B cells (2).

Comparative genomic hybridisation (CGH) was used to detect copy number aberrations in HRS cells. Chromosomal arms 1p, 6q, 7q, 11q, 12q, and 14q have shown recurrent multiple chromosomal break points. Frequent recurrent gains have been observed on chromosomes 2, 5, 9, and 12, whereas, frequent losses are reported on chromosomes 13, 21 and Y (9). The genomic instability of the HRS and LP cells will depend on the complexity of genetic alterations and poses difficulty in identifying the genetic aberrations related to pathogenesis (2).
1.2) Deregulation of multiple signalling pathways in HRS cells

In recent years, deregulated expression of several proteins and aberrant activation of a number of signalling pathways that are associated with the pathogenesis of HRS cells have been identified, namely, the transcription factors (NF-κB, STAT, AP-1, and Notch1), multiple receptor tyrosine kinase (RTKs), PI3K, and MEK/ERK pathways. These pathways are vital for the proliferation and survival of the HRS cell phenotype (2).

1.2.1) Anti-apoptotic mechanisms in HRS cells

As the HRS cells are derived from pre-apoptotic GC B cells, evading apoptosis is vital in the transformation. As mentioned above several signalling pathways are aberrantly activated to facilitate the generation of HRS cells, most importantly NF-κB.

1.2.2) NF-κB pathway activation

The NF-κB transcription factor family consists of five members, Rel A, Rel B, c-Rel, NF-κB1, and NF-κB2, which can act as homo- and/or heterodimers. In the classical NF-κB pathway, the p50/p65 dimer was kept in the cytoplasm by binding to IkBα, a NF-κB inhibitor. As a result of activating the pathway, IKK (IkB kinase complex) kinases phosphorylate IkBα and thereby induce its proteosomal degradation resulting in the translocation of NF-κB dimers into the nucleus and which subsequently activate multiple target genes. This activation can be partially mediated through receptor signalling, such as the TNF family member CD40. Other receptors involved in similar NF-κB activation include CD30, TACI, BCMA, RANK and Notch1 (15).

However, these signalling pathways are not sufficient for the strong constitutive NF-κB activation. Several genetic aberrations have been identified in HRS cells that affect the NF-κB pathway. The REL gene show copy number gains or amplification in 40% to 50% of HL cases (15). The NFKBA gene (which encodes IkBα and IkBe) has also been shown to have mutations. Recent studies have shown TNFAIP3 (Tumour Necrosis Factor, Alpha-induced Protein-3) as a novel tumour suppressor gene that is involved in the regulation of the NF-κB pathway. The TNFAIP3 gene encodes A20 protein that has ubiquitinase and deubiquitinase functions. A20 is a negative regulator of NF-κB signalling and acts upstream of the IKK kinases (15). Some of the previous studies on HL described mutations/deletions in this gene in 30%-40% of the cases (15-16).

1.3) TNFAIP3

TNFAIP3 gene expression is induced by tumour necrosis factor (TNF). TNFAIP3 encodes a zinc-finger protein, A20, which is known to inhibit NF-κB activation and TNF-mediated apoptosis (18). The TNFAIP3 gene is located on Chromosome 6q23
and has 9 exons. Coding sequence of this gene is 2373 nucleotides long, and encodes 790 amino acids (20).

*TNFAIP3* mutation seems to have an inverse correlation with the presence of Epstein-Barr virus (EBV) in HRS cells. Around 70% of EBV-ve cases showed *TNFAIP3* mutations (mostly deleterious) whereas only 12% of the EBV infected patients showed mis-sense mutations in *TNFAIP3* gene. This might indicate an alternative NF-κB activation mechanism involving A20 inactivation and EBV infection (16). Functional studies involving re-expression of A20 in A20 deficient HL cell lines resulted in reduced expression of NF-κB target genes and negatively affected cell survival, demonstrating that A20 is a tumour suppressor (16). The role of *TNFAIP3* as a tumour suppressor gene has also been demonstrated in other lymphomas, such as primary mediastinal B cell lymphoma (PMBL), which are also characterized by constitutive activation of the NF-κB pathway, as mentioned above (16).

For the better understanding of the genetic alterations and their role in the HL development, we need to isolate the tumour cells and study the variations across the whole genome.

### 1.4) Microdissection method

Microdissection techniques can be divided into three main categories: manual extraction, selective ablation, and laser capture microdissection (LCM). In this study we used LCM method to isolate the tumour cells from the fixed cell line samples. LCM was first designed by 'Lance Lotta' and co-workers in the mid 1990 at the NIH, Maryland, USA. There are 3 main types of LCMs, Infrared (IR) LCM, Ultra violet (UV) LCM, and a combined version (IR/UV LCM). In the IR LCM method, the surrounding tissue of the target specimen were filled with a melted ethylene vinyl acetate (EVA) by laser activation and then the target specimen selectively adhere to the thermoplastic membrane by low energy laser pulse. The targets are then extracted by removing the polymer from the tissue surface. This method uses low energy beams so low photo chemical effects on the specimen, although visualization in this method is fuzzy. UV LCM method, in contrary, selects and cuts the target specimen with a fine laser beam. This method avoids the unwanted debris surrounding the specimen and aided with better visualization. We used UV LCM method to microdissect the HL cells in our study (31).

### 1.5) Whole genome amplification (WGA)

A single cell usually contains ~6-7 picograms (pg) of genomic DNA (30). Most of the sequencing based methods such as whole genome sequencing or exome sequencing methods require ~3-5 micrograms (µg) of a good quality starting material (DNA) per reaction. Hence, we need to amplify the genome by more than 1000 fold, if we are working on a sample material which contains only few (1-100) cells.
Various commercial WGA kits are available to meet the above requirements, such as QIAGen’s ‘REPLI-g’, GE Healthcare’s ‘GenomiPhi’, and Sigma’s ‘GenomePlex’. These kits work on the either ‘PCR’ or ‘multiple displacement amplification’ (MDA) method. PCR-based methods use DNA polymerase from a thermophilic bacterium (*Thermus aquaticus*) and amplify the DNA by repeated ‘denaturation-elongation’ steps at various temperatures. MDA method employs ‘bacteriophage (phi 29) DNA polymerase’ for the rolling circle amplification of a DNA template at a ‘constant’ temperature (isothermal). MDA method is much slower (~18 hours) than the PCR based method (~3 hours) (29).

The above commercial kits facilitate a straight forward application of WGA method to generate the necessary quality and quantity of the product. However, the choice of the kit will depend on the research question and the sample material (old, degraded, fixed, paraffin embedded and frozen etc.) as the reliability (percentage of samples met the requirement), fidelity, and coverage will vary between these technologies. For genotyping applications, the longer MDA method (‘REPLI-g’) found to have better ‘accuracy’ and coverage than the faster PCR method (Genomeplex) (29).

1.6) **Aim**

In this study, we aimed 1) to establish ‘whole genome amplification method’ that can be potentially used in genome sequencing of HL cell lines and HL patient samples; 2) to confirm genetic changes in the *TNFAIP3* gene of HL cell lines discovered by a previous study; 3) to investigate whether there are genetic changes in the *TNFAIP3* gene of HL cell line L540; 4) to investigate whether microdissection and whole genome amplification methods will introduce any changes to the genome sequences using the KM-H2 and L-591 cell lines by comparing WGA amplified and unamplified DNA; and 5) to investigate the applicability of the whole genome amplification method to the HL patient samples.
2) Materials and Methods

2.1) COSMIC (Catalogue of Somatic Mutations in Cancer) search for the frequently mutated genes in Hodgkin’s lymphoma

COSMIC is a comprehensive database designed to gather, curate, organise, and present information on somatic mutations in cancer and put that information in the public domain (23). This database allows researchers to investigate the key oncogenes, gene fusions and structural rearrangement annotations across numerous cancer samples. COSMIC integrates the manually curated cancer mutation data from scientific literature with the output from the Cancer Genome Project (CGP) (5). I searched for the genes that are heavily mutated in the Hodgkin’s lymphoma in this database by selecting the type of cancer and the histology subtype using COSMIC web interface.

2.2) Patient samples and cell lines

This study was part of an ongoing research that has been approved by the National Children’s Cancer and Leukaemia Group (CCLG) Biological Studies Steering Group. These samples were collected by the tissue bank of the National Children’s Cancer and Leukaemia Group (CCLG). Four patient samples that were used in this study (labelled as 1.5, 6.4, 7.21 and 7.8) were already fixed and microdissected into ‘PALM membrane adhesive cap’ (PALM tubes) (50 cells in each tube) and stored at -20°C.

Human Hodgkin’s lymphoma cell lines HDLM-2 (DSMZ No: ACC 17), KM-H2 (DSMZ No: ACC 8), L-428 (DSMZ No: ACC 197), L-540 (DSMZ No: ACC 72), L-591(DSMZ No: ACC 602), and L-1236 (DSMZ No: ACC 530) were obtained from Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures, Inhoffenstraße 7B, 38124 Braunschweig, GERMANY (11).

All the 6 cultured HL cell lines, HDLM-2, KM-H2, L-428, L-540, L-591, and L-1236, were cultured and kindly provided by Eszter Nagy, a PhD student at Prof. Paul Murray’s lab, School of Cancer Sciences, College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK.

2.3) Laser microdissection and pressure catapulting of cells

Appropriate volumes of cultured KM-H2 (6.25 mL at 8X10⁴ cells/mL concentration) and L-591 (1.47 mL at 17X10⁴ cells/mL) were centrifuged at 3000 r.p.m for 5 minutes and the pellet was re-suspended with 500 µL Phosphate Buffered Saline (PBS) and 100 µL of this spun on to membrane-covered slides (PALM) using a Cytospin at 1000 r.p.m. for 5 minutes. Immediately, the slides were air dried and were fixed using cold ethanol and stained with Haematoxylin. This entire procedure was performed on ice.
Then the slides were air dried and stored at -20°C until the microdissection was performed. The slides were then taken to the Human Biomaterials Resource Centre, College of Medical and Dental Sciences, University of Birmingham, UK (10), samples were microdissected using the PALM microbeam HBO100/AX10 Laser Capture Microdissection (LCM) apparatus (Carl Zeiss PALM Microbeam, Carl Zeiss Ltd, Cambridge, UK) (27) After calibrating the ‘power and focus’ (Energy = 40, Focus= 69, LPC= -20, and Delta = 2), the cells were selected for cutting and capturing on to the caps of ‘PALM membrane adhesive cap’ tubes (PALM tubes).

Figure 1 show how a single cell on the KM-H2 PALM membrane slide was selected and captured. The slides were micro-dissected in different batches of a single cell and 50 cells onto the cap of PALM tubes. The caps of these tubes were checked under the LCM for the presence of cell sections adhered to them. The ‘PALM membrane adhesive cap’ tubes with cells were stored at -20°C until they were used in the whole genome amplification method.

2.4) DNA extraction of the HL cell lines

All the 6 HL cell lines, HDLM-2, KM-H2, L-428, L-540, L-591 and L-1236 were cultured and appropriate volumes of the cell cultures (3X10^5 cells) spun down and their pellet was re-suspended in PBS and sent to the Genomics lab, at the University of Birmingham for the DNA extraction (7).
2.5) Whole genome amplification (WGA) of the HL cell lines

Whole genome amplification of the microdissected cells was initially carried out using QIAGEN’s REPLI-g Mini Kit (25 X, Catalogue no.150023, 19) according to the manufacturer’s instructions. REPLI-g kit works on a principle known as Multiple Displacement Amplification (MDA) method where random primers (hexamers) bind to the template and generate fragments at a constant temperature with the help of a high fidelity DNA polymerase enzyme, usually ‘Φ29 DNA polymerase’. The resulting fragments of DNA are larger than conventional PCR products and with lower error frequency.

However, the REPLI-g Mini Kit didn’t give consistent amplification of the microdissected cell material. As an alternative, GenomePlex® Single Cell Whole Genome Amplification Kit (WGA4) from Sigma-Aldrich, UK (7) was used for whole genome amplification from the microdissected cells. WGA4 works on a principle of ‘non-enzymatic random fragmentation of the genomic DNA’ (28). The WGA was performed according to the manufacturer’s instructions. After the lysis procedure tubes that contained cell(s) were again checked under the microscope for the presence of cells on their caps to ensure the cells were lysed and ‘off’ their caps.

2.6) PCR

PCR primers designs were taken from a study by Roland Schmitz et al. (16) and used for the PCR and sequencing reactions (Supplementary Table 1). These primers have a melting temperature (Tm) ranging between 55°C and 74°C. Primers were ordered from Sigma-Aldrich, at 100 µM concentration (24). The primer stocks were diluted with nuclease free water as per the requirement (100 picomoles/µL for PCR, and 3.2 picomoles/ µL for sequencing reactions). The diluted primers and the stocks were stored at -20°C.

PCR was set up using QIAGEN’s multiplex PCR Kit for a total of 20 µL reaction mixture according to manufacturer’s instructions. Hot start PCR was programmed as follows: 95°C for 5 minutes (Denaturing of DNA) followed by 35 cycles of 94°C for 45 seconds, 65°C for 1 minute (Annealing) and 72°C for 1 minute (Extension) and then the final extension at 72°C for 10 minutes.

The PCR product was then validated by running 6.0 µL of the product and 3.0 µL of 100bp ladder from Invitrogen, UK (Catalogue.No.15628-019) on a 1.5% Agarose gel. The bands were checked for approximate length of the product for each set of reaction (Supplementary Table 2).
2.7) **PCR product purification**

ExoSAP-IT for PCR Product Clean-Up (catalogue no. 78200 200 UL), Affymatrix, UK, was used to clean up the PCR products according to the manufacturer’s instructions. The concentration of the clean PCR product was measured by loading 1.0 µL on the Nanodrop 1000 spectrophotometer (Thermo scientific, USA).

2.8) **Sequencing**

Clean PCR products were then diluted to the appropriate concentration (as requested by the Genomics lab) in a total of 10 µL reaction volume (Supplementary table 3) and sent to the sequencing service facility at Genomics lab, University of Birmingham, UK. These samples were later loaded on to the ABI 3730 Sequencing machine at the Genomics lab.

2.9) **Sequence Analysis**

Sequences retrieved from the Genomics lab, University of Birmingham, for the cell lines and patient sample were in the .ab1 format (format for the raw DNA data taken out from a scientific instrument and output from Applied Biosystem’s Sequence Analysis software). All these files were aligned and edited with the genomic reference sequence, (NCBI Reference Sequence: NC_000006.11, gi|224589818:138188581-138204449, 21) using ‘Sequencher 5.0 Demo’ software(17).
3) Results

3.1) COSMIC search of frequently mutated genes in Hodgkin’s lymphoma

COSMIC database was searched for the key cancer genes associated with Hodgkin’s lymphoma. The search yielded the top 20 genes most frequently mutated in the HL patient samples, namely SOCS1, TNFAIP3, NRAS, TP53, CYLD, HRAS, KDM6A, CDKN2A, CDKN2a(p14), BCR, ETV6, PCM1, NPM1, KRAS, JAK2, ATM, BRAF, PIK3CA, PTEN, and WT1. Of these 20 genes 3 genes were found to be mutated in higher percentage of patient population in which they were tested; namely, TNFAIP3, SOCS1, and NRAS (Figure 2). A literature review on the importance of these genes in HL was carried out to select a single gene for our study. All of these genes are important in the HL disease development. However, TNFAIP3 gene was selected on the basis of its central role in HL development and its location on 6q23 (this region is frequently deleted in B cell lymphomas). SOCS1 is located on chromosome 16p13.13 and the genomic instability in this region is not implicated in the HL development, so far.

