Urine Steroid Metabolomics for the Differential Diagnosis of Inborn Steroidogenesis Disorders

By

Elizabeth S. Baranowski

A thesis submitted to The University of Birmingham

for the degree of

DOCTOR OF PHILOSOPHY

Institute of Metabolism and Systems Research

College of Medical and Dental Sciences

The University of Birmingham

April 2022

Abstract

The adrenal cortex and gonads produce steroid hormones involved in salt and glucose homeostasis, blood pressure regulation, stress response and sex differentiation. These hormones are produced via a series of enzymatic steps and metabolites of steroids from each step are excreted and measurable in urine. Inborn disorders of steroidogenesis result from genetic mutations in distinct enzymes, causing a block in hormone production and lead to several forms of Congenital Adrenal Hyperplasia and differences in sex development. Each enzyme deficiency is characterised by a distinct pattern of altered excretion of individual steroid metabolites relating to the specific enzymatic block. Ratios of urine steroid metabolites can be employed as surrogates of distinct steroidogenic enzyme activities, enabling diagnosis from single spot urine samples, ideal for use with paediatric patients. widespread use in the acute setting for diagnosis of these disorders is hampered by the considerable expertise required for interpretation.

Here, we developed two approaches for assisting the detection and differentiation of inborn steroidogenic disorders. First, we have designed a clinical decision tool using biochemical analysis by established steroid precursor-to-product metabolite ratios. Second, we developed a novel steroid metabolomics approach, combining mass spectrometry-based steroid profiling with intrinsically interpretable machine learning-based data analysis. The analysis methods presented can expedite and standardise interpretation of complex urinary steroid metabolome data, making this technique more accessible to clinicians, and has excellent potential for implementation in routine clinical practice.

Acknowledgments

Firstly, I wish to acknowledge the agencies that have funded this work, the Birmingham Children's Hospital Research Foundation, and the Medical Research Council.

I would like to thank my esteemed and inspiring supervisors, Prof. Wiebke Arlt, Prof. Peter Tino and Prof. Jon Deeks. Thank you for your patience, your unwavering belief, your understanding, and for being such great role models for aspiring academics.

Prof. Wolfgang Högler, Prof. Tim Barrett and Prof. Nils Krone, without the support of whom I would not have started this research journey in the first place.

All the amazing people I have met working in the IMSR, but specifically Dr Angela Taylor, Dr Lorna Gilligan, Fozia Shaheen of SMAC who provided training and support. Hannah Ivison, thank you for your technical and emotional support. And the cookies!

I will forever be indebted to the MASSter, Prof. Cedric Shackleton, my unofficial 4th supervisor, and my official teacher, mentor and drinking friend. Thank you for your endless enthusiasm, for drawing steroids on napkins, reminding me about balance and for opening your home (and your friend's boat!) to me.

Thank you to my collaborators. My PhD counterpart and better academic half, Dr Sreejita Ghosh. I hope we continue to learn from each other, or at least continue to trade Dr Who memes. Prof. Kerstin Bunte for being far too clever and believing that you can drag me to your level. Dr Sarah Berhane, thank you for your time and kindness

Thank you to my family. My neglected partner in crime, Anthony Fearn (if you exist….). My Mum, Pam, and Stepdad, Robert, for always being my loudest cheerleaders.

I would like to dedicate this thesis to Mr Stefan J. Baranowski, whose stubbornness, weird sense of humour and poor time keeping live on in me. Gone but never forgotten.

[1] *“Do it. Then it will be done.” Hannah’s Mom.*

Table of Contents

List of Figures

List of Tables

List of Abbreviations

11-keto-DHT	11-keto-dihydrotestosterone
11-keto-T	11-keto-testosterone
11-oxo-Etio	11-oxo-etiocholanolone
11βOHAn	11β-hydroxyandrosterone
11βOHD	11β-hydroxylase deficiency
11βOHEt	11β-hydroxyetiocholanolone
15β-triolone	3α,15β,17α-trihydroxy-5β-pregnan-20-one
16α-OH-DHEA	16α-hydroxy-dehydroepiandrosterone
17αOHD	17α-hydroxylase deficiency
17HP	17α-hydroxypregnanolone
17OHP	17α-hydroxyprogesterone
17OHPreg	17α-hydroxypregnenolone
18-OH-THA	18-hydroxytetrahydro-11-dehydrocorticosterone
21OHD	21-hydroxylase deficiency
3α5α17HP	5α-17-hydroxypregnanolone
3βHSD2D	3β-hydroxysteroid dehydrogenase type 2 deficiency
3β5βTHAldo	3β,5β-tetrahydroaldosterone
5αRD	5α-reductase type 2 deficiency
5αTHA	5α-tetrahydro-11-dehydrocorticosterone
5αTHB	5α-Tetrahydrocorticosterone
5αTHB	5α-tetrahydro-corticosterone

5αTHF	5α-tetrahydrocortisol
5PD	Pregnenediol
5PT	5-pregnene-3β,17α,20α-triol
6α-hydroxyTHS	6α-hydroxytetrahydrotetrahydro-11-deoxycortisol
6αOHTHA	6α-hydroxy 11-dehydro-tetrahydrocorticosterone
6β-OHF	6β-hydroxycortisol
ACA	Adrenal cortical adenoma
ACC	Adrenal cortical carcinoma
ACTH	Adrenocorticotrophin releasing hormone
AKR	Aldo-keto reductase
AKR1C3	17β-hydroxysteroid dehydrogenase type 5
ALVQ	Angle learning vector quantization
An	Androsterone
AUC	Area under the curve
BA	Balanced accuracy
CAH	Congenital adrenal hyperplasia
CB	Cholesterol butyrate
CI	Confidence interval
CRH	Corticotrophin releasing hormone
CV	Coefficient of variation
CYB5A	Cytochrome b5
CYP	Cytochrome P450
CYP11A1	Cholesterol side chain cleavage enzyme
CYP11B1	11β-hydroxylase

CYP11B2	Aldosterone synthase
CYP17A1	17α-hydroxylase
CYP19A1	P450 Aromatase
CYP21A2	21-hydroxylase
DHEA	Dehydroepiandrosterone
DHEAS	Dehydroepiandrosterone-sulfate
DHT	5α-dihydrotestosterone
DOC	11-deoxycorticosterone
DSD	Differences in sex development
DZ	Definitive zone
E	Cortisone
EI	Electron Impact
EQA	External quality assurance
ES	External standard
Et	Etiocholanolone
F	Cortisol
FAD	Flavin adenine dinucleotide
FMN	Flavin mononucleotide
FPU	Foeto-placental unit
FZ	Foetal zone
GC	Gas chromatography
GC-MS	Gas chromatography-mass spectrometry
GMLVQ	Generalised matrix relevance learning vector quantization
GPDR	General data protection regulation
GRLVQ	Generalised relevance learning vector quantization

hCG	Human chorionic gonadotropin
HDL	High density lipoproteins
HPA axis	Hypothalamic-pituitary-adrenal axis
HSD	Hydroxysteroid dehydrogenases
HSD11B1	11β-hydroxysteroid dehydrogenase type 1
HSD11B2	11β-hydroxysteroid dehydrogenase type 2
HSD17B	17β-hydroxysteroid dehydrogenase
HSD17B3	17β-hydroxysteroid dehydrogenase type 3
HSD3B1	3β-hydroxysteroid dehydrogenase type 1
HSD3B2	3β-hydroxysteroid dehydrogenase type 2
IQR	Interquartile range
IS	Internal standard
kNN	K-nearest neighbour
LC	Liquid chromatography
LC-MS/MS	Liquid chromatography tandem mass spectrometry
LDL	Low density lipoproteins
LVQ	Learning vector quantization
m/z	Mass-to-charge ratio
MAA	Macroaveraged accuracy
MAR	Missing at random
MCAR	Missing completely at random
MNAR	Missing not at random
MS	Mass spectrometry
NADH	Nicotinamide adenine dinucleotide
NADPH	Nicotinamide adenine dinucleotide phosphate

PALVQ	Probabilistic angle learning vector quantization
PAPS	3′-phosphoadenosine 5′-phosphosulfate
PAPSS2	PAPS synthase type 2
PCOS	Polycystic ovary syndrome
PD	Pregnanediol
PD	Pregnanediol
POR	P450 oxidoreductase
PORD	P450 oxidoreductase deficiency
PPCA	Probabilistic principal component analysis
PT	Pregnanetriol
PTONE	Pregnanetriolone
QC	Quality control
ROC	Receiver operating characteristic
RSD	Relative standard deviation
SDR	Short chain dehydrogenase/reductase
SFC	Supercritical fluid chromatography
SFC-MS	Supercritical fluid chromatography-mass spectrometry
SIM	Selected ion monitoring
SMOTE	Synthetic minority oversampling technique
SRD5A2	5α-reductase type 2
SS	Stigmasterol
StAR	Steroidogenic acute regulatory protein
STARD	Standards for reporting diagnostic accuracy
Std error	Standard error
STD/std	Standard deviation

STS	Steroid sulfatase
SULT2A1	Sulfotransferase
THA	Tetrahydro- 11-dehydrocorticosterone
THAldo	3α,5β- tetrahydroaldosterone
THB	Tetrahydrocorticosterone
THB	Tetrahydrocorticosterone
THDOC	Tetrahydro-11-deoxycorticosterone
THE	Tetrahydrocortisone
THF	Tetrahydrocortisol
THS	Tetrahydro-11-deoxycortisol
TIC	Total ion chromatogram
TMSI	N-trimethylsilylimidazole
t-SNE	T-distributed stochastic neighbour embedding
TZ	Transitional zone
ULN	Upper limit of normal
ZF	Zona fasciculata
ZG	Zona glomerulosa
ZR	Zona reticularis

1. General Introduction

Sections of this chapter have been published in the following papers:

Monogenic Disorders of Adrenal Steroidogenesis.