![Genes frequently mutated in Hodgkin’s lymphoma.](image)

**Figure 2. Genes frequently mutated in Hodgkin’s lymphoma.** The red bar represents total number of samples and the blue bar represents the number of samples with mutations (Source: 23).

3.2) PCR of TNFAIP3 exons using unamplified genomic DNA

Test PCR reactions were set up with A20 E2F/E2R primers on all the six cell line DNA (Figure 3_1), and with all the TNFAIP3 gene primers on unamplified L-591 cell line to obtain optimal PCR conditions (Figure 3_2). Another PCR reaction was set up with all the primer sets for unamplified KM-H2 cell line sample. The PCR products were run on a 1.5% agarose and checked for the bands at the appropriate size (Supplementary Table 2).

As mentioned in some of the previous studies (16), a deletion of a region between intron 2 and exon 6 in the TNFAIP3 gene was observed in the unamplified sample of the KM-H2 cell line in this study (Figure 3_3).

This PCR method was repeated for the rest of the cell line and for the whole genome amplified KM-H2 and L-591, with all the TNFAIP3 gene primer sets until we obtained PCR product with appropriate band size.
Figure 3. Gel images of the PCR experiments. 1. A-F: exon 2 PCR products (462 bp) for HDLM-2, KM-H2, L-428, L-540, L-591, and L-1236 cell lines respectively; G,H: exon 7 and 9 PCR products for HDLM-2; NTC: non-template control; 2. L-591 genomic DNA PCR. Numbers 2, 3, 4/5, 6, 7.1, 7.2, 8, 9 represent the corresponding \textit{TNFAIP3} exon PCR products for L-591 genomic DNA with band sizes, 462, 306, 588, 311, 568, 653, 328, and 527 respectively (Supplementary table 2); 3. The PCR using the KM-H2 genomic DNA shows exon 3 to 6 are missing.

3.3) Whole genome amplification

Microdissected sections on the ‘PALM tubes’ caps were checked before and after the whole genome amplification procedure, to ensure the capture and lysis of the cells. Whole genome amplification of microdissected samples containing a single KM-H2 cell section, and 50 L-591 cells section was carried out using REPLI-g mini kit. Products of the WGA were run on a 2% agarose gel (Figure 4).

Figure 4. Images of a single KM-H2 cell, and 50 L-591 cells WGA products is showed. A: REPLI-g kit successfully amplified KM-H2 cell genome, NC: non template control; B: REPLI-g failed to amplify 50 L-591 cells. However, the gDNA (positive control) was amplified. NTC: non-template control.

REPLI-g kit was inconsistent in whole genome amplification process of other cell samples with a single cell and 50 cells (Figure 4), and often showing positive results for the non template control (NTC) in the reactions.
As an alternative, SIGMA’s GenomePlex Single Cell Whole Genome Amplification kit (WGA4) was used in the whole genome amplification of microdissected cell line (KM-H2 and L-591) and patient (1.5, 6.4, 7.21, and 7.8) samples (Figure 5) according to the manufacturer’s instructions. The amplified products were run on a 2% agarose gel. WGA4 kit amplified all the samples.

![Figure 5. KM-H2, L-591 and patient samples amplification using WGA4 kit. A: 10 cells and 50 cells of KM-H2; a single cell and 50 cells of L-59; PC: positive control; NC: non-template control; B: patient samples 1.5, 6.4, 7.21, and 7.8; PC: positive control; NTC: non-template control.]

### 3.4) Whole genome amplification – PCR

A test PCR with the WGA4 products of KM-H2 (10 cells, 50 cells) and L-591 (a single cell and 50 cells) microdissected samples was set up using the exon 8 primers. Only the PCR for WGA4 product of L-591 microdissected sample with 50 cells has worked (Figure 6_1).

To rule out the possibility, that, this PCR failure could have been due to the lower amount of initial DNA template concentration, another PCR test was set up with various template concentrations (0.5, 1.0, 2.0, 3.0, and 4.0 µL) of WGA4 products with the exon 8 primers. In this experiment WGA4 products of KM-H2 (10 cells, 50 cells) and 50 L-591 cells microdissection sample have worked. The L-591 single cell WGA4 product again failed to amplify this region (Figure 6_2).
Figure 6. Test PCR on WGA4 products of KM-H2 and L-591 samples. 1. A test PCR with the WGA4 products of KM-H2 (10 cells, 50 cells) and L-591 (a single cell and 50 cells) microdissected samples was set up using the exon 8 primers. NC (WGA): non-template control from WGA4 experiment; NC (PCR): non-template control for this PCR experiment; 2. A test PCR with various template concentrations (0.5, 1.0, 2.0, 3.0, and 4.0 µL) of KM-H2 (10 cells and 50 cells) and L-591 (a single cell and 50 cells) WGA4 products; NC (WGA): non-template control from WGA4 experiment; NC (PCR): non-template control for this PCR experiment.

PCR on WGA4 product of L-591 for all of the TNFAIP3 exons were successful (Figure7_1:A) and the PCR experiments on WGA4 product of KM-H2 yielded the same results as those of cell lines, such as the deletion in the region between intron 2 and exon 6 (Figure7_1:B).

WGA4 was used to amplify some Hodgkin’s lymphoma patient samples, namely, 1.5, 6.4, 7.21 and 7.8 (50 cells each). The amplified products were run on a 2% agarose gel (Figure 7_2). PCR experiments on patient samples 1.5, 7.4 and 7.8 were failed (Figure7_2:A). However, patient sample 6.4 successfully PCR amplified all the exons except exon 7.1(Figure7_2:B).

Figure 7. PCR on WGA4 products of KM-H2, L-591, and patient samples. 1. PCR on the WGA4 products of KM-H2 (50 cells) and L-591 (50 cells) microdissected samples was set up using all the exon primers. NC (WGA): non-template control from WGA4 experiment; NC (PCR): non-template control for this PCR experiment. 2. PCR on patient sample WGA4 product. A: WGA4 products for patient sample 1.5 and 7.21 have failed in the PCR; B: Patient sample 6.4 has picked up bands for all the exons except 7.1. NC: negative control; g-DNA: positive control.

3.5) Quality of DNA sequences from unamplified cell lines and amplified cell lines and patient samples

The quality of the sequences was assessed based on the percentage of the alignment (in Sequencher 5.0 software) between sequence traces obtained and the reference sequence. The quality of sequences for the unamplified cell lines was better (79.7%) than those of WGA sequences (76.5%) (Table1). Good quality sequences obtained for the rest of the unamplified cell lines (HDLM-2, L-428, L-540,
and L-1236) cell lines. Interestingly, the quality of the sequences obtained from the patient sample 6.4 were as good as the unamplified cell line samples (Table 2).

Table 1. Showing the quality of sequences for the unamplified and WGA KM-H2 and L-591

<table>
<thead>
<tr>
<th>PCR product (bp)</th>
<th>WGA (bp)</th>
<th>PASS (%)</th>
<th>UA (bp)</th>
<th>PASS (%)</th>
<th>WGA (bp)</th>
<th>PASS (%)</th>
<th>UA (bp)</th>
<th>PASS (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exon 2</td>
<td>462</td>
<td>473</td>
<td>63</td>
<td>1270</td>
<td>75</td>
<td>387</td>
<td>79.8</td>
<td>389</td>
</tr>
<tr>
<td>Exon 3</td>
<td>306</td>
<td>del</td>
<td>del</td>
<td>del</td>
<td>del</td>
<td>600</td>
<td>81.2</td>
<td>559</td>
</tr>
<tr>
<td>Exon 4-5</td>
<td>588</td>
<td>del</td>
<td>del</td>
<td>del</td>
<td>del</td>
<td>301</td>
<td>80.7</td>
<td>301</td>
</tr>
<tr>
<td>Exon 6</td>
<td>311</td>
<td>del</td>
<td>del</td>
<td>del</td>
<td>del</td>
<td>1099</td>
<td>82</td>
<td>543</td>
</tr>
<tr>
<td>Exon 7.1</td>
<td>568</td>
<td>561</td>
<td>83</td>
<td>1099</td>
<td>82</td>
<td>545</td>
<td>91.6</td>
<td>543</td>
</tr>
<tr>
<td>Exon 7.2</td>
<td>653</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>301</td>
<td>79.4</td>
<td>301</td>
</tr>
<tr>
<td>Exon 9</td>
<td>528</td>
<td>515</td>
<td>87.6</td>
<td>533</td>
<td>79.4</td>
<td>508</td>
<td>54.7</td>
<td>513</td>
</tr>
</tbody>
</table>

bp: base pairs; UA: unamplified; del: deletion; F: sequencing failure/poor quality sequences; a: exon7.1 IntR to exon 2; b: Int ex7; c: exon 3 IntF to exon 5
Table 2. Showing the quality of sequences for the unamplified and patient sample

<table>
<thead>
<tr>
<th>Exon</th>
<th>PCR product (bp)</th>
<th>L-428 UA (bp)</th>
<th>L-428 PASS (%)</th>
<th>L-540 UA (bp)</th>
<th>L-540 PASS (%)</th>
<th>L-1236 UA (bp)</th>
<th>L-1236 PASS (%)</th>
<th>HDLM-2 UA (bp)</th>
<th>HDLM-2 PASS (%)</th>
<th>Patient (6.4) sample UA (bp)</th>
<th>Patient (6.4) sample PASS (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exon 2</td>
<td>462</td>
<td>477</td>
<td>73</td>
<td>477</td>
<td>71.9</td>
<td>477</td>
<td>71.7</td>
<td>481</td>
<td>70.7</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>Exon 3</td>
<td>306</td>
<td>278</td>
<td>87</td>
<td>283</td>
<td>86.2</td>
<td>282</td>
<td>85</td>
<td>282</td>
<td>79.8</td>
<td>278</td>
<td>88.8</td>
</tr>
<tr>
<td>Exon 4-5</td>
<td>588</td>
<td>566</td>
<td>91.5</td>
<td>569</td>
<td>90.5</td>
<td>569</td>
<td>91.2</td>
<td>598</td>
<td>88.3</td>
<td>561</td>
<td>86.8</td>
</tr>
<tr>
<td>Exon 6</td>
<td>311</td>
<td>290</td>
<td>84</td>
<td>300</td>
<td>81.7</td>
<td>293</td>
<td>83.3</td>
<td>291</td>
<td>82</td>
<td>FP</td>
<td>FP</td>
</tr>
<tr>
<td>Exon 7.1</td>
<td>568</td>
<td>547</td>
<td>91.6</td>
<td>545</td>
<td>94</td>
<td>544</td>
<td>93.8</td>
<td>549</td>
<td>92</td>
<td>FP</td>
<td>FP</td>
</tr>
<tr>
<td>Exon 7.2</td>
<td>653</td>
<td>637</td>
<td>84</td>
<td>639</td>
<td>85</td>
<td>636</td>
<td>90.4</td>
<td>642</td>
<td>89.1</td>
<td>629</td>
<td>93.5</td>
</tr>
<tr>
<td>Exon 8</td>
<td>328</td>
<td>303</td>
<td>88.8</td>
<td>302</td>
<td>90.4</td>
<td>301</td>
<td>87</td>
<td>300</td>
<td>88.7</td>
<td>299</td>
<td>91</td>
</tr>
<tr>
<td>Exon 9</td>
<td>528</td>
<td>512</td>
<td>88.7</td>
<td>512</td>
<td>88.9</td>
<td>513</td>
<td>89.9</td>
<td>512</td>
<td>88.9</td>
<td>500</td>
<td>94.6</td>
</tr>
</tbody>
</table>

bp: base pairs; UA: unamplified; del: deletion; F: sequencing failure/poor quality sequences; FP: PCR failure

3.6) Comparison of WGA amplified and unamplified DNA sequence

To investigate whether microdissection and whole genome amplification methods will introduce any changes to the genome sequences, exon sequence of TNFAIP3 gene of WGA amplified and unamplified DNA from the KM-H2 and L-591 cell lines were compared. No sequence differences were found.

3.7) Analysis of the TNFAIP3 mutations in HL cell lines and patient sample

All of the seven nucleotide changes that were observed in this study were illustrated in the Table 3. Nucleotide T deletion was observed at the base position 8751 (intron 5-6) (gi| reference sequence) in L-540 and L-591 (both unamplified and WGA amplified) (Table 3).
Figure 8. Chromatograms showing nucleotide changes observed in the cell lines and patient samples. The chromatogram shows A. T446G transversion for the L-591 (unamplified) cell line, B. G491A transition for the L-1236 cell line, C. A1434T transversion for the patient 6.4; and D. C8751A transversion for the L-591 and L-540 cell lines.
Some of the nucleotide changes observed in this study match with the previous studies on the HL cell lines (16), such as the deletion of the region between intron2 and exon 6 in the KM-H2 cell line, and mutation in L-591 at coding base 491 (G491A) which has changed the amino acid Tryptophan to STOP codon. A1434T transversion and T3003A transitions were observed in the patient sample 6.4.

Mutations observed in L-1236 (G491A) and patient sample 6.4 (A1434T) were searched (http://siftdna.org) for potential implication on the protein function. The results show that both of these mutations are ‘novel’ with tolerated (A1434T) and damaging (G491A) effect. We could not obtain population frequency for any of the genetic changes observed in this study.

**Table 3.** Mutations found in cell lines and patient sample

<table>
<thead>
<tr>
<th>Sample</th>
<th>EBV</th>
<th>Subtype</th>
<th>Nucleotide change</th>
<th>Amino acid change</th>
</tr>
</thead>
<tbody>
<tr>
<td>KM-H2 (B)</td>
<td>−</td>
<td>NS</td>
<td>Δ intron 2–exon 6</td>
<td>Frameshift aa 99</td>
</tr>
<tr>
<td>L-540 (g)</td>
<td>−</td>
<td>NS</td>
<td>C8751A</td>
<td>Intron 5-6</td>
</tr>
<tr>
<td>L-1236 (g)</td>
<td>+</td>
<td>NS</td>
<td>G491A</td>
<td>W142STOP</td>
</tr>
<tr>
<td>L-591 (B)</td>
<td>+</td>
<td>NS</td>
<td>C8751A</td>
<td>Intron 5-6</td>
</tr>
<tr>
<td>Patient 6.4</td>
<td>−</td>
<td>−</td>
<td>A1434T</td>
<td>T463S</td>
</tr>
<tr>
<td>Patient 6.4</td>
<td>−</td>
<td>−</td>
<td>T3003A</td>
<td>3' UTR</td>
</tr>
</tbody>
</table>

**NS:** Nodular Sclerosis; **MC:** Mixed Cellular; **Δ:** deletion; **g:** unamplified; **B:** found in both amplified and unamplified; **a:** position on the coding sequence (22); **b:** corresponding to PDB accession no.NP_006381; **c:** corresponding to the genomic reference sequence, (NCBI Reference Sequence: NC_000006.11, gi|224589818:138188581-138204449 (21)).
4) Discussion

In this study, we initially used REPLI-g kit to whole genome amplify the cell line samples. However, due to inconsistencies in the amplification results and unspecific amplification of samples, we replaced it with the WGA4 kit. WGA4 kit amplified DNA of all the microdissected cell lines samples and patient samples. However, WGA products for patient samples 1.5, 7.28, and 7.8 failed to amplify all of the TNFAIP3 exons in the following PCR experiments. Patient sample 6.4 amplified all of these exons except the exon 6 and 7.1.

Better quality sequences were obtained for the cell line (>76.5%) and patient sample 6.4 (90.9%). WGA method gave good coverage for the TNFAIP3 gene in these samples. However, sequencing reactions for exon 3 (L-591) and exon 7.2 (KM-H2 and L-591) had consistently failed.

After analysing the mutation spectra for the whole genome ‘amplified’, and ‘unamplified’, KM-H2 and L-591 cell lines, we came to a conclusion that microdissection and whole genome amplification methods didn’t introduce any changes to the sequences. However, differences in the quality of the sequences obtained and the failure to generate sequences for all the exons for these amplified samples had hampered our investigation.

Our study for the first time revealed C8751A transversion in the TNFAIP3 gene for the L-540 cell line. The same mutation was observed in the L-591 cell lines in this study.

Our results were compared to a previous study (16) that looked into the mutation spectra for the TNFAIP3 gene in the same cell lines. We observed same nucleotide changes in the cell lines KM-H2 (deletion of intron 2–exon 6 region), and L-1236 (G491A). However, improvement in the quality of the sequence and the coverage for the region of interest in the WGA products will result in a better whole genome or exome sequencing of the target sample.

The functional role of the genetic changes observed in our study was investigated in http://siftdna.org. The results show that the mutations observed in L-1236 (G491A) and patient sample 6.4 were ‘novel’ with tolerated (A1434T) and damaging (G491A) effect. We could not obtain population frequency for any of the genetic changes observed in this study.

This study enabled us to amplify DNA from very few numbers of cells (~50 cells) which will be sufficient enough to use in the future exome/whole genome sequencing applications to understand the role of genetic alterations in Hodgkin’s lymphoma.
References


6. Genomics lab, University of Birmingham, Birmingham, UK. (http://www.genomics.bham.ac.uk/sequencing.htm).


17. Sequencher 5.0 (http://genecodes.com/).


23. COSMIC (Catalogue of Somatic Mutations in Cancer) (http://www.sanger.ac.uk/cosmic).


### 6) SUPPLEMENTARY NOTES

**Table 1.** Primer sequences used for genomic DNA amplification of TNFAIP3

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Sequence</th>
<th>Amplification of</th>
</tr>
</thead>
<tbody>
<tr>
<td>A20E2exF</td>
<td>TGCCTACAGATCACGGGTAATGACAAG</td>
<td>exon 2</td>
</tr>
<tr>
<td>A20E2NewexF</td>
<td>GGAGTCGTATTAAAGTCAGGCTAA</td>
<td>exon 2</td>
</tr>
<tr>
<td>A20E2NewexR</td>
<td>GCCAAAAGAAAAACACAACAGAAC</td>
<td>exon 2</td>
</tr>
<tr>
<td>A20E2intF</td>
<td>GTTTCTGCAGGCAGCTATAGAGG</td>
<td>exon 2</td>
</tr>
<tr>
<td>A20E2R</td>
<td>AGCTTCATGAATGGGGATCCAGCAG</td>
<td>exon 2</td>
</tr>
<tr>
<td>A20E3exF</td>
<td>ACCATTCGTCGCCCTAGAATAGCAG</td>
<td>exon 3</td>
</tr>
<tr>
<td>A20E3intF</td>
<td>ACCTTTGCTGGTCTTACATGCAG</td>
<td>exon 3</td>
</tr>
<tr>
<td>A20E3R</td>
<td>TATGCCCAACCATGGAGCTCTGTTAG</td>
<td>exon 3</td>
</tr>
<tr>
<td>A20E4-5exF</td>
<td>TGAATAATTGTAGAGTGATGTCAAGAC</td>
<td>exon 4/5</td>
</tr>
<tr>
<td>A20E4-5intF</td>
<td>TACAGGGAGTACAGGATACATCCAGC</td>
<td>exon 4/5</td>
</tr>
<tr>
<td>A20E4-5R</td>
<td>GGAATACCTGATGTTCTGAGTCTAG</td>
<td>exon 4/5</td>
</tr>
<tr>
<td>A20E6exR</td>
<td>AATCACTCTACTTTACTATGGCCTTG</td>
<td>exon 6</td>
</tr>
<tr>
<td>A20E6F</td>
<td>TGAGATCTACTTACTATGGCCTTG</td>
<td>exon 6</td>
</tr>
<tr>
<td>A20E6intR</td>
<td>TCAGGTGGCTGAGGTAAAGACAG</td>
<td>exon 6</td>
</tr>
<tr>
<td>A20E7.1exF</td>
<td>GGTCTTACATTTCATGGGCATCCAC</td>
<td>exon 7</td>
</tr>
<tr>
<td>A20E7.1intF</td>
<td>GAGCTAATGATGAAAAATCTTGTGTTG</td>
<td>exon 7</td>
</tr>
<tr>
<td>A20E7.1R</td>
<td>CAAATACGGTTGTGCTGACACATTCC</td>
<td>exon 7</td>
</tr>
<tr>
<td>A20E7.2exR</td>
<td>CAGTTCTGCGCTGACTGCTACATG</td>
<td>exon 7</td>
</tr>
<tr>
<td>A20E7.2F</td>
<td>CTCCTCGGGAGAAAGCCTATGAGC</td>
<td>exon 7</td>
</tr>
<tr>
<td>A20E7.2intR</td>
<td>GACAAAACCCCTCTGGACACGCAG</td>
<td>exon 7</td>
</tr>
<tr>
<td>A20E8exR</td>
<td>ATGAGGAGACAGAACCTGGGCAGAG</td>
<td>exon 8</td>
</tr>
<tr>
<td>A20E8F</td>
<td>ACTGTCAGCATCTCTGTATCGGTG</td>
<td>exon 8</td>
</tr>
<tr>
<td>A20E8intR</td>
<td>TGTCATGCTGCGTAAAGAACGCTC</td>
<td>exon 8</td>
</tr>
<tr>
<td>A20E9exF</td>
<td>GTAGACTCCACACTCTCCAATGAG</td>
<td>exon 9</td>
</tr>
<tr>
<td>A20E9intF</td>
<td>GTGCTCTCCTAAGAATGTGACG</td>
<td>exon 9</td>
</tr>
<tr>
<td>A20E9R</td>
<td>GGGTTACCAACCTGAGCATCGTG</td>
<td>exon 9</td>
</tr>
<tr>
<td>A20E9Renew</td>
<td>CGGGTTACCAACCTGAGCATCGTG</td>
<td>exon 9</td>
</tr>
</tbody>
</table>

**E2-E9:** exon2-exon9; **F:** forward primer; **R:** reverse primer.
Table 2. List of the primers used and their PCR products

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Position on gDNA</th>
<th>Product (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A20 E2intF-&gt;E2R</td>
<td>3690-&gt;4152</td>
<td>462</td>
</tr>
<tr>
<td>A20 E2NewF-&gt;E2NewR</td>
<td>3712-&gt;4110</td>
<td>398</td>
</tr>
<tr>
<td>A20 E3intF-&gt;E3R</td>
<td>7336-&gt;7642</td>
<td>306</td>
</tr>
<tr>
<td>A20 E4-5intF-&gt;E4-5R</td>
<td>8193-&gt;8781</td>
<td>588</td>
</tr>
<tr>
<td>A20 E6intF-&gt;E6R</td>
<td>9875-&gt;9914</td>
<td>311</td>
</tr>
<tr>
<td>A20 E7.1intF-&gt;E7.1R</td>
<td>10918-&gt;11486</td>
<td>568</td>
</tr>
<tr>
<td>A20 E7.2intF-&gt;E7.2R</td>
<td>11966-&gt;12012</td>
<td>653</td>
</tr>
<tr>
<td>A20 E8intF-&gt;E8R</td>
<td>12888-&gt;12925</td>
<td>328</td>
</tr>
<tr>
<td>A20 E9intF-&gt;E9R</td>
<td>13522-&gt;14049</td>
<td>527</td>
</tr>
<tr>
<td>A20 E9intF-&gt;E9Rnew</td>
<td>13522-&gt;14050</td>
<td>528</td>
</tr>
</tbody>
</table>

E2-E9: exon2-exon9; F: forward primer; R: reverse primer; gDNA: genomic DNA.

Table 3. Template requirements for sequencing at the Genomics lab

<table>
<thead>
<tr>
<th>Template DNA</th>
<th>DNA Quantity (ng)</th>
</tr>
</thead>
<tbody>
<tr>
<td>single-stranded DNA</td>
<td>25-50 (*)</td>
</tr>
<tr>
<td>double-stranded DNA</td>
<td>150-300 (*)</td>
</tr>
<tr>
<td>PCR : 100-200 bp</td>
<td>1.3 (***)</td>
</tr>
<tr>
<td>PCR : 200-500 bp</td>
<td>3-10</td>
</tr>
<tr>
<td>PCR : 500-1000 bp</td>
<td>5-20</td>
</tr>
<tr>
<td>PCR : 1000-2000 bp</td>
<td>10-40</td>
</tr>
<tr>
<td>PCR : &gt;2000 bp</td>
<td>20-50 (*)</td>
</tr>
</tbody>
</table>

(*) : for BigDye® Terminator versions 1.0, 2.0 and 3.0 higher quantities necessary

(**) : rule of thumb: PCR product length (in bp) / 50 = amount in ng.
Comparing and contrasting different approaches to identifying predictive biomarker combinations

BY

PRADEEP RAMAGIRI

This project is submitted in partial fulfilment of the requirements for the award of the MRes in Biomedical Research

School of Cancer Sciences
College of Medical and Dental Sciences
University of Birmingham
August 2012
Abstract

Adrenocortical carcinoma (ACC) is a heterogeneous malignancy with an incident rate of about 0.7-2.0 cases per million people per year. However, it accounts for more than 15% of the total adrenal incidentalomas registered every year. Distinguishing malignant ACC from benign Adrenocortical Adenoma (ACA) has been a major problem. Hence, in this study we attempt to statistically design a biomarker tool based on steroid metabolic excretion data from ACA and ACC patients. Data in question is of 32 distinct adrenal derived steroid secretions obtained from 102 ACA patients and 45 ACC patients. Variable selection was performed by forward selection, backward search, stepwise selection, and all subset combination methods using logistic regression. An investigation into prediction model building was carried out and an assessment was made on the model fitness using various criteria, such as log-likelihood, deviance, AIC, and Wald’s statistics. Sensitivity and specificity for some of the best models was evaluated using ROC curve. Some of the best prediction models from all subset combination models were chosen to inspect all the patient samples in the data. Models with 6, 7, and 8 variable subsets were shown to have best prediction capability with perfect sensitivity and specificity. However, investigation into these variable subsets revealed ‘overfitting’. Models with 4 and 5-variable subset combinations have relatively high prediction capability with a significant sensitivity and specificity (AUC=99.3% and AUC=99.8%, respectively). The strength of the estimates in these models was relatively stable when compared to 6-8 variable subset models.
Acknowledgement

I sincerely thank Prof. Jon Deeks and Dr. Wenbin Wei for their kind support and guidance.
# Table of contents

1) **Introduction** ..................................................................................................................... 7  
   1.1) Evidence based medicine ................................................................................................. 7  
   1.2) Statistical modelling for prediction .................................................................................. 8  
   1.3) Logistic regression .......................................................................................................... 9  
   1.4) Building prediction models ........................................................................................... 10  
      1.4.1) Coding ......................................................................................................................... 10  
      1.4.2) Model selection .......................................................................................................... 10  
   1.5) Challenges involved in statistical prediction modelling ................................................... 11  
   1.6) Aim of the study ............................................................................................................ 12  

2) **Material and Methods** .................................................................................................. 13  
   2.1) Patient data ....................................................................................................................... 13  
   2.2) Coding ............................................................................................................................... 13  
   2.3) The Logistic regression model for binary responses ....................................................... 14  
   2.4) Data was processed using R language ............................................................................. 14  
   2.5) SPSS .................................................................................................................................. 14  
   2.6) Stata ................................................................................................................................... 14  
   2.7) Bluebear ........................................................................................................................... 14  
   2.8) Building prediction models ............................................................................................. 15  
      2.8.1) Variable selection ....................................................................................................... 15  
         a) Forward selection .......................................................................................................... 15  
         b) Backward elimination ..................................................................................................... 16  
         c) Stepwise regression ......................................................................................................... 16  
         d) All subset combinations ............................................................................................... 16  
   2.9) Criteria for model fitting ............................................................................................... 17  
      2.9.1) Log likelihood ratio ................................................................................................. 17  

2.9.2) Akaike Information Criteria (AIC) ................................................................. 17
2.9.3) Assessing the contribution of predictors (Wald’s statistics) ......................... 18
2.10) Area under the curve (AUC) ............................................................................. 18
2.11) Correlation test ............................................................................................... 18

3) Results ................................................................................................................... 19

3.1) Data analysis ...................................................................................................... 19
3.2) Coefficients of the predictive variables .............................................................. 21
  3.2.1) Raw data .................................................................................................... 21
  3.2.2) Log form of changed data ...................................................................... 22
  3.2.3) Comparing the coefficients of raw and log transformed data ............... 22
3.3) Model fitting ..................................................................................................... 24
  3.3.1) Forward selection .................................................................................... 24
  3.3.2) Backward search and bidirectional search ............................................. 24
  3.3.3) All subset combinations (from best data set) ......................................... 26
3.4) Area under the ROC curve .............................................................................. 27
3.5) Correlation among the best predictive variables ............................................ 29
  3.5.1) Best predictor variables from the subset models .................................... 29
  3.5.2) Correlation ............................................................................................... 30
  3.5.3) Inspection of predictor variables from each subset ................................. 31

4) Discussion ........................................................................................................... 33

References ............................................................................................................... 35

Supplementary material ......................................................................................... 37
Table of figures

Figure 1. Inter quartile range (IQR) for predictor variables in ACC and ACA patients...........21

Figure 2. Area under the ROC curve, values for 2-variable (A) through to 7-variable (G) subset models (as mentioned in the Section 3.3.3) and forward selection model (H)..........28

Figure 3. Frequencies of the variables in 2-8 variable subset models were shown here........29

Figure 4. Scatter plot depicting the model fit for each patient........................................31
1) Introduction

Adrenal tumours are one of the most prevalent forms of cancer with an incident rate of 2% in general population and about 7% in the older generation (70 years or above age groups). Incidental discovery of Adrenal tumour (Adrenal incidentalomas) during computed tomography (CT) and autopsy resulted in a rapid increase in their prevalence rate (1, 8, and 12). Adrenal tumours cause serious health problems in patients, such as Cushing’s syndrome, Conn’s syndrome (hyperaldosteronism), virilisation in females, feminisation in males, and multiple endocrine neoplasia (MEN) (1, 3, 8, 12-13).