Baranowski ES, Arlt W, Idkowiak J. Hormone Research in Paediatrics. 2018 Jul;89(5):292-310. doi: 10.1159/000488034

Steroid Metabolome Analysis in Disorders of Adrenal Steroid Biosynthesis and Metabolism. Storbeck KH, Schiffer L, Baranowski ES, et al. Endocrine Reviews 2019 Dec;40(6):1605-1625. doi: 10.1210/er.2018-00262

Human Steroid Biosynthesis, Metabolism and Excretion are Differentially Reflected by Serum and Urine Steroid Metabolomes: A Comprehensive Review. Schiffer L, Barnard L, Baranowski ES, et al. J Steroid Biochem Mol Biol. 2019 Nov;194:105439. doi: 10.1016/j.jsbmb.2019.105439.

1.1. Adrenal Gland Structure and Function

The adrenal glands are two triangular shaped organs situation on the upper pole of each kidney. They contain two macroscopically distinct parts, the outer cortex, and the inner medulla, arising from different embryological origins- the mesenchyme and the neural crest cells respectively. The adrenal medulla secretes catecholamines, whereas the adrenal cortex is responsible for the production of steroid hormones.

The adrenal cortex produces 3 classes of steroid hormone: mineralocorticoids, glucocorticoids, and precursors for active androgens. Mineralocorticoids are responsible for salt and water balance, effected by sodium transport via the kidneys, which is important for blood pressure regulation. (Fuller & Young, 2005). Glucocorticoids are a pivotal part of the body's stress response, with a functional role in increasing vascular tone, glucose homeostasis via action on enzymes critical for glycogenolysis and gluconeogenesis, and immune system modulation (Cruz-Topete & Cidlowski, 2015; Timmermans et al., 2019). Androgens are sex hormones responsible for external masculinisation, development of phenotypically male secondary sexual characteristics during puberty and have a role in sexual function (Forest, 1983). These three steroid hormone subtypes are synthesised from cholesterol as a common substrate via a series of enzymatic reactions. Biosynthesis of adrenal steroid hormones in the adult human adrenal cortex occurs in three functional zones which are distinct both histologically and functionally by differential enzyme expression. Mineralocorticoids are primarily produced in the outermost Zona Glomerulosa layer, Glucocorticoids predominantly in the middle Zona Fasciculata

layer and Androgen Precursors predominantly in the innermost Zona Reticularis layer (Schiffer et al., 2019) **(Figure 1.1)**

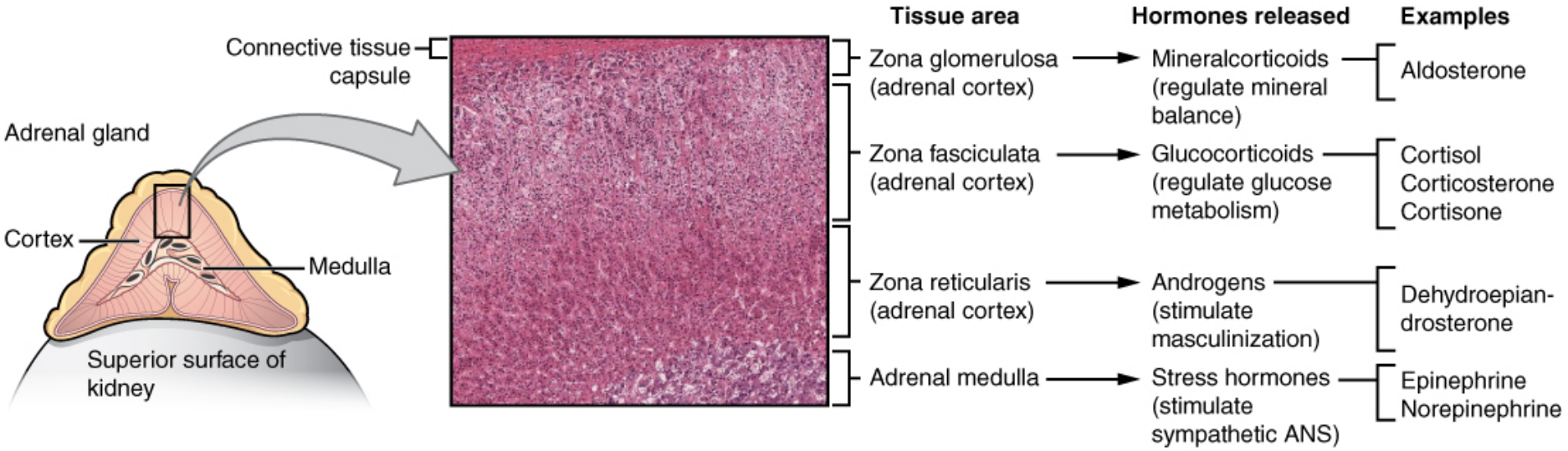


Figure 1.1: Adrenal Zonation

The histological zones observed in the adrenal gland and the steroid hormones they produce. (OpenStaxCollege)

1.2. Steroid Core Structure

Being derived from a common substrate (cholesterol), all steroid hormones have a closely related molecular structure with a common "steroid backbone". The basic core structure is typically composed of seventeen carbon atoms, bonded in a cyclopentanophenanthrene 4-ring structure. This comprises three six carbon cyclohexane rings (rings A, B and C) and one five-carbon cyclopentane ring (the D ring). Biochemically altered through various enzymatic reactions, the different adrenal steroid hormones vary by the functional groups attached to this 4-ring backbone and by the oxidation state of the rings (Carlson, 1989) **(Figure 1.2).**

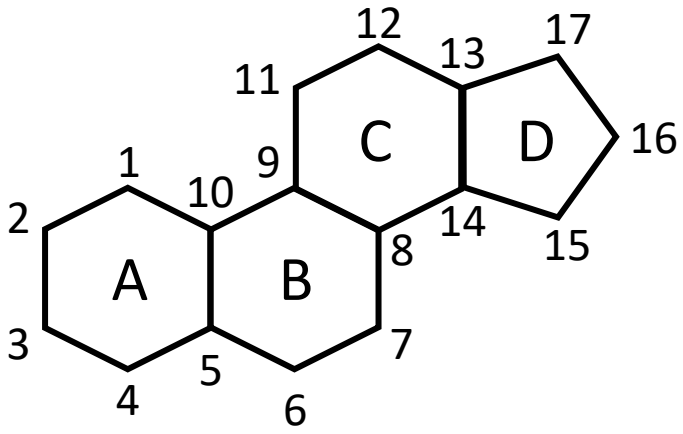


Figure 1.2: Steroid Core Structure

The arrangement of three six carbon cyclohexane rings and one cyclopentane ring which is common to all steroids. The positions of the seventeen carbon atoms constituting this structure are shown here numbered.

1.3. Adrenal Steroidogenic Enzymes

There are two major classes of enzyme critical for steroid biosynthesis in humans: Cytochrome P450 (CYP) enzymes and Hydroxysteroid dehydrogenase (HSD) enzymes. In the process of steroidogenesis, CYP enzymes catalyse hydroxylation reactions and Carbon-Carbon bond cleavage. HSD enzymes catalyse the conversion of a given hydroxysteroid to its corresponding ketosteroid counterpart and back again (Schiffer et al., 2019).

1.3.1. Cytochrome P450 (CYP) Enzymes

The CYP enzymes are a large group of cysteinato-heme containing oxidative enzymes which are found naturally in all forms of life (Meunier et al., 2004). They are named according to the carbon atom they act upon. The term “P450” is derived from the term “Pigment 450”. This group of enzymes are so named because, in their reduced states complexed with carbon monoxide, they absorb light at 450 nm (Miller & Auchus, 2011). For oxidation reactions, CYPs use electrons donated by nicotinamide adenine dinucleotide phosphate (NADPH) to insert one oxygen atom from molecular oxygen into a substrate, while the other oxygen atom is reduced to water (Meunier et al., 2004). There are two types of CYP enzyme important for human steroidogenesis, termed “type 1” and “type 2”. The two types are distinct by intracellular location and by the way electrons are transferred from NADPH to facilitate the oxidative reaction (Miller & Auchus, 2011; Schiffer et al., 2019). CYP type 1 enzymes are situated in the inner mitochondrial membrane. Electron donation from NADPH for type 1 CYP enzymes requires ferredoxin reductase (a flavoprotein, also known as adrenodoxin

reductase within the adrenal gland) and ferredoxin (a small iron-sulphur protein, also known as adrenodoxin within the adrenal gland). Ferredoxin reductase oxidizes NADPH and transfers electrons to ferredoxin which in turn carries and delivers the electrons to the CYP enzyme (Meunier et al., 2004; Schiffer et al., 2019). Type 2 CYP enzymes are situated in the endoplasmic reticulum. Electrons are donated to Type 2 CYP enzymes by a flavin adenine dinucleotide (FAD) and flavin mononucleotide (FMN) containing reductase called P450 Oxidoreductase (POR). The FAD component facilitates the receipt of electrons from NADPH and the FMN component facilitates the transfer to the CYP enzyme (Meunier et al., 2004; Schiffer et al., 2019).