Adrenal tumours can be divided into two groups based on their origin, ‘Adrenocortical adenoma’ (ACA) and ‘Adrenocortical carcinoma’ (ACC) of Adrenal Cortex origin, and ‘Neuroblastoma’ and ‘Pheochromocytoma’ of Adrenal Medulla origin. The majority of adrenal tumours are benign, but still cause some serious health problems. Only 10% of the Pheochromocytomas and all ACC are malignant. ACC is an aggressive form of cancer which can occur in all age groups (1, 8, 12).

Most often ACC are not diagnosed, until they have grown very large and/or metastasize to other organs, due to their location deep in the retro peritoneum. Surgical removal of ACC is carried out as part of the main treatment but is not feasible with many patients, due to complications associated with the hormonal imbalances after surgical removal of the tumour. It is very difficult to distinguish ACC from ACA based on histopathological and biochemical methods. Some of the imaging techniques, such as $^{18}$F Fluorodeoxyglucose positron emission tomography, offer better diagnostic capability with higher specificity (91%) and sensitivity (97%) (13). But this technology is very expensive and sparsely available. Hence, it is pivotal to differentiate ACC from ACA by some other means.

1.1) Evidence based medicine

Traditionally, medicine has been very much subjective. This trend has been changing in recent years, where ‘evidence based’ medicine is gaining popularity among physicians and health policy makers. Evidence based medicine can be best explained as ‘a conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients’. Clinical prediction models may provide the evidence based input by providing estimates of the individual probabilities of risks and benefits. Clinical prediction models consider a number of characteristics related to the disease, patient and the treatment (7).
Prediction models may help design preventive interventions for patients with high risk of having or developing a disease. From a clinical point of view, prediction models may provide valuable information to patients and their physicians on the probability of a diagnosis or a prognostic outcome.

### 1.2) Statistical modelling for prediction

In data analysis, statistical models summarise patterns in the data. Prediction mainly estimates risks of disease development and survival rates for the patients. Prediction also tests hypotheses, such as the importance of certain predictor variables in a disease and the correlation between variables (14).

Statistical model for Prediction can be distinguished into 3 main classes: regression, classification, and neural networks according to Ewout W. Steyerberg (10). However, majority of other statisticians discern the prediction modelling methods into two classes: Regression analysis (for continuous variables) and classification (for non-continuouos variables). Regression analysis investigates the relationship between a ‘dependent/outcome’ variable and one or more ‘independent/predictor’ variable(s) by employing several techniques for modelling and analysis. Dependent/outcome variable can be defined as the observable result, such as the presence or absence of cancer, when any one of the independent variable is manipulated (5). Independent variables are usually the measurements of an effect, such as response to treatment/drug, and effect of temperature on durability of materials. Regression helps understand how the value of a dependent variable changes with changes to any one of the independent variables (10).

Classification method identifies a set of categories/sub-populations to which a new observation belongs, based on ‘training set’ of data containing observations whose category membership is known. The individual observations are analysed into a set of quantifiable properties, known as explanatory variables, features etc. These properties may be of categorical, ordinal, integer valued, or real valued. Neural networks can be considered as the non-linear extensions of linear logistic models. The generalised non-linear models (GNLM) are implemented as neural networks. In this method, outcome is related to the non-linear combinations of the predictor variables, in contrast to the methods where the outcome is related to simple linear combination of estimated regression coefficients and predictor values (14).

For binary outcome such as the diagnostic outcome (presence or absence of a disease) there are several methods available for prediction modelling. Logistic regression facilitates as a quite flexible model to derive predictions. In this approach, interactions and nonlinearity can be incorporated. Other models such as Naive Bayes can be seen as a simplified version of logistic regression, which ignores
correlations between predictors. The choice of the model depends on various factors, such as the sample size and research question(s) (14).

This study used logistic regression models to distinguish ACC from ACA patients in the data.

1.3) Logistic regression

Logistic regression is a flexible method as it can incorporate categorical and continuous variables, non-linear transformations, and interaction terms. Logistic regression is a form of ‘generalised linear models’. Similar to linear regression model, the binary outcome ‘Y’ is linked to a linear combination of a set of predictors and regression coefficients ‘β’ (4). Logistic ‘link’ function is used to restrict predictions between ‘0’ and ‘1’.

Logistic regression works by relating the log odds (logit) to a linear combination of the explanatory variables. Logistic regression was applied to predict the transformation of the outcome (otherwise we could predict unrealistic values of the probability (i.e. outside 0 to 1 range). Log odds value lies between minus infinity and plus infinitive. Logistic regression can be interpreted using odds and odds ratio as below,

\[
\log\text{it}(P) = \ln\left(\frac{P}{1-P}\right)
\]

Therefore, \( \log_e (\text{odds}) = a+bx \)

Where, 'a' is an intercept where predictor variable \( x=0 \), 'b' regression coefficient that relate to 'x', and 'e' is the base of natural logarithm.

In the case of one predictor variable 'P' the logistic regression equation from which the probability of outcome is predicted is given by:

\[
P = \frac{e^{a+bx}}{1 + e^{a+bx}}
\]

In logistic regression, estimates of the coefficients are obtained in an iterative procedure known as 'maximum-likelihood'.
Maximum likelihood is generally used to estimate $\beta_i$ regression coefficients in a standard logistic regression model. $\beta_i$ regression coefficients can be interpreted as the effect of a 1-unit increase in $x_i$ while keeping the other predictors in the model constant. In logistic regression model, $\beta_i$ is unadjusted or univariate for a single predictor, but with multiple predictors, $\beta_i$ is adjusted or conditional on the values of other predictors (14).

1.4) **Building prediction models**

In the development of prediction models some important considerations have to be made. The very first and the most important one is the ‘research question’. Some of the other important questions that need to be addressed are: What do we know about the predictors? How are the predictors in the study defined? What is the outcome? How to deal with the missing values? The other important considerations involve coding of the predictors, specifying a model, ‘overfitting’, and validation of the model (14).

1.4.1) Coding

As in many cases the raw data is often not in an appropriate form for entering in the regression models. Hence, they need to be manipulated. This process is called ‘coding’. Data analysis usually starts with obtaining an impression of the data. This includes finding the missing values and the distribution of the predictors. Descriptive analyses, such as frequency distribution and Inter-quartile (IQ) range are quite useful to find the distribution of predictor variables in the data (14).

1.4.2) Model selection

Selection of the predictor variables can also be made based on the subject knowledge. The list of the number of variables in the study can be reduced by a literature review. Variables can be selected based on their distributions. Variables with relative importance can be deleted if they have a large number of missing values (14).

Variables are often strongly correlated with each other. This ‘collinearity’ is one of the major ‘variance inflation factors’, which degrades the precision of estimate coefficients (11). The number of variables in a model must not depend only on the statistical significance but also on the number of questions it will answer. A parsimonious model is not necessarily a better model, as it might not answer all the biological research questions. However, it can be the case when pre-specified
models were compared. A model with smaller number of variables is easier to interpret and practice.

Validation of the prediction models on new subjects, outside the subjects on which they were built, is the main aim of outcome prediction. A major threat to this validation is ‘Overfitting’, where predictions are not valid for the new subjects even though the data under study are well described. These models might have been built by capitalising on the specifics and idiosyncrasies of the sample (9). ‘Overfitting’, can also be defined as the ‘curse of dimensionality’, or as fitting a statistical model that has too many degrees of freedom (10). As a result of overfitting, too ‘optimistic’ impression of a model will be achieved in new subjects from the underlying population. Optimism = True performance (underlying population) - Apparent performance (estimated performance in the sample) (14).

1.5) Challenges involved in statistical prediction modelling

Statistical prediction modelling faces various challenges, such as model uncertainty, and limited sample size.

Model uncertainty arises when a model is not fully pre-specified, before fitting it to a data set. Iterative model checking and model modifications are often followed in this case. On the other hand, standard statistical methods assume that a model was pre-specified. Hence, parameter estimates such as regression coefficients, their corresponding standard errors, 95% confidence intervals, and p-values are largely unbiased. When structure of the model was at least partly based on findings in the data, bias may occur, and usually the uncertainty of conclusions drawn from the model is underestimated (2, 6).

A large sample size is necessary to address most of the scientific questions with empirical data. In the majority of the prediction studies ‘effective sample size’ is much smaller than indicated by the total number of subjects in a study (9). For example, a study containing 1000 patients with an event incidence rate of 1% will have an effective sample size of 10. We may not be able to derive a reliable prediction model from a small sample size. A large sample size is desirable to study several aspects of prediction modelling, such as gene-disease associations and multivariable prognostic modelling. A small size renders in making strong modelling assumptions. With a small samples size we may have to assume the linearity of the predictor variable and no/less correlation between the predictor variables. Hence, we may have a limited power to test the deviations from these model assumptions. Sample size may restrict our scientific approach and may dictate on what we can achieve as more complex questions, such as ‘what are the most important predictors in the model?’ can be solved with a large sample data (14, 15).
1.6) Aim of the study

This study was aimed at designing prediction models by using the differences in the steroid and steroid metabolite excretion levels between the ACC and ACA patients. In this study we investigated and compared different methods for selecting variables. We also aimed to investigate statistical significance of the estimates in the prediction models and examine the estimates for these models on all of the patients in the data.
2) Material and Methods

2.1) Patient data

The data used in this study was provided by Wiebke Arlt group from ‘Centre of Endocrinology, Diabetes, and Metabolism, University of Birmingham, Birmingham, UK’. The data contains urinary steroid metabolomes, analysed in a single run by gas chromatography/mass spectrometry (GC/MS), of 24 hour urine samples from adrenal tumour patients. These samples were collected, with informed consent, in six specialist referral centres participating in the European Network for the study of Adrenal Tumours (ENS@T; www.ensat.org) between 2006 and 2009. Appropriate ethical approval was obtained from the local ethical boards. These samples were identified as either benign ACA or malignant ACC with the help of histological, biochemical, clinical examination as well as imaging.

Wiebke Arlt group have used the same data to design a biomarker tool by applying ‘Generalized Matrix Learning Vector Quantization (GMLVQ)’ method to differentiate ACC from ACA. They have published their research findings in ‘Journal of Clinical Endocrinol Metab’, in 2011(17). This group have also used logistic regression method, but concluded GMLVQ was a better method for building prediction models with this data set. In this study, I have made an attempt to learn and apply logistic regression method to build alternative prediction models.

The data in the present study contains 32 predictor variables. These variables were grouped into Androgen metabolites, Androgen precursor metabolites, Mineralocorticoid metabolites, Mineralocorticoid precursor metabolites, Glucocorticoid precursor metabolites, and Glucocorticoid metabolites, based on their biochemical nature. The detailed description of these 32 variables is available in the Supplementary Table 2.

2.2) Coding

The data taken from the Wiebke Arlt group’s study (17) was labelled as ‘raw data’ and the predictor variables in the raw data were labelled from P1 to P32. Data was manipulated by replacing any ‘0’ values with ‘0.5’, so that when log transformed, the patient measurements with ‘0’ values were not denoted as ‘NA’. Log transformation expands the smaller values and squeezes the bigger values thus making the distribution more symmetric. Predictor variables in the log-transformed changed data were labelled from s1 to s32. The mean values observed for the log-transformed
changed data were less than that of the log transformed raw data. Best variables from both forms of the data set were taken to build prediction models.

2.3) The Logistic regression model for binary responses

A logistic regression model helps us to predict the probability of a particular outcome in relation to a list of predictor (independent) variables. In our study, the outcome has two categories. The predicted value lies between ‘0’ and ‘1’ (in this study, ‘0’ represents the presence of Adrenocortical adenoma (ACA) and 1 represents presence of Adrenocortical carcinoma (ACC). Hence, Binary logistic regression was applied. The outcome variable will be the proportion (probability) of individuals with the characteristic (i.e., ACC or ACA). This approach is a form of ‘discriminant’ analysis.

2.4) Data was processed using R language

The majority of statistical problems in this study were performed using R language on R platform (R version 2.13.1 and R Commander version 1.7-3). All of the functions run in R were obtained from the ‘Comprehensive R Archive Network (CRAN) (16). The scripts that were used to perform the statistical analysis were detailed in the Supplementary Table3.

2.5) SPSS

IBM SPSS software was used to generate prediction models using ‘stepwise forward selection’ method and ‘stepwise backward search’ method.

2.6) Stata

Stata V11.0 was used to generate ROC curves using the ‘lroc’ command following fitting the logistic models.

2.7) Bluebear

‘Bluebear’ cluster is a parallel computing service provided by the ‘IT services’ at the University of Birmingham, Birmingham, UK. It consists of 1500 processing cores and approximately 150 terabytes of user disk space with a sophisticated cluster computing system. Computationally intensive, all subset combinations method was performed using logistic regression on Bluebear platform. This method generated
subset combinations that had significant values for assessing measure of fit, such as 'deviance' and 'AIC'.

**2.8) Building prediction models**

In the previous sections, I have estimated and interpreted the coefficients in the logistic regression model. To build a 'best prediction model' for distinguishing ACA from ACC using the dataset, I needed to devise a strategy to a) select predictor variables for the model, and b) employ a set of methods to assess 'fit' of the models by analysing the estimates in the prediction models.

**2.8.1) Variable selection**

One of the aims in this study is to build the most parsimonious prediction model that still can describe the data. Prediction models with less number of variables can be numerically more stable and more easily generalized. Addition of more variables to the prediction model will increase the estimated standard errors and as a result the prediction model becomes dependent on the observed data. This approach can lead to 'over fitting' of the prediction model, where idiosyncrasies in the data are fitted rather than the 'generalizable' patterns. As a result, this model is not applicable to new patients.

It is important to include the relevant variables in the prediction model, to gain as much control of confounding as possible within the given data set. But, care must be taken for not 'over fitting' the model. This can be very real in our case as the number of variables is relatively large compared to that of patients.

In this study, potentially important variables were selected based on their estimated coefficients, estimated standard error, the likelihood ratio test for significance of the coefficient, and Wald statistic. There are various techniques available for choosing variables for a regression model. One of the popular methods that can be used to build prediction models is generalized linear regression, which employs 'forward selection', 'backward elimination', 'stepwise regression ', and 'best subsets selection'(Supplementary Table3).

a) **Forward selection**

Forward selection is one of the simplest data-driven prediction models building approach, where variables are added one at a time. The ‘significance’ of the variable is a measure of the statistical significance of the coefficients for that variable. In logistic regression, the significance is assessed with the help of likelihood ratio chi-square test. Hence, at any single step in the procedure, the most 'significant' variable
will produce the greatest change in the log-likelihood as compared to a model not containing the variable. The ‘significance’ of the variables is usually tested based on a pre-set P-value. The conventional P-value set up is at 0.05, but can choose a customary P-value level at say 0.10 or 0.15, to explore the nature of this method (4).