1.3.2. Hydroxysteroid Dehydrogenase (HSD) Enzymes

Hydroxysteroid Dehydrogenase enzymes (HSDs) are a group of short chain dehydrogenase enzymes; they are named according to the number assigned to the carbon they act upon. Multiple isoforms exist for each HSD enzyme, arising from distinct genes. There are two subgroups of HSD enzyme defined by their structure: short-chain dehydrogenase/reductase (SDR) and aldo-keto reductase (AKR) (Miller & Auchus, 2011). Both families are dependent on a cofactor in the form of nicotinamide adenine dinucleotide (NADH/NAD+) or nicotinamide adenine dinucleotide phosphate (NADPH/NADP+) to act as an electron acceptor/donor to catalyse the oxidoreduction of steroids. Unlike CYPs, HSDs generally catalyse bidirectional reactions, and as such are involved in both synthesis and inactivation of steroid hormones by conversion of hydroxysteroids to their corresponding ketosteroids or the other way round (Penning, 1997). The dehydrogenase reaction to oxidise a hydroxysteroid to a ketosteroid mainly depends on the presence of NAD+ as the required cofactor. The reductase reaction

to reduce a ketosteroid to an hydroxysteroid mainly depends on the presence of NADPH as the required co-factor (Penning, 1997). There tends to be an observed directional preference in vivo for these bidirectional reactions which is dependent on the relative affinity of an enzyme for each co-factor, co-factor availability and cellular redox status (Schiffer et al., 2019). The two 3β-hydroxysteroid dehydrogenase isozymes (HSD3B1 and HSD3B2) are exceptional within the HSD family because they catalyse an irreversible dehydrogenase reaction by additional isomerisation of a Δ^5 double bond to a Δ^4.(Luu-The et al., 1991; Thomas et al., 1995)

1.4. Adrenal Steroidogenesis

A schematic for the process of adrenal steroidogenesis is seen in **Figure 1.3**. All adrenal steroid hormones are produced from a common precursor – cholesterol. The predominant supply of cholesterol for the adrenal gland is derived from dietary sources (Gwynne & Strauss, 1982) but, if necessary, endogenous cholesterol can be synthesised within the adrenal gland from acetate (Mason & Rainey, 1987). The current understanding of the crucial preliminary step required for adrenal steroidogenesis is described by Miller WL in a 2017 review (Miller, 2017). Plasma low-density lipoproteins (LDL) and high-density lipoproteins (HDL) are taken up via receptor mediated endocytosis. Intracellular lysosomes facilitate the hydrolysis of LDL cholesterol esters to free cholesterol by Lysosomal Acid Lipase. Adrenocorticotrophin releasing hormone (ACTH) produced by the pituitary gland increases the availability of free cholesterol for use in steroidogenesis by three different methods. 1) Inhibiting enzymes which would re-esterify free cholesterol (cholesterol acyl transferase), 2) upregulating enzymes involved in freeing stored cholesterol (hormone-sensitive

lipase) and 3) upregulating enzymes which encourage LDL endocytosis (3-hydroxy-3-methylglutaryl co-enzyme A reductase). The steroidogenic acute regulatory protein (StAR) shifts cholesterol from the outer to the inner mitochondrial membrane. Cholesterol side chain cleavage enzyme (CYP11A1) is located on the inner mitochondrial membrane. It performs 3 consecutive modifying reactions upon the cholesterol molecule: firstly, 20α-hydroxylation, secondly, 22R-hydroxylation, and finally, carbon side-chain cleavage of C20-C22 which results in the production of pregnenolone (Chung et al., 1986; Shikita & Hall, 1973). CYP11A1 is expressed in all 3 layers of the adrenal cortex (Suzuki et al., 2000). Production of zone specific downstream steroid hormones is dependent on the differential expression of specific combinations of steroidogenic enzymes.

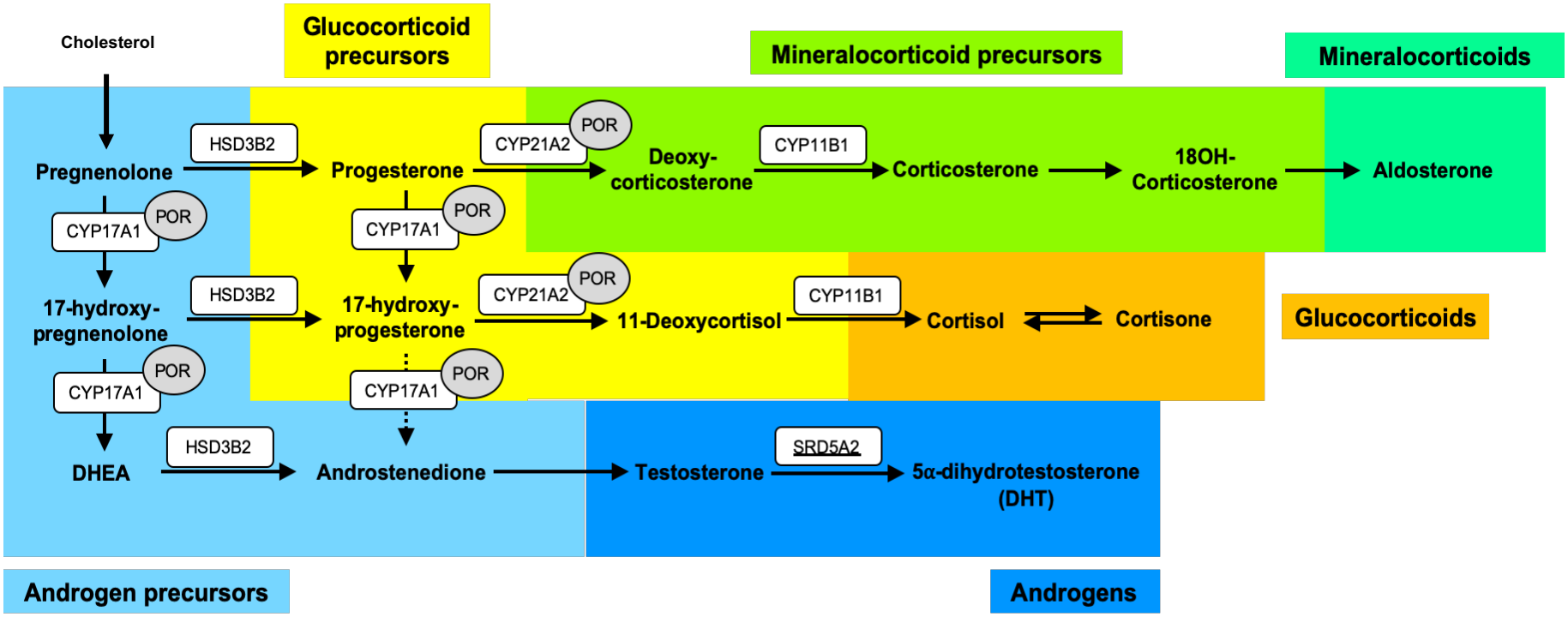


Figure 1.3: Adrenal and Gonadal Steroidogenesis

Enzymes are indicated by white boxes and the co-factor P450 oxidoreductase is indicated by the grey circles. Only enzymes or co-factors investigated in this manuscript are shown. Enzymes shown, if defective lead to Congenital Adrenal Hyperplasia (Hargitai et al.) +/- Differences in Sex Development (Engels et al.) and underlined enzyme defect leads to DSD in isolation.

1.4.1. Mineralocorticoid Production in the Zona Glomerulosa

Enzymes required to produce mineralocorticoids are preferentially expressed in the zona glomerulosa, namely 3β-hydroxysteroid dehydrogenase type 2 (HSD3B2), 21-hydroxylase (CYP21A2) and the two isoforms of 11β-hydroxylase (CYP11B1 and CYP11B2). The lack of expression of 17α-hydroxylase (CYP17A1) in the zona glomerulosa means that steroid hormone precursors are committed to mineralocorticoid biosynthesis in this zone (Coulter & Jaffe, 1998; Suzuki et al., 2000). HSD3B2 generates progesterone from pregnenolone (Schiffer et al., 2019). It has 2 catalytic activities: it oxidizes the hydroxyl group at carbon 3 to a keto group with its dehydrogenase activity, and then subsequently changes the carbon-carbon double bond from the position between C5/6 ("Δ^5") to C4/5 ("Δ^4") with its Δ4/Δ5 isomerase activity (Lachance et al., 1990; Thomas et al., 1989). 21-hydroxylase (CYP21A2) then converts progesterone via hydroxylation on carbon 21 to the potent mineralocorticoid precursor 11-deoxycorticosterone (DOC) (Schiffer et al., 2019). Further activation of DOC to aldosterone is performed in 3 subsequent steps. Firstly, 11β-hydroxylation of DOC yields corticosterone, secondly, 18-hydroxylation of corticosterone forms 18-hydroxycorticosterone, and finally the 18-methyl oxidation of corticosterone generates aldosterone. The first two steps can be performed by both isoforms of 11β-hydroxylase (CYP11B1 and CYP11B2), whereas the final step is performed exclusively by aldosterone synthase (CYP11B2) (Baranowski et al., 2018; Kawamoto et al., 1992; Mulatero et al., 1998; Schiffer et al., 2019).

1.4.2. Glucocorticoid Production in the Zona Fasciculata

Within both the zona fasciculata and glomerulosa there are high amounts of HSD3B2 expressed, facilitating the conversion of pregnenolone to progesterone. CYP17A1 is also highly expressed in the zona fasciculata and competes with HSD3B2 for substrates to direct production towards glucocorticoid and sex steroid biosynthesis (Suzuki et al., 2000). With its 17α-hydroxylase activity, CYP17A1 introduces a hydroxyl group on carbon position 17 of pregnenolone and progesterone to form 17α-hydroxypregnenolone (17OHPreg) and 17α-hydroxyprogesterone (17OHP), respectively (Chung et al., 1987; Yanase et al., 1991). HSD3B2 is also able to catalyse the conversion of 17OHPreg to 17OHP with similar efficiency to the conversion of pregnenolone and DHEA (Lee et al., 1999). 17OHP, the principal glucocorticoid precursor, is converted to 11-deoxycortisol, by CYP21A2. Finally, CYP11B1 converts 11-deoxycortisol in one single oxidative reaction to the major active glucocorticoid, cortisol (Schiffer et al., 2019).

1.4.3. Androgen Precursor Production in the Zona Reticularis and the Classical Androgen Pathway

The Zona Reticularis layer of the human adrenal gland develops during adrenarche, which occurs between the ages of 6 and 10 years. There is strong expression of CYP17A1, abundant presence of both the co-factors POR and cytochrome b5 (CYB5A), and relatively scant expression of HSD3B2 (Auchus & Rainey, 2004; Nakamura, Gang, et al., 2009). This enables the zona reticularis to produce robust amounts of the androgen precursor dehydroepiandrosterone (DHEA) (Miller, 2002).

CYP17A1 has a second catalytic activity in addition to 17α-hydroxylation, 17,20 lyase activity, whereby it cleaves 3 carbon atoms from position 17 of the hydroxylated compounds 17OHPreg and 17OHP, yielding DHEA and androstenedione (Zuber et al., 1986). Both catalytic activities of CYP17A1 require electron transfer from NADPH, which is facilitated by the redox enzyme P450 oxidoreductase (POR) (Miller & Auchus, 2011). CYP17A1 17,20 lyase activity is greatly enhanced by additional interaction with the small haemoprotein CYB5A, which is solely expressed in the zona reticularis (Akhtar et al., 2005; Auchus et al., 1998; Lee-Robichaud et al., 1995). In normal physiology, CYP17A1 17,20 lyase activity has a 100-fold higher substrate preference for 17OHPreg than for 17OHP, which means that under physiological conditions most androgen synthesis in the adrenal goes through DHEA in the classical "Δ^5 pathway" (Auchus et al., 1998; Fluck et al., 2003). The mechanisms by which CYB5A facilitates the 17,20 lyase activity of CYP17A1 are not entirely understood. It has been shown that CYB5A can donate electrons to cytochrome P450 enzymes including CYP17A1 (Duggal et al., 2016; Zhang et al., 2008), but it has been suggested that it has only indirect involvement in electron transfer (Auchus et al., 1998). It has further been suggested that CYB5A allosterically modulates CYP17A1 activity. Recent work indicates direct binding between CYB5A and CYP17A1, inducing a conformational change in the molecule which positively impacts ligand entry (Bart & Scott, 2017; Estrada et al., 2014). Unlike cortisol and aldosterone which leave the adrenal cortex as "end products" with high affinities to their respective receptors, DHEA is inert and requires further conversion to active sex steroids in the gonads and peripheral target tissues (Labrie et al., 2003). The majority of DHEA leaves the adrenal in the form of its sulfate ester, DHEAS, which is by far the most abundant steroid hormone in the post-adrenarchal circulation in higher primate species (Bernstein et al., 2012; Rege et

al., 2013). The sulfation of DHEA is catalyzed in the zona reticularis by the enzyme DHEA sulfotransferase (SULT2A1) (Suzuki et al., 2000). DHEA sulfation requires activated sulfate in the form of 3′-phosphoadenosine 5′-phosphosulfate (PAPS), generated and delivered to SULT2A1 by PAPS synthase type 2 (PAPSS2) (Mueller et al., 2015). The conversion of DHEA to androstenedione is catalyzed by HSD3B2, and greater than 50% of circulating plasma androstenedione is of adrenal origin (Abraham, 1974; Arlt et al., 1998; Judd et al., 1974). It was proposed that adrenal androstenedione is produced in the overlapping area between the zona fasciculata and reticularis, where CYP17A1 / POR / CYB5A and HSD3B2 are all expressed (Nakamura et al., 2011). HSD3B1, expressed in many non-steroidogenic tissues (i.e., liver, skin, adipose, kidney, and others) also generates androstenedione from DHEA (Labrie et al., 1994; Martel et al., 1994). Additionally, ovarian HSD3B2 may also contribute to androstenedione generation in premenopausal women (Abraham, 1974; Arlt et al., 1999; Arlt et al., 1998; Haning et al., 1991). Further extra-adrenal downstream conversion towards potent, androgen receptor-activating androgens is performed by 17β-hydroxysteroid dehydrogenases (HSD17Bs) (Labrie et al., 1994; Miller & Auchus, 2011). In the testes, Leydig cells express both HSD3B2 and HSD17B3 which catalyse the production of testosterone from DHEA via androstenedione and androstenediol (Schiffer et al., 2018). Testosterone generation is also catalysed in small proportions in the adrenal cortex by the enzyme 17β-hydroxysteroid dehydrogenase type 5 (AKR1C3) (Nakamura, Hornsby, et al., 2009). Finally, peripheral 5α-reduction by 5α-reductases yield 5α-dihydrotestosterone (DHT), the most potent androgen (Miller & Auchus, 2011; Schiffer et al., 2018). This pathway is known as the classical pathway to DHT biosynthesis **(Figure 1.4**).

1.4.4. Adrenal Precursors of the 11-Oxygenated Androgen Pathway

Substantial quantities of androstenedione are converted to 11β-hydroxyandrostenedione, and the significance of this has been brought to the forefront by recent work. 11β-hydroxylation of androstenedione, catalyzed by CYP11B1, forms 11β-hydroxyandrostenedione. This was originally thought to be an inert biproduct of androgen synthesis but has been revealed to be a precursor to other 11-oxygenated androgens, an alternative class of potent active androgens (Fig. 1.4) (Pretorius et al., 2017). Huge quantities of 11-oxygenated androgens are secreted and have been found in greater quantities than adrenally derived androstenedione before and after ACTH stimulation (Rege et al., 2013). 11β-hydroxyandrostenedione is converted to 11-keto-androstenedione by 11β-hydroxysteroid dehydrogenase type 2 (HSD11B2), which is further converted to 11-keto-testosterone (11-keto-T) by the action of AKR1C3. 11-keto-T serves as a substrate for 11β-hydroxysteroid dehydrogenase type 1 (HSD11B1) for the generation of 11-keto-dihydrotestosterone (11-keto-DHT) (Storbeck et al., 2013). The in vitro androgen receptor-transactivating activity of 11-keto-T and 11-keto-DHT is similar to that of testosterone and DHT (Pretorius et al., 2016). The significance of this additional androgen class is wide-ranging. Recent studies have shown that they play a key role in androgen excess states, representing the majority of circulating androgens in polycystic ovary syndrome (PCOS) (O'Reilly et al., 2017) and found to be increased in Congenital Adrenal Hyperplasia (Hargitai et al.) due to 21-hydroxylase deficiency (21OHD) (Turcu et al., 2016). In addition, 11-oxygenated androgens are thought to contribute to the pathogenesis of castration resistant prostate cancer (Barnard et al., 2020).