The model building starts with adding the most ‘significant’ variable to an empty model. Then at each step, each variable that is not already in the model is tested for the inclusion in the model. The most ‘significant’ of these variables is added to the model, so long as its P-value is below the pre-set level. This method is iterated until none of remaining variables are ‘significant’ when added to the model. This multiple use of hypothesis testing nature can lead to higher ‘type 1 error rate’ for a variable, where an ‘unnecessary’ chance is given for a variable inclusion. Because of this error prone nature, forward selection is usually a very good exploratory method (4). In forward selection, each addition of a variable to a model can also render one or more of the variables in the model ‘insignificant’ due to their ‘correlation’ with each other.

b) Backward elimination

In this method, an initial screening of the variables for their ‘significance’ is done based on the pre-set P-value. Then a model is fitted with all the variables of interest. Then at each step, the least ‘significant’ variable is excluded based on the pre-set P-value. This procedure is repeated until we end up with a model containing all the ‘significant’ variables.

But, this model building approach can omit a variable in the initial steps which can be a very ‘significant’ variable to include in the final reduced model. This flaw can be avoided by using the ‘stepwise regression’ method.

c) Stepwise regression

Stepwise regression method is a bidirectional selection method, which uses both forward selection and backward elimination. The model either starts with an empty model or a full model and the predictor variables are selected either for inclusion or exclusion from the model in each step based entirely on statistical criteria. Missing values restrict the number of available cases to this method, especially, when a full model is employed.

d) All subset combinations

An alternate method is to fit all possible regression models, and to evaluate these models according to some criterion, such as deviance or AIC. In this method a
number of best regression models can be selected. However, the fitting of all possible regression models is very computer intensive and time consuming.

2.9) **Criteria for model fitting**

Log-likelihood, Deviance, AIC, and Wald’s statistics were used to assess the fit of the prediction model.

2.9.1) **Log likelihood ratio**

The log likelihood ratio (LLR) corresponds to a difference in log likelihoods. LLR is maximum at the maximum-likelihood estimate (MLE) and equals zero. MLE is the value that corresponds to the largest possible likelihood of the event ($\pi$)

$$\text{Log (LR)} = L(\pi) - L(\text{MLE})$$

Log-likelihood indicates how much unexplained information is there after the model has been fitted. The deviance statistic -2LL is a goodness of fit indicator. Large values of log-likelihood indicates poor fitting. Log-likelihood can be used to compare the state of logistic regression model against baseline state. A baseline state is when there is only the constant in the prediction model.

$$\chi^2 = -2 LL_R - (-2 LL_F) = -2 \ln \left( \frac{\text{likelihood}_R}{\text{likelihood}_F} \right)$$

Where, ‘R’ is simple model and ‘F’ is a complex model (usually has an extra variable). The parameters in the simple model must be a proper subset of the parameters in the complex model. This model follows $\chi^2$ distribution with k degrees of freedom where k is the number of predictors in the new model. $\chi^2$ was used to test whether addition of a variable will reduces the goodness-of-fit measure.

2.9.2) **Akaike Information Criteria (AIC)**

Akaike Information Criteria (AIC) is used to measure the relative ‘goodness of fit’ of a prediction model. This criteria work on the basis of trade off between ‘accuracy’ and ‘complexity’ of the model.
The general equation for AIC can be given as below,

$$AIC = 2k - 2\ln(L)$$

Where, ‘k’ is the number of parameters in the model and ‘L’ is the maximum value of likelihood function for the estimated model.

The model with the least AIC value is considered as a better model. AIC includes a penalty for increasing the number of estimated parameters. This would discourage ‘over fitting’ (increase in the parameters improves goodness of fit, regardless of the parameters used in generating the data).

2.9.3) Assessing the contribution of predictors (Wald’s statistics)

The contribution of a predictor can be assessed by Wald statistics (z). Wald statistics is defined as the coefficient of the predictor (b) divided by the standard error (SE) of the coefficient, b/SE(b).

2.10) Area under the curve (AUC)

Area under the curve, otherwise known as the area under the receiver operating characteristic (ROC) curve, is one of the traditional measures for binary and survival outcomes. It’s a visualisation method for discriminant analysis. ROC curve plots the sensitivity (true positive rate) against the 1-specificity (false positive rate) for consecutive cut-offs for the probability of an outcome. A model with AUC=1 is a perfect model with greater specificity and sensitivity. AUC=0.5 is as good as guessing (15).

2.11) Correlation test

Correlation is a method of analysis for association between two continuous variables. The degree of ‘association’ is measured by ‘correlation coefficient (r)’.

Pearson product-moment correlation coefficient

This is a standard method of measuring the association between variables. This method leads to a quantity called ‘r’ which can take any value from -1 to +1, r=0 means no correlation, and r=1 means perfect correlation. This correlation coefficient ‘r’ measures the degree of ‘straight line’ association between the values of the two variables. Pearson’s correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations.
3) Results

3.1) Data analysis

Initial interrogation of the data revealed that it is positively skewed with some missing values (total 57, all of them in ACC patients), 0 (total 13, mostly in ACC patients), 1 (total 16, mostly in the ACA patients) with a highest measured value of 812807 (for variable P3 in patient 10) (Table1). We employed logistic regression method in our study which makes no ‘assumptions’ on the distribution of the data.

Table1. Raw data containing ‘zeroes’ and ‘missing values’

<table>
<thead>
<tr>
<th>Variable</th>
<th>Missing</th>
<th>‘Zero’ value</th>
</tr>
</thead>
<tbody>
<tr>
<td>P3</td>
<td>1(10)</td>
<td>1(139)</td>
</tr>
<tr>
<td>P5</td>
<td>1(1)</td>
<td></td>
</tr>
<tr>
<td>P6</td>
<td>2(11,15)</td>
<td>1(30)</td>
</tr>
<tr>
<td>P9</td>
<td>10(1-4,10-15)</td>
<td>4(45,48,59,73)</td>
</tr>
<tr>
<td>P10</td>
<td>1(30)</td>
<td></td>
</tr>
<tr>
<td>P11</td>
<td>11(1-4,10-16)</td>
<td>2(34-35)</td>
</tr>
<tr>
<td>P12</td>
<td>7(1-4,10,11,15)</td>
<td>1(34)</td>
</tr>
<tr>
<td>P13</td>
<td>10(1-4,10-15)</td>
<td></td>
</tr>
<tr>
<td>P15</td>
<td>11(1-4,10-16)</td>
<td>2(24,34)</td>
</tr>
<tr>
<td>P18</td>
<td>3(12,13,14)</td>
<td>1(96)</td>
</tr>
<tr>
<td>P21</td>
<td>10(1-4,10-15)</td>
<td>1(34)</td>
</tr>
<tr>
<td>P32</td>
<td>1(1)</td>
<td></td>
</tr>
</tbody>
</table>

Predictor variables in the raw data with missing values and zeroes (the corresponding patient number in parentheses).

To study the distribution of the predictor variables in the data, the median and quartiles were calculated (Table2). The median and IQR values for these predictor variables were found to be larger in the ACC patients. Figure1 shows the difference in distribution between the ACC and ACA patients. IQR values for ACC were found to be significantly larger than in ACA. This will further strengthen our theory, that steroid and steroid metabolite excretions vary significantly between ACC and ACA patients.
Table 2. IQR and median values for the predictor variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Common name</th>
<th>MEDIAN</th>
<th></th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>An</td>
<td>1130</td>
<td>632</td>
<td>1898</td>
</tr>
<tr>
<td>P2</td>
<td>Etio</td>
<td>3671</td>
<td>803.5</td>
<td>6029</td>
</tr>
<tr>
<td>P3</td>
<td>DHEA</td>
<td>612</td>
<td>58</td>
<td>15725</td>
</tr>
<tr>
<td>P4</td>
<td>16α-OHDHEA</td>
<td>653</td>
<td>201.5</td>
<td>2566</td>
</tr>
<tr>
<td>P5</td>
<td>5-PT</td>
<td>1900.5</td>
<td>121</td>
<td>7292.5</td>
</tr>
<tr>
<td>P6</td>
<td>5-PD</td>
<td>3412</td>
<td>257</td>
<td>13460.5</td>
</tr>
<tr>
<td>P7</td>
<td>THA</td>
<td>112</td>
<td>93.5</td>
<td>137</td>
</tr>
<tr>
<td>P8</td>
<td>5α-THA</td>
<td>76</td>
<td>88</td>
<td>114</td>
</tr>
<tr>
<td>P9</td>
<td>THB</td>
<td>147</td>
<td>105</td>
<td>308.5</td>
</tr>
<tr>
<td>P10</td>
<td>5α-THB</td>
<td>155</td>
<td>221.5</td>
<td>275</td>
</tr>
<tr>
<td>P11</td>
<td>3α5β-THALDO</td>
<td>24</td>
<td>22</td>
<td>36.25</td>
</tr>
<tr>
<td>P12</td>
<td>THDOC</td>
<td>102.5</td>
<td>15.5</td>
<td>184.75</td>
</tr>
<tr>
<td>P13</td>
<td>5α-THDOC</td>
<td>22</td>
<td>4</td>
<td>53.5</td>
</tr>
<tr>
<td>P14</td>
<td>PD</td>
<td>839</td>
<td>137.5</td>
<td>2262</td>
</tr>
<tr>
<td>P15</td>
<td>3α5α-17HP</td>
<td>18.5</td>
<td>9.5</td>
<td>31.25</td>
</tr>
<tr>
<td>P16</td>
<td>17HP</td>
<td>511</td>
<td>120</td>
<td>658</td>
</tr>
<tr>
<td>P17</td>
<td>PT</td>
<td>1484</td>
<td>372.5</td>
<td>3215</td>
</tr>
<tr>
<td>P18</td>
<td>PTONE</td>
<td>32</td>
<td>18.5</td>
<td>86.25</td>
</tr>
<tr>
<td>P19</td>
<td>TSH</td>
<td>2151</td>
<td>122</td>
<td>4346</td>
</tr>
<tr>
<td>P20</td>
<td>F</td>
<td>245</td>
<td>85</td>
<td>534</td>
</tr>
<tr>
<td>P21</td>
<td>6β-OH-F</td>
<td>356</td>
<td>133</td>
<td>1176</td>
</tr>
<tr>
<td>P22</td>
<td>THF</td>
<td>2836</td>
<td>1811</td>
<td>3791</td>
</tr>
<tr>
<td>P23</td>
<td>5α-THF</td>
<td>852</td>
<td>1264.5</td>
<td>1174</td>
</tr>
<tr>
<td>P24</td>
<td>α-cortol</td>
<td>557</td>
<td>355.5</td>
<td>1289</td>
</tr>
<tr>
<td>P25</td>
<td>β-cortol</td>
<td>740</td>
<td>536</td>
<td>1018</td>
</tr>
<tr>
<td>P26</td>
<td>11β-OH-An</td>
<td>653</td>
<td>552.5</td>
<td>1520</td>
</tr>
<tr>
<td>P27</td>
<td>11β-OH-Et</td>
<td>366</td>
<td>265</td>
<td>1479</td>
</tr>
<tr>
<td>P28</td>
<td>E</td>
<td>164</td>
<td>126</td>
<td>250</td>
</tr>
<tr>
<td>P29</td>
<td>THE</td>
<td>3701</td>
<td>3478.5</td>
<td>4638</td>
</tr>
<tr>
<td>P30</td>
<td>α-cortolone</td>
<td>1840</td>
<td>1340</td>
<td>1899</td>
</tr>
<tr>
<td>P31</td>
<td>β-cortolone</td>
<td>677</td>
<td>665.5</td>
<td>813</td>
</tr>
<tr>
<td>P32</td>
<td>11-oxo-Et</td>
<td>483.5</td>
<td>401</td>
<td>1623.25</td>
</tr>
</tbody>
</table>

Common names of the variables in the study and their distribution in ACC and ACA patients are shown here. Inter quartile range (IQR) in ACC patients are showing higher values.

Steroid profiling of the samples revealed the differences between ACA and ACC patients (Table 2, Figure 1). Higher excretion of active androgens (An and Etio) and androgen precursor metabolites (5-PD, DHEA, and 16α-OHDHEA) were observed in ACC. In mineralocorticoid precursor metabolites, only THDOC and 5α-THDOC show elevated levels of excretion in ACC. ACC also showed increased levels of Glucocorticoid precursor metabolites (PD, PT, and 17-HP), and Glucocorticoid metabolites (TSH, THE, 6β-OH-F, and α-cortol) excretion when compared to ACA (Table 2, Figure 1).
Figure 1. Inter quartile range (IQR) for predictor variables in ACC and ACA patients. The Blue and Red columns represent ACC and ACA patients in the study respectively.

Analysis of the sum of steroid subclass metabolites had revealed elevated levels of glucocorticoid metabolites excretion in 82% of ACC, compared to only 29% of ACA. An excess of androgen precursor, either combined with elevated levels of glucocorticoid precursors (36%) or both glucocorticoid and mineralocorticoid precursor (33%) was seen in the majority (67%) of ACC patients. In 9% of the ACC and 1% ACA patients, isolated glucocorticoid precursor excess was observed. In 2% of ACC and ACA patients elevated levels of isolated androgen precursor metabolites was observed. No excessive levels of isolated androgen metabolite were found.

There was no effect of age, sex, tumour size, or presence of metastasis on the level of metabolic excretion observed, in these ACC and ACA patients (Weibke Arlt et al).

3.2) Coefficients of the predictive variables

3.2.1) Raw data

Coefficients of the predictive variables in the raw data were analysed using R (Table3). Estimations for coefficients were made on both raw data and log transformed changed data based on log-likelihood ratio, Akaike Information Criteria (AIC), Deviance, p-value and Wald’s statistics. Based on log-likelihood ratio of the variables, P19, P5, P6, P13, P2, P21, P12, P17, P14, and P3 have significant predictive capabilities. However, the order of significance based on the p-value and deviance was P5, P6, P19, P2, P13, P17, P14, P3, P16, and P4 (Table3).
Based on AIC criteria, the order of significance obtained was P5, P6, P19, P13, P2, P21, P12, P17, P14 and P3. These four selection methods yielded the same five highly significant predictive variables at the top of the table. However, Wald’s statistics \([\text{Pr}(>|z|)]\) yielded P2, P5, P13, P16, P17, P6, P12, P14, P24, and P19 variables with significant predictive capability (Table3). The differences in the output can be due to the differences in the way they trade off statistical significance with the number of parameters (as in case of AIC).

3.2.2) Log form of changed data

Raw data was manipulated by replacing ‘0’ value with ‘0.5’ so that when log transformed the patient measurements with ‘0’ values were not denoted as ‘NA’. The analysis of log transformed changed data revealed that the most significant variables for log-likelihood, deviance, p-value, and AIC were the same, namely, s19, s5, s14, s6, s16, s17, s2, s3, s4, and s20 in the order of most significance first. However, for Wald’s statistics the most significant variables, although identical to the other methods, were yielded in a different order: s14, s19, s5, s6, s2, s16, s17, s3, s4, and s20 (Table3).