1.4.5. The Alternative Pathway to DHT Synthesis

The earlier described pathway leading to the production of DHT is described as the "classical" androgen pathway. Through work looking particularly at the Tammar Wallaby, an alternative pathway to DHT synthesis which bypasses the conventional androgen precursors was discovered which uses progesterone and 17OHP as the preliminary substrates (Shaw et al., 2006; Wilson et al., 2003) **(Figure 1.4**). This alternative or "backdoor" pathway to active androgen synthesis is observed in human physiology both in healthy and disease states, with roles proposed in healthy male foetal genital development and in conditions of androgen excess such as CAH due to 21OHD. It has also been shown to explain the genital ambiguity observed in genetic females with P450 Oxidoreductase deficiency (Jones et al., 2017; Miller & Auchus, 2019; Reisch et al., 2019). Through the action of 5α-reductase type 1, Progesterone is 5α-reduced to 5α-dihydroprogesterone and 17OHP is 5α-reduced to 17α-hydroxydihydroprogesterone. These substrates are then further 3α-reduced by AKR1C enzymes to allopregnanolone and 17α-hydroxyallopregnanolone. Similar to the conversion of Progesterone to 17OHP by the action of CYP17A1, allopregnanolone can also be converted to 17α-hydroxyallopregnanolone which is further converted to androsterone through the 17,20 lyase activity of CYP17A1. Androsterone undergoes a reversible 17β-reduction reaction catalysed by AKR1C3 to produce androstanediol which then finally undergoes a reversible 3α-oxidation reaction in peripheral tissues to produce active androgen DHT (Reisch et al., 2019; Schiffer et al., 2019).

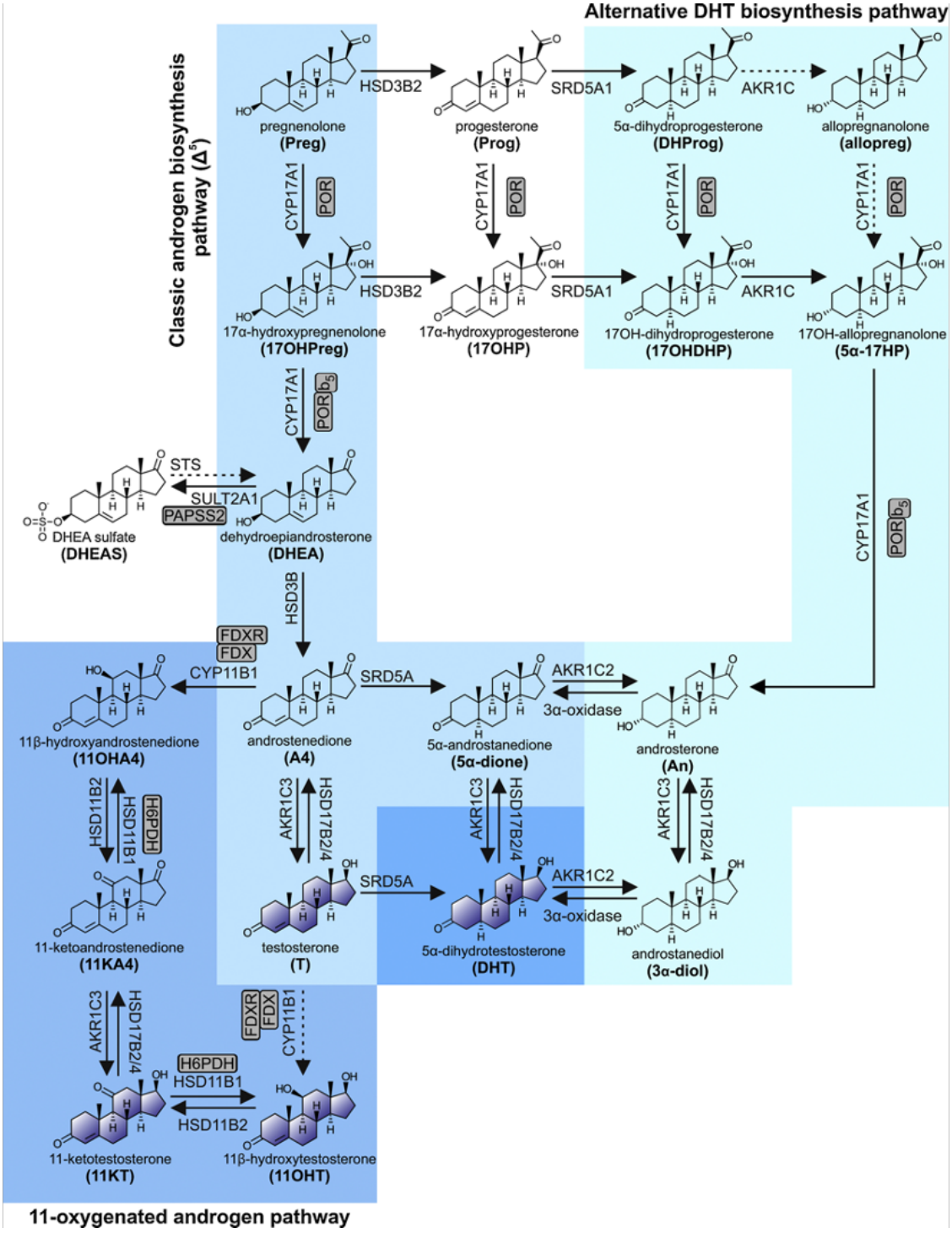


Figure 1.4: Three pathways to active androgen synthesis

*The classical pathway to DHT synthesis, the alternative DHT biosynthesis pathway and the 11-oxygenated androgen pathway (*Schiffer et al., 2019*)*

1.5. Steroidogenesis from Conception to Adult Life

Key changes in adrenal structure and function over the life course have been well described in the literature and are summarised in **Figure 1.5**. Antenatally, the foetal adrenal gland is both structurally and functionally quite different to the healthy adult adrenal gland. The foetal adrenal gland rapidly grows throughout pregnancy and by full term is disproportionately large, up to 10-20 times greater than the size of an adult adrenal gland. Each foetal adrenal gland weighs approximately 3-5g (Keene, 1927), which is a similar weight to adult adrenal glands which have a mean combined weight of 11.8g (range 5.8 – 19.9g) as determined from autopsy findings (Lam et al., 2001). 80% of the foetal adrenal weight is attributed to the inner "Foetal Zone" (FZ). The remaining adrenal mass is attributed to the outer "Definitive Zone" (Zalas et al.) and a third smaller intermediate "Transitional Zone" (TZ) which develops after mid-gestation (Ishimoto & Jaffe, 2011). The FZ lacks HSD3B1/2 and highly expresses CYP17A1 so is primarily responsible for Δ^5 steroid production (Johnston et al., 2018). The TZ and DZ express the required enzymes for glucocorticoid and mineralocorticoid production and in late pregnancy begin to resemble adult zona fasciculata and glomerulosa respectively (Hill et al., 2010). The proportionally large size of the FZ reflects the substantial production of androgen precursors, which are destined to be aromatised to oestrogens in the placenta. Postnatally, the foetal adrenal rapidly decreases in size due to atrophy of the FZ (Ishimoto & Jaffe, 2011).

The steroid production and excretion in pregnancy is remarkable. It was noticed in the 1930s that pregnanediol and oestriol were the main maternal urinary products, and each could be excreted at levels greater than 50mg/day at the end of gestation. It was

early recognized that pregnanediol was the metabolite of placental progesterone, but it took until the 1960s for oestriol to be designated as a metabolite of foetal adrenal DHEAS. It was at this time that the concept of the foetoplacental unit was developed, in large part through the work of Egon Diczfalusy and others (Diczfalusy, 1964).

1.5.1. The Foeto-Placental Unit

The foeto-placental unit (FPU) represents the collaborative steroidogenic function of the placenta and foetus. Individually, they express different steroidogenic enzymes but together they can produce the full spectrum of mineralocorticoids, glucocorticoids and androgens. Owing to a lack of CYP17A1 and high expression of HSD3B1 and Aromatase (CYP19A1), the placenta is principally responsible for production of progesterone (following cessation of progesterone production from the corpus luteum around 8 weeks gestation), and aromatisation of foetal androgen precursors to oestrogens (Pezzi et al., 2003). A schematic illustrating the main steroidogenic outputs of the relative components of the FPU is illustrated in **Figure 1.6**.

1.5.2. Progesterone Production in the Placenta

CYP11A1 is highly expressed in the placenta, foetal adrenals (FZ and TZ mostly) and testes (Johnston et al., 2018; Pezzi et al., 2003). Cholesterol is delivered to the placenta and crosses the mitochondrial membrane prior to CYP11A1 action. This occurs independent of StAR protein, instead using MNL64 (a constitutive protein with START domain), so even in lipoid adrenal hyperplasia due to StAR mutations pregnenolone is produced (Bose et al., 1996; Bose et al., 2000). In the placenta,

pregnenolone is efficiently converted to progesterone by HSD3B1 (Pezzi et al., 2003; Tuckey, 2005). Virtually all progesterone within the FPU and maternal compartment originates in the placenta. Production increases rapidly, and by term reaches up to 250mg/day (Simpson & MacDonald, 1981). Placental progesterone circulates around the FPU but ultimately almost all returns to the maternal compartment- principally excreted as pregnanediol (5β-pregnane-3α,20α-diol) glucuronide. Maternal urine contains many other progesterone metabolites such as pregnanediol epimers, epimeric pregnanolones, 6α- and 16α-pregnanolones and, following 20-reduction, related pregnanetriols (Miller, 1998).

1.5.3. DHEA-S Production in FZ and Aromatization in Placenta

The foetal adrenal exclusively utilizes foetal cholesterol for steroid synthesis and in the FZ synthesis follows the Δ^5 pathway (Savchuk et al., 2017). Foetal adrenal and liver express high levels of SULT2A1 and as such many steroids including C-21 Δ^5 are sulfated (O'Shaughnessy et al., 2013). Extensive sulfation is likely protective, sparing the foetus from high levels of androgen precursors (Hill et al., 2014). Within the FZ, pregnenolone sulfate is subject to 17α-hydroxylation (CYP17A1) producing 17α-hydroxypregnenolone sulfate, which further yields abundant DHEA-sulfate (DHEAS) (Savchuk et al., 2017). While a sulfate pathway to DHEAS is emphasized here, to what extent this steroid is formed through either an unconjugated or sulfate pathway is still debated. The majority of DHEAS is 16α-hydroxylated in the foetal liver by CYP3A7 forming 16α-hydroxyDHEA sulfate (16αOHDHEAS) (O'Shaughnessy et al., 2013). This circulates to the placenta to be deconjugated by steroid sulfatase (Yamamoto et al.) and oxidised by HSD3B1 forming 16α-hydroxyandrostenedione. This is reduced

by HSD17B1 forming 16α-hydroxytestosterone which is aromatised by CYP19A1 to oestriol (Hill et al., 2010; Takeyama et al., 2000). A smaller fraction of DHEAS is aromatised to oestrone/oestradiol. During pregnancy, oestriol production increases >500-fold; oestriol glucuronide excretion can reach 100mg/day, mean 25mg/day, at least 90% formed from foetal precursors (Beischer et al., 1969).

1.5.4. TZ and DZ Produce Small Quantities of Downstream Corticoids

By the end of first trimester, the TZ and DZ express all the enzymes required for glucocorticoid and mineralocorticoid synthesis. There is a transient peak in foetal cortisol production during weeks 8-9 at the border of FZ and DZ associated with appearance of HSD3B2 (regulated by orphan receptor NGF1B) (Goto et al., 2006). This coincides with the period of sexual differentiation, possibly indicating a role for cortisol to inhibit ACTH-controlled androgen synthesis, thus protecting the female foetus (Goto et al., 2006). Cortisol synthesis then recedes before returning permanently around 20 weeks' gestation, most circulating as inactive cortisone due to HSD11B2 activity (Johnston et al., 2018; Naccache et al., 2016; Travers et al., 2018).

Deoxycorticosterone can be synthesised in the placenta from early gestation, probably from foetal 21-hydroxypregnenolone-sulfate (Guerami et al., 1988). Further downstream mineralocorticoid production in the foetal adrenal is debated. CYP11B1 is expressed primarily in the TZ, whereas CYP11B2 is scantily expressed in the DZ during the third trimester (Naccache et al., 2016). Small amounts of corticosterone and aldosterone may be produced towards the end of gestation but deoxycorticosterone,

while a weaker mineralocorticoid, remains throughout pregnancy the major mineralocorticoid produced (Johnston et al., 2018; Travers et al., 2018).

1.5.5. Active Androgen Production Occurs Primarily in Foetal Testes

The sulfate ester of DHEA, DHEAS, is the most abundantly produced steroid in the foetal compartment. At the end of the first trimester, androstenedione and testosterone derivatives are detectable in the foetal adrenal. Foetal testes begin to produce testosterone around 9 weeks post conception, and testicular 5α-reductase type 2 (SRD5A2) converts testosterone to DHT peripherally for male sexual differentiation. Precursors of DHT produced via the backdoor pathway and enzymes required for this pathway are expressed in foetal adrenal and testes (Fukami et al., 2013; Hill et al., 2014; Morel et al., 2016; Reisch et al., 2019; Savchuk et al., 2017).

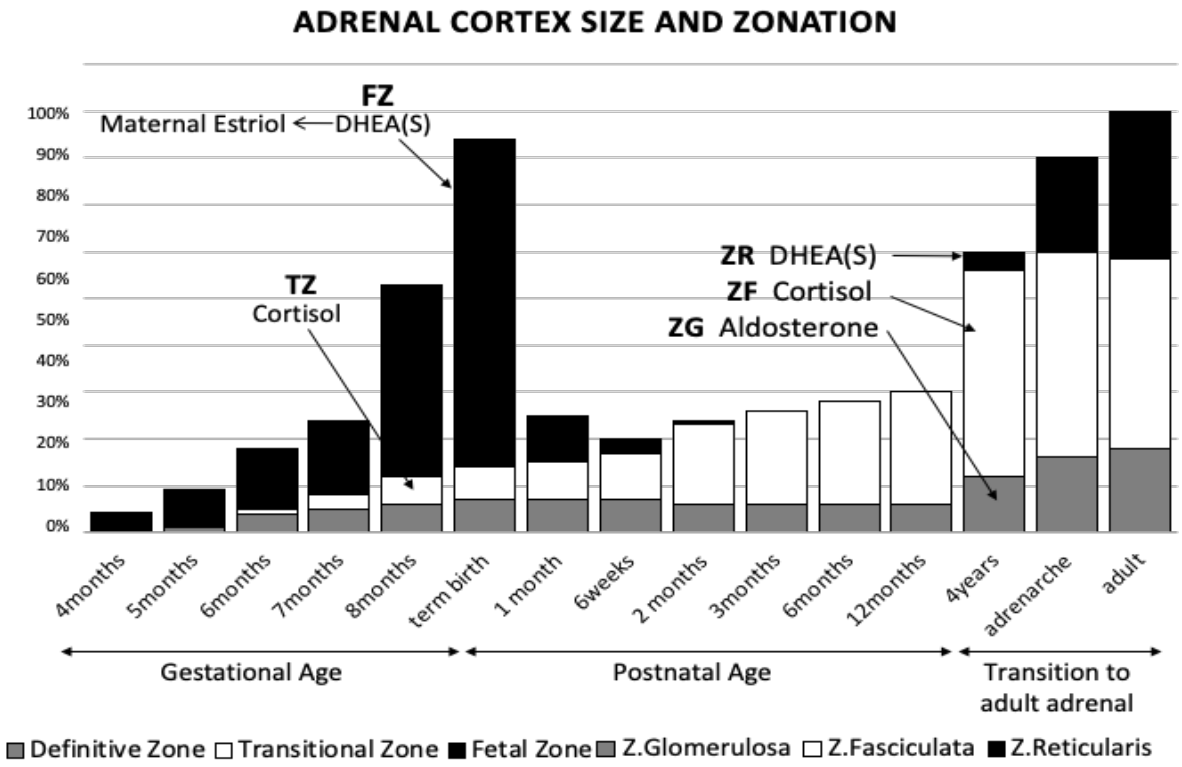


Figure 1.5: Approximate representation of relative size and zonation of the adrenal cortex as it develops through life, with indication of the major products of each zone

The foetal zone (FZ) constitutes the majority of the foetal adrenal cortex in utero. The size of the FZ is proportional to the output of its major product, DHEAS (converted to estriol in maternal compartment). The FZ involutes postnatally. The definitive and transitional zones (DZ and TZ) of the foetal adrenal postnatally develop into zona glomerulosa (ZG) and zona fasciculata (Diczfalusy) respectively. The zona reticularis (ZR) is apparent from approximately 4 years of age.