3.2.3) Comparing the coefficients of raw and log transformed data

The estimates of the coefficients from raw data and log transformed changed data were then analysed to check which scale of the data has most significant predictor variables. The investigation revealed estimates for coefficients for predictor variables, s10, s12, s14, s16, s17, s20, s21, and s23 from the log-scale were more significant than their counterparts in the raw data. A new predictor data set (Best data set) was designed which is a concoction of the most significant predictor variables from both scales (Table3). Best data set was used in the prediction modelling.
Table 3. Coefficients for Raw data and log transformed data

<table>
<thead>
<tr>
<th>Var</th>
<th>Wald's p-value</th>
<th>AIC</th>
<th>Change in Deviance</th>
<th>Var</th>
<th>Wald's p-value</th>
<th>AIC</th>
<th>Change in Deviance</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>&lt; 0.0001</td>
<td>167.1</td>
<td>17.99</td>
<td>s1</td>
<td>&lt; 0.0001</td>
<td>173.32</td>
<td>11.78</td>
</tr>
<tr>
<td>P2</td>
<td>&lt; 0.0001</td>
<td>117.64</td>
<td>67.45</td>
<td>s2</td>
<td>&lt; 0.0001</td>
<td>128.76</td>
<td>56.33</td>
</tr>
<tr>
<td>P3</td>
<td>&lt; 0.0001</td>
<td>138.71</td>
<td>46.38</td>
<td>s3</td>
<td>&lt; 0.0001</td>
<td>139.67</td>
<td>45.42</td>
</tr>
<tr>
<td>P4</td>
<td>&lt; 0.0001</td>
<td>142.39</td>
<td>42.7</td>
<td>s4</td>
<td>&lt; 0.0001</td>
<td>148.52</td>
<td>36.57</td>
</tr>
<tr>
<td>P5</td>
<td>&lt; 0.0001</td>
<td>81.58</td>
<td>100.82</td>
<td>s5</td>
<td>&lt; 0.0001</td>
<td>110.58</td>
<td>74.52</td>
</tr>
<tr>
<td>P6</td>
<td>&lt; 0.0001</td>
<td>90.58</td>
<td>92.13</td>
<td>s6</td>
<td>&lt; 0.0001</td>
<td>121.06</td>
<td>64.03</td>
</tr>
<tr>
<td>P7</td>
<td>0.0548</td>
<td>181.41</td>
<td>3.69</td>
<td>s7</td>
<td>0.20</td>
<td>183.84</td>
<td>1.25</td>
</tr>
<tr>
<td>P8</td>
<td>0.2951</td>
<td>184</td>
<td>1.1</td>
<td>s8</td>
<td>0.3862</td>
<td>184.34</td>
<td>0.75</td>
</tr>
<tr>
<td>P9</td>
<td>&lt; 0.0001</td>
<td>169.46</td>
<td>10.83</td>
<td>s9</td>
<td>0.9875</td>
<td>185.09</td>
<td>0</td>
</tr>
<tr>
<td>P10</td>
<td>0.1669</td>
<td>183.18</td>
<td>1.91</td>
<td>s10</td>
<td>0.0507</td>
<td>181.27</td>
<td>3.82</td>
</tr>
<tr>
<td>P11</td>
<td>0.1115</td>
<td>154.42</td>
<td>2.53</td>
<td>s11</td>
<td>&lt; 0.0001</td>
<td>168.19</td>
<td>16.91</td>
</tr>
<tr>
<td>P12</td>
<td>&lt; 0.0001</td>
<td>126.39</td>
<td>41.32</td>
<td>s12</td>
<td>&lt; 0.01</td>
<td>176.99</td>
<td>8.1</td>
</tr>
<tr>
<td>P13</td>
<td>&lt; 0.0001</td>
<td>99.83</td>
<td>59.88</td>
<td>s13</td>
<td>&lt; 0.01</td>
<td>176.05</td>
<td>7.57</td>
</tr>
<tr>
<td>P14</td>
<td>&lt; 0.0001</td>
<td>137.11</td>
<td>47.99</td>
<td>s14</td>
<td>&lt; 0.0001</td>
<td>116.33</td>
<td>68.76</td>
</tr>
<tr>
<td>P15</td>
<td>&lt; 0.0001</td>
<td>143.78</td>
<td>13.17</td>
<td>s15</td>
<td>0.3367</td>
<td>184.17</td>
<td>0.92</td>
</tr>
<tr>
<td>P16</td>
<td>&lt; 0.0001</td>
<td>140.54</td>
<td>44.55</td>
<td>s16</td>
<td>&lt; 0.0001</td>
<td>123.16</td>
<td>61.93</td>
</tr>
<tr>
<td>P17</td>
<td>&lt; 0.0001</td>
<td>129.42</td>
<td>55.67</td>
<td>s17</td>
<td>&lt; 0.0001</td>
<td>123.74</td>
<td>61.35</td>
</tr>
<tr>
<td>P18</td>
<td>0.019</td>
<td>172.34</td>
<td>5.5</td>
<td>s18</td>
<td>0.4318</td>
<td>184.48</td>
<td>0.62</td>
</tr>
<tr>
<td>P19</td>
<td>&lt; 0.0001</td>
<td>98.27</td>
<td>86.83</td>
<td>s19</td>
<td>&lt; 0.0001</td>
<td>99.978</td>
<td>85.12</td>
</tr>
<tr>
<td>P20</td>
<td>&lt; 0.0001</td>
<td>165.88</td>
<td>19.21</td>
<td>s20</td>
<td>&lt; 0.0001</td>
<td>159.03</td>
<td>26.06</td>
</tr>
<tr>
<td>P21</td>
<td>&lt; 0.0001</td>
<td>124.57</td>
<td>35.14</td>
<td>s21</td>
<td>0.5764</td>
<td>184.78</td>
<td>0.31</td>
</tr>
<tr>
<td>P22</td>
<td>&lt; 0.0001</td>
<td>168.45</td>
<td>16.64</td>
<td>s22</td>
<td>&lt; 0.001</td>
<td>171.63</td>
<td>13.46</td>
</tr>
<tr>
<td>P23</td>
<td>0.571</td>
<td>184.77</td>
<td>0.32</td>
<td>s23</td>
<td>0.0115</td>
<td>178.7</td>
<td>6.39</td>
</tr>
<tr>
<td>P24</td>
<td>&lt; 0.0001</td>
<td>161.1</td>
<td>23.99</td>
<td>s24</td>
<td>&lt; 0.0001</td>
<td>164.63</td>
<td>20.46</td>
</tr>
<tr>
<td>P25</td>
<td>&lt; 0.001</td>
<td>171.96</td>
<td>13.13</td>
<td>s25</td>
<td>0.0247</td>
<td>180.05</td>
<td>5.05</td>
</tr>
<tr>
<td>P26</td>
<td>&lt; 0.0001</td>
<td>167.53</td>
<td>17.56</td>
<td>s26</td>
<td>0.0156</td>
<td>179.25</td>
<td>5.84</td>
</tr>
<tr>
<td>P27</td>
<td>&lt; 0.0001</td>
<td>164.36</td>
<td>20.73</td>
<td>s27</td>
<td>0.0156</td>
<td>179.24</td>
<td>5.85</td>
</tr>
<tr>
<td>P28</td>
<td>&lt; 0.01</td>
<td>176.15</td>
<td>8.94</td>
<td>s28</td>
<td>0.0102</td>
<td>178.49</td>
<td>6.6</td>
</tr>
<tr>
<td>P29</td>
<td>0.0165</td>
<td>179.35</td>
<td>5.74</td>
<td>s29</td>
<td>0.4242</td>
<td>184.45</td>
<td>0.64</td>
</tr>
<tr>
<td>P30</td>
<td>&lt; 0.01</td>
<td>177.11</td>
<td>7.98</td>
<td>s30</td>
<td>0.2222</td>
<td>183.6</td>
<td>1.49</td>
</tr>
<tr>
<td>P31</td>
<td>&lt; 0.01</td>
<td>176.95</td>
<td>8.14</td>
<td>s31</td>
<td>0.4004</td>
<td>184.39</td>
<td>0.71</td>
</tr>
<tr>
<td>P32</td>
<td>&lt; 0.0001</td>
<td>163.71</td>
<td>19</td>
<td>s32</td>
<td>0.3732</td>
<td>184.3</td>
<td>0.79</td>
</tr>
</tbody>
</table>

Estimates for the predictive variables from the raw data and log transformed changed data was shown in the table. **Var:** predictor variables
3.3) Model fitting

The predictor variables were selected by employing logistic regression using various statistical approaches, such as forward selection, backward search, stepwise selection, and all subset selection methods. These methods were performed on R, SPSS, and Stata software platforms as mentioned in the materials and methods section.

3.3.1) Forward selection

Forward selection method was performed in the SPSS software as I was having algorithm convergence problems in R. Stepwise forward selection had fitted s12, s17, s20, and s23 variables into a model (Table 4).

### Table 4. Variable equation in forward selection

<table>
<thead>
<tr>
<th>Variables in the Equation</th>
<th>OR</th>
<th>CI</th>
<th>Lower limit</th>
<th>Upper limit</th>
<th>Lower limit</th>
<th>Upper limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1⁰</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s12</td>
<td>1.786</td>
<td>0.323</td>
<td>30.515</td>
<td>0</td>
<td>5.968</td>
<td>3.17</td>
</tr>
<tr>
<td>s12</td>
<td>-7.572</td>
<td>1.229</td>
<td>37.99</td>
<td>0</td>
<td>0.001</td>
<td>3.94</td>
</tr>
<tr>
<td>s23</td>
<td>-1.047</td>
<td>0.306</td>
<td>11.684</td>
<td>0.001</td>
<td>0.351</td>
<td>0.19</td>
</tr>
<tr>
<td>s17</td>
<td>1.193</td>
<td>0.468</td>
<td>6.487</td>
<td>0.011</td>
<td>3.297</td>
<td>1.32</td>
</tr>
<tr>
<td>s23</td>
<td>-1.312</td>
<td>0.35</td>
<td>14.028</td>
<td>0</td>
<td>0.269</td>
<td>0.14</td>
</tr>
<tr>
<td>Step 2⁰</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s12</td>
<td>1.574</td>
<td>0.369</td>
<td>18.224</td>
<td>0</td>
<td>4.825</td>
<td>2.34</td>
</tr>
<tr>
<td>s17</td>
<td>1.193</td>
<td>0.468</td>
<td>6.487</td>
<td>0.011</td>
<td>3.297</td>
<td>1.32</td>
</tr>
<tr>
<td>s23</td>
<td>-1.312</td>
<td>0.35</td>
<td>14.028</td>
<td>0</td>
<td>0.269</td>
<td>0.14</td>
</tr>
<tr>
<td>Step 3⁰</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s12</td>
<td>1.574</td>
<td>0.369</td>
<td>18.224</td>
<td>0</td>
<td>4.825</td>
<td>2.34</td>
</tr>
<tr>
<td>s17</td>
<td>1.193</td>
<td>0.468</td>
<td>6.487</td>
<td>0.011</td>
<td>3.297</td>
<td>1.32</td>
</tr>
<tr>
<td>s23</td>
<td>-1.312</td>
<td>0.35</td>
<td>14.028</td>
<td>0</td>
<td>0.269</td>
<td>0.14</td>
</tr>
<tr>
<td>Step 4⁰</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>-7.223</td>
<td>2.956</td>
<td>5.97</td>
<td>0.015</td>
<td>0.001</td>
<td></td>
</tr>
</tbody>
</table>

a. Variable(s) entered on step 1: s12; b. Variable(s) entered on step 2: s23; c. Variable(s) entered on step 3: s17; and d. Variable(s) entered on step 4: s20.

3.3.2) Backward search and bidirectional search

Backward and bidirectional search in R yielded the same predictor variables in the model and failed to fit full 32 variable models in all the software I used in the study. Algorithms in these programs software did not converge. This can be blamed on the
smaller number of events (45) compared to the number of predictor variables. Hence, this method cannot be applied in our study (Table 5). Backward and bidirectional search yielded a model with 18 variables (P2 + P4 + P5 + P8 + P11 + P13 + P15 + P18 + P19 + P22 + P26 + P27 + P28 + P31 + s10 + s12 + s20 + s21) with an AIC value of 38.

Table 5. Bidirectional search method

<table>
<thead>
<tr>
<th>Step</th>
<th>Deviance</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>7.76</td>
<td>36.45</td>
</tr>
<tr>
<td>+ s12</td>
<td>9.03</td>
<td>55.09</td>
</tr>
<tr>
<td>+ s10</td>
<td>8.91</td>
<td>53.23</td>
</tr>
<tr>
<td>+ P31</td>
<td>8.76</td>
<td>50.95</td>
</tr>
<tr>
<td>+ P2</td>
<td>8.65</td>
<td>49.33</td>
</tr>
<tr>
<td>+ P22</td>
<td>8.56</td>
<td>47.79</td>
</tr>
<tr>
<td>+ P5</td>
<td>8.44</td>
<td>45.86</td>
</tr>
<tr>
<td>+ P27</td>
<td>8.42</td>
<td>45.57</td>
</tr>
<tr>
<td>+ P4</td>
<td>8.37</td>
<td>44.82</td>
</tr>
<tr>
<td>+ P19</td>
<td>8.28</td>
<td>43.36</td>
</tr>
<tr>
<td>+ P13</td>
<td>8.28</td>
<td>43.28</td>
</tr>
<tr>
<td>+ P28</td>
<td>8.28</td>
<td>43.26</td>
</tr>
<tr>
<td>+ P18</td>
<td>8.23</td>
<td>42.46</td>
</tr>
<tr>
<td>+ P8</td>
<td>8.15</td>
<td>41.15</td>
</tr>
<tr>
<td>+ P15</td>
<td>8.14</td>
<td>41.01</td>
</tr>
<tr>
<td>+ s21</td>
<td>8.13</td>
<td>40.91</td>
</tr>
<tr>
<td>- P6</td>
<td>7.76</td>
<td>38.45</td>
</tr>
<tr>
<td>- P30</td>
<td>7.75</td>
<td>38.39</td>
</tr>
<tr>
<td>- P25</td>
<td>7.75</td>
<td>38.39</td>
</tr>
<tr>
<td>- s23</td>
<td>7.75</td>
<td>38.38</td>
</tr>
<tr>
<td>- P29</td>
<td>7.75</td>
<td>38.35</td>
</tr>
<tr>
<td>- P3</td>
<td>7.75</td>
<td>38.34</td>
</tr>
<tr>
<td>- P24</td>
<td>7.75</td>
<td>38.34</td>
</tr>
<tr>
<td>- P9</td>
<td>7.75</td>
<td>38.32</td>
</tr>
<tr>
<td>- P7</td>
<td>7.75</td>
<td>38.23</td>
</tr>
<tr>
<td>- P32</td>
<td>7.74</td>
<td>38.22</td>
</tr>
<tr>
<td>- P1</td>
<td>7.72</td>
<td>37.87</td>
</tr>
<tr>
<td>- s16</td>
<td>7.72</td>
<td>37.86</td>
</tr>
<tr>
<td>- s14</td>
<td>7.72</td>
<td>37.85</td>
</tr>
<tr>
<td>+ P26</td>
<td>7.94</td>
<td>37.60</td>
</tr>
<tr>
<td>- s17</td>
<td>7.71</td>
<td>37.58</td>
</tr>
<tr>
<td>+ s20</td>
<td>7.92</td>
<td>37.28</td>
</tr>
<tr>
<td>+ P11</td>
<td>7.90</td>
<td>36.97</td>
</tr>
</tbody>
</table>

‘+’addition of the variable to the model; ‘-’ is deletion
3.3.3) All subset combinations (from best data set)

This was a CPU intensive approach, where all the possible subset combinations were calculated from the 32 variables. Logistic regression analysis was carried out on the best data set raw data using ‘Generalized Linear Model’ (GLM) method in R. A script was written to obtain subsets containing 2 to 8 variable set combinations using the ‘Bluebear’ (Supplementary Table 3). This method resulted in the subset combinations which had significant values for the assessing measure of fit such as ‘deviance’ and/or ‘AIC’.

The time scales given below were for running the ‘glm’ (generalized linear model) script that was on a regular desktop. The script for 2-variable subset models yielded 496 combinations in less than 10 minutes. Best of these combinations was (A): P5 + P19, with deviance=57.5 and AIC=63.5. The script for 3-variable subset yielded 4960 different combinations in less than 30 minutes. The most significant model derived from the 3-variable subset combination was (B): P5 + P19+P31, with deviance=39.8 and AIC=47.8 (Table 6).

35,960 4-variable subset combinations were calculated, and the output was produced in less than 3 hours. The predictive significance in these models had improved significantly as compared to 2, 3-variable subset models, with the best model (C): P2+P19+ s21+P31. There were 201,376 different 5-variable subset models that can be obtained from 32 variables in 10 hours. The best model in these combinations was (D): P2+P6+P11+P19+P31 (Table 6).

The ‘glm’ script for 6-variable combination subset took more than 40 hours to run. So, for the variations in this script and for the large numbered (7 and above) variable subsets, the scripts were run in R on the ‘Bluebear’ platform. The script for 6-variable subset models returned 906,192 with the most significant model (E): P2+P6+P19+s21+P25+P30. The predictive capability has significantly increased for these models (Table 6).

The scripts for 7 and 8-variable subset model returned 3,365,856 and 10,518,300 combinations respectively. The best models for these combinations were (F):P2+P6+P19+s21+P22+P25+P30, and (G): P2+P3+P6+P15+P19+P22+P25+P30, respectively. These scripts took around 40 hours each to return the results on the Bluebear.

The significance of the model subsets increased with the increase in the number of variables in the model in all the subset combinations with deviance. However, the significance measure ‘AIC’ increased up to 6 variable subset model and then started declined on addition of further variables to the model (Table 6).
Table 6. Best 3 models from each subset combination

<table>
<thead>
<tr>
<th>V1</th>
<th>V2</th>
<th>V3</th>
<th>V4</th>
<th>V5</th>
<th>V6</th>
<th>V7</th>
<th>V8</th>
<th>Deviance</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>6</td>
<td>19</td>
<td>21</td>
<td>28</td>
<td>31</td>
<td></td>
<td></td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>11</td>
<td>19</td>
<td>25</td>
<td>30</td>
<td></td>
<td></td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>19</td>
<td>21</td>
<td>25</td>
<td>30</td>
<td></td>
<td></td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>11</td>
<td>19</td>
<td>22</td>
<td>25</td>
<td>29</td>
<td></td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>15</td>
<td>19</td>
<td>22</td>
<td>25</td>
<td>30</td>
<td></td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>19</td>
<td>21</td>
<td>22</td>
<td>25</td>
<td>30</td>
<td></td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>17</td>
<td>19</td>
<td>21</td>
<td>22</td>
<td>25</td>
<td>30</td>
<td>0</td>
<td>18</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>6</td>
<td>15</td>
<td>19</td>
<td>22</td>
<td>25</td>
<td>30</td>
<td>0</td>
<td>18</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>6</td>
<td>19</td>
<td>21</td>
<td>22</td>
<td>25</td>
<td>30</td>
<td>0</td>
<td>18</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>11</td>
<td>19</td>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td>13.256</td>
<td>25.256</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>13</td>
<td>19</td>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td>14.179</td>
<td>26.179</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>19</td>
<td>21</td>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td>14.247</td>
<td>26.247</td>
</tr>
<tr>
<td>2</td>
<td>19</td>
<td>21</td>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>24.281</td>
<td>34.281</td>
</tr>
<tr>
<td>6</td>
<td>19</td>
<td>21</td>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>24.625</td>
<td>34.625</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>19</td>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>25.073</td>
<td>35.073</td>
</tr>
<tr>
<td>5</td>
<td>19</td>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>39.832</td>
<td>47.832</td>
</tr>
<tr>
<td>5</td>
<td>19</td>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>39.977</td>
<td>47.977</td>
</tr>
<tr>
<td>6</td>
<td>19</td>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>40.089</td>
<td>48.089</td>
</tr>
<tr>
<td>5</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>57.521</td>
<td>63.521</td>
</tr>
<tr>
<td>5</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>59.007</td>
<td>65.007</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>63.268</td>
<td>69.268</td>
</tr>
</tbody>
</table>

The coefficients (deviance and AIC) of the best 3 models from each subset combination was shown here. V1 to V8 are variables 1 to 8 in the model. The change in deviance has improved significantly with the addition of an extra variable. However, AIC value degraded with the addition of an extra variable to the 6-variable subset model.

Investigations on the significance of estimates for each predictor variable were carried using the best prediction model of each subset combination.

3.4) Area under the ROC curve

The area under the receiver operating characteristic (ROC) curve was used to measure the discrimination and visualize the sensitivity and specificity of the prediction models in the study. The ROC curves for the best models from 2-8 variable subsets (A to G) and forward selection model were drawn. The sensitivity and specificity for 2-variable subset was found to be 94%. A similar sensitivity and specificity (AUC= 94.8%) was found for the stepwise forward selection model.
(s12+s17+s20+s23). The measurements for the sensitivity and specificity has increased with the addition of each variable, 3-variable subset (97.8%), 4-variable subset (99.3%), 5-variable subset (99.8%), and attained a perfect 100% sensitivity and specificity from 6-variable subsets onwards (Figure 2).

![Figure 2. Area under the ROC curve values for 2-variable (A) through to 7-variable (G) subset models (as mentioned in the Section 3.3.3) and forward selection model (H).]
3.5) Correlation among the best predictive variables

3.5.1) Best predictor variables from the subset models

Best predictor variables were investigated by checking the most frequent variables in the best subset models. An investigation into the 12 best models from the 2 to 8 variable subsets was carried out to find the most frequent variables in these best models subsets. As depicted in Figure 3, the most frequent variables in these best subset models were, P2, P5, P6, P11, P15, P19, s21, P22, P25, P30, and P31.

Figure 3. Frequencies of the variables in 2-8 variable subset models were shown here. Here, bar diagrams A-G are showing the variable frequencies in 2-8 variable subsets.
3.5.2) Correlation

An investigation on the correlation among the best predictive variables was carried out. Strong association between the predictor variables will undermine the significance for the estimates and also result in the inflation of variance. Analysis revealed high correlation among variables from the same steroid group as compared to the other steroid types. Predictor variable P2 has higher correlation with P6, P22, and s21. Variables P5, P3 and P6 have high correlation with each other. P6, P2, P5, and P3 were highly correlated. Similarly, predictor variables P22, P25, P31, P30, s21, and P2 had high correlation with each other (Table 7&8).

**Table 7.** Pearson product-moment

<table>
<thead>
<tr>
<th></th>
<th>P2</th>
<th>P3</th>
<th>P5</th>
<th>P6</th>
<th>P11</th>
<th>P15</th>
<th>P19</th>
<th>P22</th>
<th>P25</th>
<th>P30</th>
<th>P31</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P3</td>
<td>0.44</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P5</td>
<td>0.40</td>
<td>0.71</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P6</td>
<td>0.57</td>
<td>0.51</td>
<td>0.54</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P11</td>
<td>0.28</td>
<td>0.19</td>
<td>0.32</td>
<td>0.18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P15</td>
<td>0.37</td>
<td>0.20</td>
<td>0.19</td>
<td>0.36</td>
<td>0.10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P19</td>
<td>0.32</td>
<td>0.31</td>
<td>0.20</td>
<td>0.39</td>
<td>0.00</td>
<td>0.15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P22</td>
<td>0.56</td>
<td>0.07</td>
<td>0.16</td>
<td>0.27</td>
<td>0.15</td>
<td>0.16</td>
<td>0.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P25</td>
<td>0.56</td>
<td>0.12</td>
<td>0.18</td>
<td>0.33</td>
<td>0.23</td>
<td>0.22</td>
<td>0.22</td>
<td>0.91</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P30</td>
<td>0.36</td>
<td>0.05</td>
<td>0.18</td>
<td>0.36</td>
<td>0.27</td>
<td>0.25</td>
<td>0.78</td>
<td>0.77</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P31</td>
<td>0.49</td>
<td>0.01</td>
<td>0.13</td>
<td>0.33</td>
<td>0.18</td>
<td>0.20</td>
<td>0.09</td>
<td>0.87</td>
<td>0.88</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td>s21</td>
<td>0.55</td>
<td>0.17</td>
<td>0.29</td>
<td>0.44</td>
<td>0.27</td>
<td>0.32</td>
<td>0.23</td>
<td>0.69</td>
<td>0.65</td>
<td>0.72</td>
<td>0.63</td>
</tr>
</tbody>
</table>

Significant correlations were highlighted in yellow (high) and red (low) colour. The identical half of the table was trimmed to make it look more aesthetically appealing.

**Table 8.** Pair-wise p-values

<table>
<thead>
<tr>
<th></th>
<th>P2</th>
<th>P3</th>
<th>P5</th>
<th>P6</th>
<th>P11</th>
<th>P15</th>
<th>P19</th>
<th>P22</th>
<th>P25</th>
<th>P30</th>
<th>P31</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P3</td>
<td>&lt;0.0001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P5</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P6</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P11</td>
<td>0.0011</td>
<td>0.0249</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P15</td>
<td>&lt;0.0001</td>
<td>0.8369</td>
<td>0.0308</td>
<td>&lt;0.0001</td>
<td>0.2281</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P19</td>
<td>0.0002</td>
<td>0.0002</td>
<td>0.0266</td>
<td>&lt;0.0001</td>
<td>0.9954</td>
<td>0.0843</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P22</td>
<td>&lt;0.0001</td>
<td>0.3933</td>
<td>0.0583</td>
<td>&lt;0.0016</td>
<td>0.0852</td>
<td>0.066</td>
<td>0.099</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P25</td>
<td>&lt;0.0001</td>
<td>0.1604</td>
<td>0.0365</td>
<td>&lt;0.0001</td>
<td>&lt;0.01</td>
<td>0.0184</td>
<td>0.0108</td>
<td>&lt;0.0001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P30</td>
<td>&lt;0.0001</td>
<td>0.5970</td>
<td>0.0403</td>
<td>0.0003</td>
<td>&lt;0.01</td>
<td>&lt;0.01</td>
<td>&lt;0.01</td>
<td>&lt;0.0001</td>
<td>0.00001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P31</td>
<td>&lt;0.0001</td>
<td>0.9094</td>
<td>0.1344</td>
<td>&lt;0.0001</td>
<td>0.0366</td>
<td>0.0224</td>
<td>0.2774</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
<td></td>
</tr>
<tr>
<td>s21</td>
<td>&lt;0.0001</td>
<td>0.0549</td>
<td>0.0007</td>
<td>&lt;0.0001</td>
<td>&lt;0.0013</td>
<td>&lt;0.0001</td>
<td>&lt;0.01</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

P2 to s21 were predictor variables from the best models. The identical half of the table was trimmed to make it look more aesthetically appealing.
3.5.3) Inspection of predictor variables from each subset

The prediction capability of the best model from each subset was analysed with the help of ‘scatter plot’ graph (Figure 4). When these models were validated on each of the patients in the data set, a clear variation in the predictive capabilities between these models was found. Variable subsets 2, 3, 4 and 5 showed inconsistency in the predictive capability. However, variable subsets 6, 7, and 8 showed a perfect prediction capability.

![Figure 4. Scatter plot depicting the model fit for each patient.](image)

Each scatter plot shows fitness of a subset model on each patient’s measurements in the data (number 1-45 are ACC patients, P=1, and patient number 46-147 are ACA patients P=0). The subset model represented here is the best model (A-G) from each variable subset.

The estimates for the best predictor variables from all the subset models were obtained by logistic regression. These estimates were then standardized by dividing the predictor variable measurements with their respective Inter Quartile Range (IQR) values. The strength of the predictor variable estimates was examined starting from their own estimates and throughout the model building procedure. A good model will have its estimates gaining/losing gradually in this process. A sudden multi-fold hike or drop can be a result of ‘overfitting’.

The strength of the estimates was found to be unstable with the addition of a variable to 5-variable subset model, as it is evident in the case of P19, where the significance of the estimate increased gradually up to 5 variable subset, a sudden 19-fold
increase in the 6 variable subset model, and a 75% drop in the strength of estimate in the 7 variable subset. The strength of the estimates for individual predictor variable in the majority of subset models was random as well. This phenomenon of ‘overfitting’ can be blamed on the small sample size, interaction, and collinearity (Table 9).

Table 9. Inspection of predictor variable’s estimates in each subset model

<table>
<thead>
<tr>
<th></th>
<th>Intercept</th>
<th>Estimate</th>
<th>IQR</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>P5</td>
<td>-3.11</td>
<td>2.17</td>
<td>414.00</td>
<td>1.70</td>
<td>2.83</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P19</td>
<td>-2.47</td>
<td>1.20</td>
<td>381.00</td>
<td>0.93</td>
<td>2.99</td>
<td>5.93</td>
<td>11.44</td>
<td>197.09</td>
<td>47.36</td>
<td></td>
</tr>
<tr>
<td>P31</td>
<td>-1.45</td>
<td>0.40</td>
<td>560.00</td>
<td>-3.63</td>
<td>-6.96</td>
<td>-22.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P2</td>
<td>-2.76</td>
<td>1.57</td>
<td>1339.00</td>
<td>4.21</td>
<td>8.84</td>
<td>128.99</td>
<td>66.09</td>
<td>83.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>s21</td>
<td>-7.20</td>
<td>1.29</td>
<td>1.14</td>
<td>0.87</td>
<td>28.83</td>
<td>26.31</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P6</td>
<td>-2.90</td>
<td>1.61</td>
<td>641.00</td>
<td>2.71</td>
<td>66.22</td>
<td>63.55</td>
<td>64.07</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P11</td>
<td>-1.34</td>
<td>0.14</td>
<td>22.00</td>
<td>-2.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P25</td>
<td>-1.49</td>
<td>0.44</td>
<td>543.00</td>
<td>-187.28</td>
<td>-179.79</td>
<td>-196.57</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P30</td>
<td>-1.58</td>
<td>0.47</td>
<td>1089.00</td>
<td>-170.32</td>
<td>-201.90</td>
<td>-120.34</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P22</td>
<td>-1.44</td>
<td>0.38</td>
<td>1967.00</td>
<td>130.39</td>
<td>141.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P3</td>
<td>-1.52</td>
<td>0.36</td>
<td>433.00</td>
<td>-6.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P15</td>
<td>-1.72</td>
<td>0.57</td>
<td>19.50</td>
<td>-24.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The estimates for each predictor variables in the best model from each subset were analysed. Here, IQR, Inter Quartile Range, A is 2-variable subset: P5+P19, B is 3-variable subset: P5+P19+P31, C is 4-variable subset: P2+P19+s21+P31, D is 5-variables subset: P2+P6+P11+P19+P31, E is 6-variable subset: 2+P6+P19+s21+P25+P30, F is 7-variable subset: P2+P6+P19+s21+P22+P25+P30, and G is 8-variable subset: P2+P3+P6+P15+P19+P22+P25+P30.
4) Discussion

Analysis of urinary steroids and steroid metabolite data from adrenal tumour patients revealed that some of the steroids and steroid metabolites have a distinct secretion pattern between ACC and ACA patients. In ACC patients, a significantly higher excretion of An, Etio (active androgens), 5-PD, DHEA, 16α-OHDHEA (androgen precursor metabolites), THDOC, 5α-THDOC (mineralocorticoid precursor metabolites) PD, PT, 17-HP, THS (Glucocorticoid precursor metabolites), THE, 6β-OH-F, and α-cortol (Glucocorticoid metabolites) was observed when compared to ACA (Table1&Figure1).