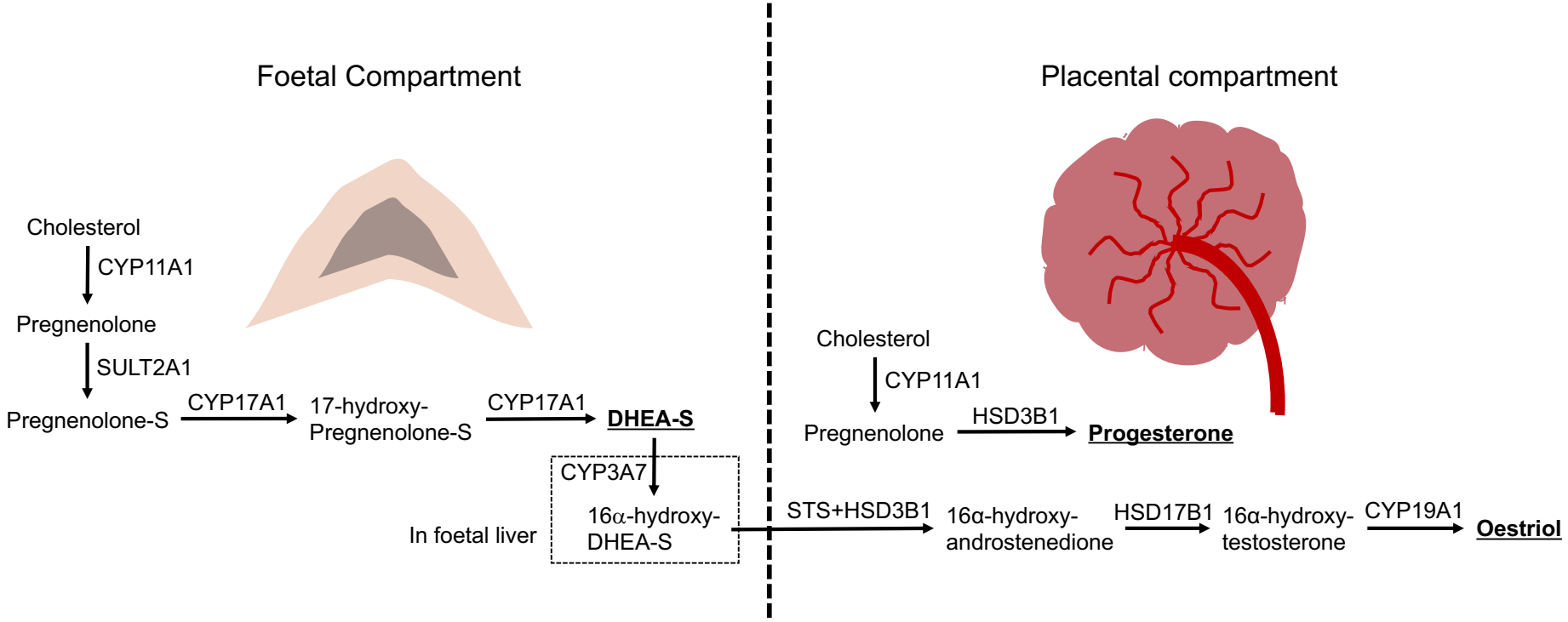


Figure 1.6: Synergistic functioning of the foeto-placental unit components

*The left panel illustrates the principal steroid production from cholesterol within the foetal compartment. DHEA-S (**bold**) is the main steroid hormone produced by the foetal adrenal*. The right panel illustrates the principal steroid hormones *(**bold**)* produced within the placental compartment; progesterone produced from cholesterol and oestriol produced from androgen precursors produced within the foetal compartment. Arrows indicate an enzymatic reaction, with the responsible enzyme stated above it.

1.5.6. Transition to Post-natal Adrenal Function- The Neonate and Young Infant

The neonatal urine steroid metabolome in the days after birth is an extension of foetoplacental and maternal metabolism- containing remnants of progesterone metabolites, oestrogens and the dominant FZ 3β-hydroxy-Δ^5 steroids (Shackleton et al., 1979). The transition from pre-natal to post-natal adrenal structure and function would typically be observed over the first 3-6 months of life, with the most dramatic transition expected to be observed within the first month of life and an adult pattern of steroid metabolism eventually reached by 1 year of age (Dhayat et al., 2015; Shackleton et al., 1979). A study evaluating adrenal gland size at autopsy found there were two phases of foetal zone involution: a rapid phase in the first two weeks of life followed by a slower involution phase from 3 weeks of age onwards (Bocian-Sobkowska, 2000). This involution of the foetal zone of the adrenal cortex zone is accompanied by a concurrent rapid reduction in DHEAS production (Grueters & Korth-Schutz, 1982; Kojima et al., 1981). In infants born prematurely, 3β-hydroxy-Δ^5 steroids persist for longer than in babies born at term, and the reduction in adrenal size is at a lower rate than observed in term infants, suggesting the involution of the foetal zone and the adrenal gland might be determined by gestation as well as post-natal age (Bocian-Sobkowska, 2000; Hata, Hata, et al., 1988; Hata, Nagata, et al., 1988; Midgley et al., 1998; Midgley et al., 1996).

A summary of neonatal steroid metabolism has been published by Shackleton et al (Shackleton et al., 1979). The major steroid group contains the 3β-hydroxy-Δ^5 steroids, and this group is dominated by 16α-hydroxyDHEA and its isomers 16β-hydroxyDHEA and 16-oxo-androstenediol. Other important steroids are the androstenetriolones and

androstenetetrols (Shackleton & Taylor, 1975) showing that additional hydroxylation at 15β- and 18- are significant in the newborn. Occasionally, 5-androstene-3β,16α,18-triol-17-one is the largest component of a neonate's metabolome. Hydroxylation at 15β- is also important in pregnene (and pregnane) metabolism (Caulfield et al., 2002; Kraan et al., 1993). With the involution of the FZ, the excretion of 3β-hydroxy-Δ^5 steroids reaches low levels by 3 months of age. Most Δ^5 steroids are excreted as monosulfates, exceptions being 21-hydroxypregnenolone, 16β-hydroxyDHEA, 5-pregnene-3β,20α-diol and the 5-androstene-3β,17-diols which are dominantly disulfates (Shackleton et al., 1979).

The complexity of neonatal corticosteroid metabolism was realised after cortisone acetate administration to a baby with 21-hydroxylase deficiency and identification of urinary metabolites (Taylor et al., 1978). Two features were evident. Firstly, metabolites were almost exclusively those with an 11-carbonyl (i.e., metabolites of cortisone). Reduced metabolites of cortisol like THF (tetrahydrocortisol) and 5αTHF and the cortols were minor. Secondly, the absence of 11β-hydroxylated metabolites was quantitatively made up by extensive 1β-and 6α-hydroxylation. Additional hydroxylation also applies to the metabolites of corticosterone, 6α-hydroxy THA is important with near absence of THB (tetrahydrocorticosterone) and 5αTHB during the first month of life. Production and excretion of the 1β-and 6α- hydroxylated metabolites is minimal by 3 months and has ceased by 6 months. Proportional reductive and oxidative 11β–HSD activity reaches mature levels by 6 months also (Derks & Drayer, 1978; Shackleton et al., 1979; Taylor et al., 1978).

Testicular androgen production declines after birth, but then becomes temporarily apparent again during the central gonadotrophin driven event, “mini-puberty”, which occurs approximately between 1 and 6 months of life. Mini-puberty is also observed in female infants to a lesser degree (Becker & Hesse, 2020; Forest et al., 1973).

1.5.7. Steroidogenesis from Infancy to Adulthood

Cortisol production seemingly increases in the first year of life if traditional non-polar cortisol metabolites are quantified from urine (Rogers et al., 2014). If body surface area is accounted for, however, cortisol production remains quite static throughout infancy, childhood and adolescence (Dhayat et al., 2015; Miller & Auchus, 2011). Plasma concentrations of other steroid hormones, including aldosterone, 17OHPreg, 17OHP, DOC and Androstenedione, gradually decline over the first 6-12 months of life (Dhayat et al., 2015; Garagorri et al., 2008; Shimozawa et al., 1988). In the first year of life, despite substantial changes occurring with age, there are fewer sex-specific differences observed in adrenal steroidogenesis (Dhayat et al., 2015; Homma et al., 2003; Malunowicz et al., 1997; Rogers et al., 2014). Testicular steroidogenesis declines following mini-puberty and remains quiescent until the time of true puberty. Following the rapid decline of adrenal DHEA and DHEAS production after birth, concentrations remain low until the onset of adrenarche, which is typically observed clinically from around 6 years of age. From 2 years of age, increases in serum DHEA, DHEA-S, Pregnenolone, 17OHPreg and 17OHP can be observed, becoming more pronounced over 4 years of age with accompanying increases in 17α–hydroxylase and 17,20 lyase activity and decreased HSD3B2 activity by 6 years (Kim et al., 2020). Focal islands of zona reticularis (ZR) begin to appear around 3 years of age and further

increases until a thin continuous ZR layer appears around 6 years of age, progressively increasing in depth until 13 years of age. From 5 years of age, CYP17A1, SULT2A1, POR and CYB5A expression in the zona reticularis become detectable and dramatically increase until reaching a plateau after 13 years of age (Suzuki et al., 2000). Progressive development of the continuous ZR layer positively correlates with DHEA and DHEAS production (Reiter et al., 1977). In additional to DHEA production, adrenarche is also associated with observed increasing concentrations of the urinary metabolites of androstenedione (Kelnar & Brook, 1983). Maximal concentrations of DHEA and DHEAS are observed between 20 and 30 years of age (Sulcova et al., 1997). Beyond the 3rd decade of life there is a decline in DHEA and DHEAS levels, falling by up to 80% by age 70 years in both men and women, with a higher rate of decline up to 60 years of age and lower rate of decline beyond that (Labrie et al., 1997; Sulcova et al., 1997). Expression of CYB5A similarly decreases with age in both sexes, reflecting the decline in androgen production (Tezuka et al., 2021). The phenomenon of adrenal androgen precursor decline is termed adrenopause. The ZR reduces in size following the process of adrenopause. Similarly, the zona glomerulosa (ZG) has also been found to undergo functional changes with aging, with a reduction in the area expressing CYP11B2 and increased autonomous aldosterone producing cell clusters. The reduction in CYP11B2 expression being more marked in men than women (Nanba et al., 2017; Tezuka et al., 2021). Morphologically, the zona fasciculata (Diczfalusy) increases in size with age, in stark contrast to the observed decline in the ZG and ZR (Nonaka et al., 2020; Tezuka et al., 2021). Early morning serum concentrations of cortisol and mineralocorticoid intermediates corticosterone and 11-deoxycorticosterone showed no significant age variation in one study (Tezuka et al., 2021). However, another study described a significant age-