This study was aimed at designing a biomarker tool by exploiting the differences in steroid and steroid metabolites excretion levels between ACC and ACA patients. The choice of the statistical method, regression analysis, was made basing on some criteria, such as the performance of the method, distribution of the data, missing values, sample size, and expertise.

Logistic regression analysis revealed that some of the variables have higher predictive capability. However, the order of significance was slightly different depending on the measurement criteria, for example, log-likelihood, deviance, p-value, and AIC criteria gave s19 (THS), s5 (5-PT), s14 (PD), s6 (5-PD), s16 (17HP), s17 (PT), s2 (Etio), s3 (DHEA), s4 (16α-OH DHEA), and s20 (F) in the order of most significance first. However, for Wald’s statistics the most significant variables, although identical to the other methods, were yielded in a different order: s14, s19, s5, s6, s2, s16, s17, s3, s4, and s20 (Table3). The differences in the output can be attributed to the random presence of large measurement values for the predictor variables in the data, which renders Wald’s statistics insignificant (type2 error) (13).

This study used stepwise regression methods and all subset combination for fitting the models. Among stepwise regression methods, only forward selection method has worked with this data. Forward selection method fitted s12, s17, s20, and s23 variables into a model. This model has a significant sensitivity and specificity (AUC=94.8%). However, a thorough investigation into the strength and stability of model estimates, interaction and collinearity between variables, and validation on a new data set is required. Backward selection method failed to fit any models as the number of events in the study is only 45 (ACC=45) and there were 32 predictor variables to fit. A large sample would have been ideal for the application of other testing methods.

Among the methods employed, all subset combination methods yielded most parsimonious models with better statistical significance (Table5). However, this method is computationally intensive and will require more time and high performance computing, such as cluster computing. Inspection of subset models on each of the
patient sample in the data revealed an increase in the predictive capability with the increase in the number of variables (Figure 2, Supplementary table 3). The sensitivity and specificity measure for these models using ROC curve showed a perfect predictive capability in the larger subset models (from 6-variable subset and above) (Figure 2). However, an investigation into the significance of estimates in each model revealed a possible ‘overfitting’ in these models.

Among the subset models 4 and 5-variable subset models showed a significant prediction capability with a sensitivity and specificity (AUC) measure of 99.3 % and 99.8%. The strength of the variable estimates in these models remained stable throughout the process. However, a closer examination into the ‘collinearity’ and ‘interaction’ between the predictor variables in the model is necessary.

Backward and bidirectional search yielded a model with 18 variables (P2 + P4 + P5 + P8 + P11 + P13 + P15 + P18 + P19 + P22 + P26 + P27 + P28 + P31 + s10 + s12 + s20 + s21) with an AIC value of 38. Although this model could describe the variables in the model, this method might degrade the precision of the estimates of predictor variables in the model.

This study has been an exploratory one. But, future studies can benefit from employing a larger sample population, a better strategy to deal with the missing values, a literature review and an expert advice in reducing the number of predictor variables, and by comparing with the alternative mathematical models such as Trees and neural networks at the same time (2, 6, 9, and 10).
References


### Supplementary Table 1. Demographic and clinical characteristics of adrenal tumour (AT) patients

<table>
<thead>
<tr>
<th></th>
<th>ACA (n=102)</th>
<th>ACC (n=45)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median age (range) at time of urine collection (yr)</td>
<td>60 (19-84)</td>
<td>55 (20-80)</td>
</tr>
<tr>
<td>Sex (male, female)</td>
<td>39, 63</td>
<td>24, 21</td>
</tr>
<tr>
<td>Tumour load at the time of collection</td>
<td>AT (n=102)</td>
<td>AT, no metastasis (n=9)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AT+metastasis (n=26)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ACC metastasis after removal of primary tumour (n=10)</td>
</tr>
<tr>
<td>Maximum diameter of AT at the time of urine collection (median and range)</td>
<td>26 (9-78) mm</td>
<td>90 (14-230) mm</td>
</tr>
<tr>
<td>Surgical removal of AT</td>
<td>24/102 (24%)</td>
<td>30/45 (67%)</td>
</tr>
<tr>
<td>Median Weiss score(a)</td>
<td>1 (0-2) (n=15)</td>
<td>6 (3-9) (n=21)</td>
</tr>
<tr>
<td>Duration of follow-up (median and range)</td>
<td>52 (26-201) months since diagnosis (n=102)</td>
<td>14 (1–187) months from diagnosis to death due to metastatic ACC in deceased patients (n=35)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>45 (30–100) months since diagnosis in alive patients (n=10)</td>
</tr>
</tbody>
</table>

\(a\) The Weiss system scores the presence or absence of nine histopathological features (Weiss score range 0–9); scores under 3 are indicative of a benign adrenal tumour, scores of 3 are borderline, and scores of 4 and above are indicative of malignancy (14).

\(b\) Seven of the 10 surviving patients suffer from metastatic disease. The three remaining patients have not shown evidence of recurrence yet (all three initially presented with early-stage disease [ENS@T II (13); primary tumour diameters 80, 89, and 160 mm; histology indicative of ACC with Weiss scores of 5, 7, and 4, respectively; current survival times 45, 51, and 42 months, respectively).

** ** Taken from Wiebke et al. 2011 (Reference 17).
**Supplementary Table 2.** Description of the steroid and steroid metabolites used in the study

<table>
<thead>
<tr>
<th>No.</th>
<th>Abbreviation</th>
<th>Common name</th>
<th>Chemical name</th>
<th>Metabolite of</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Androgen metabolites</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>An</td>
<td>Androsterone</td>
<td>5α-androstan-3α-ol-17-one</td>
<td>Androstenedione, testosterone, 5α-dihydro testosterone</td>
</tr>
<tr>
<td>2</td>
<td>Etio</td>
<td>Etiocholanolone</td>
<td>5β-androstan-3α-ol-17-one</td>
<td>Androstenedione, testosterone</td>
</tr>
<tr>
<td><strong>Androgen precursor metabolites</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>DHEA</td>
<td>Dehydroepiandrosterone</td>
<td>5-androsten-3β-ol-17-one</td>
<td>DHEA + DHEA sulphate (DHEAS)</td>
</tr>
<tr>
<td>4</td>
<td>16α-OHDHEA</td>
<td>16α-hydroxy-DHEA</td>
<td>5-androstene-3β, 16α-diol-17-one</td>
<td>DHEA + DHEAS</td>
</tr>
<tr>
<td>5</td>
<td>5-PT</td>
<td>Pregnenetriol</td>
<td>5-pregnene-3β, 17, 20α-triol</td>
<td>17-hydroxy pregnenolone</td>
</tr>
<tr>
<td>6</td>
<td>5-PD</td>
<td>Pregnenediol</td>
<td>5-pregnene-3β, 20α-dioland 5, 17, (20)-pregnadien-3β-ol</td>
<td>pregnenolone</td>
</tr>
<tr>
<td><strong>Mineralocorticoid metabolites</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>THA</td>
<td>Tetrahydro-11-dehydro corticosterone</td>
<td>5β-pregnane-3α, 21-diol, 11, 20-dione</td>
<td>corticosterone, 11-dehydro corticosterone</td>
</tr>
<tr>
<td>8</td>
<td>5α-THA</td>
<td>5α-tetrahydro-11-dehydro corticosterone</td>
<td>5α-pregnane-3α, 21-diol-11, 20-dione</td>
<td>corticosterone, 11-dehydro corticosterone</td>
</tr>
<tr>
<td>9</td>
<td>THB</td>
<td>Tetrahydro corticosterone</td>
<td>5β-pregnane-3α, 11β, 21-triol-20-one</td>
<td>corticosterone</td>
</tr>
<tr>
<td>10</td>
<td>5α-THB</td>
<td>5α-tetrahydro corticosterone</td>
<td>5α-pregnane-3α, 11β, 21-triol-20-one</td>
<td>corticosterone</td>
</tr>
<tr>
<td>11</td>
<td>3α5β-THALDO</td>
<td>Tetrahydro aldosterone</td>
<td>5β-pregnane-3α, 11β, 21-triol-20-one-18-al</td>
<td>aldosterone</td>
</tr>
<tr>
<td><strong>Mineralocorticoid precursor metabolites</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>THDOC</td>
<td>Tetrahydro-11-deoxy corticosterone</td>
<td>5β-pregnane-3α, 21-diol-20-one</td>
<td>11-deoxy corticosterone</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>--------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>5α-THDOC</td>
<td>5α-tetrahydro-11-deoxy corticosterone</td>
<td>5α-pregnane-3α, 21-diol-20-one 11-deoxy corticosterone</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>PD</td>
<td>Pregnanediol</td>
<td>5β-pregnane-3α, 20adiol progesterone</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>3α5α-17HP</td>
<td>3α, 5α-17-hydroxypregnanolone</td>
<td>5α-pregnane-3α, 17α-diol-20-one 17-hydroxy progesterone</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>17HP</td>
<td>17-hydroxy pregnanolone</td>
<td>5β-pregnane-3α, 17α,20α-triol 17-hydroxy progesterone</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>PT</td>
<td>Pregnanetriol</td>
<td>5β-pregnane-3α, 17α,20α-triol 17-hydroxy progesterone</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>PTONE</td>
<td>Pregnanetriolone</td>
<td>5β-pregnane-3α, 17,20α-triol-11-one 21-deoxycortisol</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>THS</td>
<td>Tetrahydro-11-deoxycortisol</td>
<td>5β-pregnane-3α, 17, 21-triol-20-one 11-deoxycortisol</td>
<td></td>
</tr>
</tbody>
</table>

**Glucocorticoid precursor metabolites**

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>F</td>
<td>Cortisol</td>
<td>4-pregnene-11β, 17, 21-triol-3, 20-dione cortisol</td>
</tr>
<tr>
<td>21</td>
<td>6β-OH-F</td>
<td>6β-hydroxy-cortisol</td>
<td>4-pregnene-6β, 11β, 17,21-tetrol-3, 20-dione cortisol</td>
</tr>
<tr>
<td>22</td>
<td>THF</td>
<td>Tetrahydrocortisol</td>
<td>5β-pregnane-3α, 11β,17, 21-tetrol-20-one cortisol</td>
</tr>
<tr>
<td>23</td>
<td>5α-THF</td>
<td>5α-tetrahydrocortisol</td>
<td>5α-pregnane-3α, 11β,17, 21-tetrol-20-one cortisol</td>
</tr>
<tr>
<td>24</td>
<td>α-cortol</td>
<td>α-cortol</td>
<td>5β-pregnan-3α, 11β, 17,20α, 21-pentol cortisol</td>
</tr>
<tr>
<td>25</td>
<td>β-cortol</td>
<td>β-cortol</td>
<td>5β-pregnan-3α, 11β, 17,20β, 21-pentol cortisol</td>
</tr>
<tr>
<td>26</td>
<td>11β-OH-An</td>
<td>11β-hydroxy androsterone</td>
<td>5α-androstane-3α, 11β-diol-17-one cortisol (+androgens)</td>
</tr>
<tr>
<td>27</td>
<td>11β-OH-Et</td>
<td>11β-hydroxy etiocholanolone</td>
<td>5β-androstane-3α, 11β-diol-17-one cortisol (+androgens)</td>
</tr>
</tbody>
</table>
**Table 3. Scripts used in model selection methods**

#### Stepwise regression

```r
> fit <-
glm(cancer~P1+P2+P3+P4+P5+P6+P7+P8+P9+P11+P13+P15+P18+P19+P22+P24+P25+P26+P27 +P28+P29+P30+P31+s10+s12+s14+s16+s17+s20+s21+s23,data=AT)
> step <- stepAIC(fit, direction="both")
```

Start: AIC=60.3

```
cancer ~ P1 + P2 + P3 + P4 + P5 + P6 + P7 + P8 + P9 + P11 + P13 + P15 + P18 + P19 + P22 + P24 + P25 + P26 + P27 + P28 + P29 + P30 + P31 + s10 + s12 + s14 + s16 + s17 + s20 + s21 + s23
```  

#### Stepwise backward elimination

```r
> fit <-
glm(cancer~P1+P2+P3+P4+P5+P6+P7+P8+P9+P11+P13+P15+P18+P19+P22+P24+P25+P26+P27 +P28+P29+P30+P31+s10+s12+s14+s16+s17+s20+s21+s23,data=AT)
> step <- stepAIC(fit, direction="backward")
```

Start: AIC=60.3

```
cancer ~ P1 + P2 + P3 + P4 + P5 + P6 + P7 + P8 + P9 + P11 + P13 + P15 + P18 + P19 + P22 + P24 + P25 + P26 + P27 + P28 + P29 + P30 + P31 + s10 + s12 + s14 + s16 + s17 + s20 + s21 + s23
```
Script used to generate all subset combinations in the blue bear

```r
###data
data<-read.csv(file.choose(), header = T, sep = ",")

###Number of predictors
preds<-32

###Number chosen
choose<-4

library(gtools)

###Number of combinations
#library(combinat)
#library(R.basic)
#n_choose<-nChooseK(preds,choose)
n_choose<-choose(preds,choose)

###Time count
time.start<-Sys.time()

###Combinations
pred_combs<-combinations(n=preds,r=choose)

###Fitting models
glm_output<-rep(0, n_choose)
for (i in 1:n_choose)
{
glm_m_output<-glm(data$cancer~data[, pred_combs[i,1]]+data[, pred_combs[i,2]]+data[, pred_combs[i,3]]+data[, pred_combs[i,4]], family="binomial")
glm_output[i]<-glm_m_output$deviance
}

###deviance output and ordering
deviance_mod<-cbind(pred_combs, glm_output)
deviance_mod[order(deviance_mod[,5]),]
deviance_mod<-as.data.frame(deviance_mod)

time.end<-Sys.time()

print(difftime(time.end, time.start))

write.table(deviance_mod, file="C:\Prj2\Models_deviance\AT_chgv_logdeviance_mod3.csv", sep="", row.names=FALSE, quote = FALSE)
```