related decline in all plasma steroid hormones apart from 21-deoxycortisol, with the strongest negative correlations with advancing age observed for pregnenolone, DHEA/S, cortisone, 11-deoxycortisone, androstenedione, 17OHP and corticosterone (Eisenhofer et al., 2017). Other studies aiming to assess the effect of aging on cortisol production/exposure have produced similarly discrepant results (Feldman et al., 2002; Lupien et al., 1996; Moffat et al., 2020; Seeman et al., 1997) but it is likely this is due to variation in the methods employed to assess cortisol response. It seems with age there is an attenuated morning rise in glucocorticoid concentrations but with an accompanying higher nocturnal nadir leading to a flattening of the usual daily circadian rhythm and an overall increase in the total daily production of cortisol (Yiallouris et al., 2019). Eisenhofer et al (Eisenhofer et al., 2017), in addition to age related changes, also evaluated sex related differences in early morning plasma steroid hormone concentrations. In adult women, plasma concentrations of progesterone vary according to menstrual cycle and are higher than observed in adult men. Plasma concentrations of mineralocorticoid steroids (aldosterone, corticosterone and 11-deoxycorticosterone) show no significant difference between men and women. Similarly, the upstream steroid hormone precursor pregnenolone also showed no significant sex related difference. Other glucocorticoid, androgen and respective precursors are produced in higher concentrations in men than women. Progesterone and Testosterone concentrations showed the most marked sex-related differences. Interestingly, while androgens for the majority show a decline with age, 11-oxygenated androgens have been shown not to decline with age in women (Nanba et al., 2019). 11-oxygenated androgen precursors 11-hydroxyandrostenedione and 11-hydroxytestosterone remain stable for both men and women with age, whereas the

potently active 11-oxygenated androgen 11-keto-T has been shown to decline in men but not women (Davio et al., 2020).

1.6. Inborn Disorders of Steroidogenesis

Congenital disorders affecting the process of steroid hormone production, through the inheritance of a genetic mutation affecting one of a critical group of enzymes, are otherwise termed Inborn disorders of steroidogenesis. They comprise two main groups of disorders: congenital adrenal hyperplasia (Hargitai et al.), which manifests with adrenal insufficiency with or without differences in sex development (Engels et al.), or disorders which manifest with DSD in isolation. They primarily present in childhood and require rapid diagnosis to avoid high mortality from adrenal insufficiency, and to support timely gender assignment in DSD. However, these disorders can present throughout the life course and to a variety of health care professionals so both paediatric and adult specialities are required to be vigilant. At birth, there may be overt or suspected genital ambiguity (an apparently phenotypic male with bilateral undescended testes or apparently phenotypic female with palpable gonad/s, for example). In the early neonatal period, patients can present with salt wasting adrenal crisis. Patients can present in childhood with premature adrenarche/pubarche, or in adolescence with delayed puberty or rapid virilisation. In adulthood, patients can present with menstrual abnormalities, clinical signs of androgen excess, or with infertility.

1.6.1. Current Diagnostic Pathway

Patients presenting with symptoms suggestive of an inborn disorder of steroidogenesis should be assessed by suitably experienced medical professionals. While practice varies across the world, expert consensus statements support assessment utilising a multidisciplinary team approach including (but not limited to) endocrinologists, surgeons, geneticists, and psychologists. A comprehensive suggested approach to clinical assessment of a child presenting with DSD is outlined in the 2016 consensus statement by Lee et al (Lee et al., 2016). Typically, for both CAH and DSD, diagnosis is sought both biochemically and genetically where possible. There have been massive developments in molecular genetics in recent years which are rapidly being translated into routine clinical practice. Specific recommendations regard genetic testing in the context of DSD are discussed in a publication by Audi L et al from 2018 (Audi et al., 2018). Biochemical assessment provides functional information to assist with diagnosis and is made through measurement of serum hormone levels or by measuring urinary steroid hormone metabolites. For serum hormone measurements, dynamic function testing is often helpful to test the relevant hormonal axis, where hormones are variably secreted across a day or where hormones are normally secreted in very low concentrations. Urinary steroid metabolite profiling is a valuable investigative tool, particularly for the elucidation of enzymatic disorders, as it provides a comprehensive description of an individual's entire steroidogenic function. However, access to this technique is widely variable across Europe and the world. Relatively few centres exist with sufficient expertise to run and analyse samples, as such, demand and cost are high. Steroid hormone analysis in the context of DSD has been reviewed by Kulle et al (Kulle et al., 2017).

1.6.2. Congenital Adrenal Hyperplasia

Congenital Adrenal Hyperplasia (Hargitai et al.) comprises a group of diseases resulting from inherited inactivating genetic mutations in an enzyme or co-factor required for adrenal steroidogenesis. They are inherited in an autosomal recessive manner and characteristically result in defective cortisol biosynthesis with variable effects on the other steroid hormone pathways depending on the site of the defective enzyme within the steroidogenesis pathway. By means of a negative feedback loop between the adrenal gland and the hypothalamus and pituitary, insufficient circulating cortisol prompts the release of Corticotrophin Releasing Hormone (CRH) and ACTH which in turn stimulates the adrenal cortex and the steroidogenesis pathway. Due to the impaired mutant enzyme resulting in an “enzymatic block” at a particular point in the pathway, this overstimulation of the adrenal cortex leads to an accumulation of steroid hormone precursors and increased production of any unaffected arm of the steroidogenesis pathway. Chronic overstimulation by ACTH also causes hypertrophy of the adrenal cortex, a finding which led to the naming of the condition. Steroidogenic enzymes implicated in the pathophysiology of CAH include 21-hydroxylase (CYP21A2), 11β-hydroxylase (CYP11B1), 17α-hydroxylase/17,20 lyase (CYP17A1), and 3β-hydroxysteroid dehydrogenase type 2 (HSD3B2). Mutations in POR, an electron donor enzyme serving CYP21A2 and CYP17A1, cause the most recently identified variant of CAH (Baranowski et al., 2018).

1.6.2.1. 21-hydroxylase Deficiency

First described by DeCreccio in 1865, the most common cause of CAH is due to 21-hydroxylase deficiency (21OHD), which accounts for 90-95% of cases (Claahsen-van der Grinten et al., 2022; El-Maouche et al., 2017). The majority of 46,XX individuals who present with genital ambiguity will have a diagnosis of CAH secondary to 21OHD. Data from newborn screening would suggest that the incidence of the classical form is 1 in 14,000 to 1 in 18,000 live births but this varies according to ethnicity and location from approximately 1 in 6000 to 1 in 27,000 (Baranowski et al., 2018; Claahsen-van der Grinten et al., 2022; Krone & Arlt, 2009). Impaired CYP21A2 enzyme activity leads to reduced or absent mineralocorticoid and glucocorticoid production. Hypothalamic-Pituitary-Adrenal (HPA) axis activation via negative feedback leads to increasing ACTH production, which results in concurrent androgen excess as the only unaffected arm of the adrenal steroidogenesis pathway. There is accumulation of precursor steroids prior to the enzymatic block, most notably 17OHP, which is used as the primary serum marker steroid for diagnosis and neonatal screening (Speiser et al., 2010; Therrell et al., 1998). 17OHPreg, Progesterone and Pregnenolone are also found to be raised as upstream precursor steroids (Miller & Auchus, 2011). CYP11B1 converts 17OHP to 21-deoxycortisol in the absence of 21-hydroxylase. This has been hailed as an additional characteristic serum biomarker of the condition and measurement has been proposed to improve diagnostic accuracy of screening (Held et al., 2022; Miller, 2019). All 3 pathways to active androgen synthesis (“Classical”, “Alternative”, and “11-oxygenated pathways”) have been described as upregulated due to the accumulation of relevant precursor steroids feeding all 3 pathways (Jones et al., 2017; Kamrath et al., 2012; Turcu et al., 2016) **(Figure 1.4).**

The urine steroid profile classically shows marked reduction in cortisol metabolite concentrations, the degree of which depends on whether the mutation results in an abolished or partially functional enzyme. Downstream mineralocorticoid steroid metabolites are also greatly diminished. There is marked elevation of 17α-hydroxyprogesterone metabolites, 17α-hydroxypregnanolone (17HP) and pregnanetriol (PT). There is also marked increase of the 21-deoxycortisol metabolite, pregnanetriolone (PTONE). Ratios of 17α-hydroxyprogesterone metabolites 17HP, PT and PTONE over combinations of glucocorticoid metabolites (such as THE (Tetrahydrocortisone), THF, 5αTHF) have been proposed as diagnostically helpful and shown to be abnormally raised even in late onset forms (Shackleton et al., 1986; Storbeck et al., 2019). Metabolites from all 3 pathways to active androgen synthesis have been found to be raised: androsterone and etiocholanolone from the classical pathway to DHT synthesis, 5α-17-hydroxypregnanolone (3α5α17HP) representing the "alternative pathway" to DHT synthesis and 11β-hydroxyandrosterone (11βOHAn) representing the 11-oxygenated androgen pathway (Jones et al., 2017; Kamrath et al., 2012; Turcu et al., 2016). In neonates, diagnosis can be made from day 1 of life by assessing their urinary steroid metabolome. PTONE and 17HP will be elevated, but PT is less informative in early neonatal life (Wudy et al., 2000). There is raised 16α-hydroxypregnenolone (Homoki & Teller, 1982). This can be particularly emphasised when taken in proportion to 16α-hydroxyDHEA; usually the two metabolites are excreted in equal amounts in healthy babies, but in babies with 21-hydroxylase deficiency 16α-hydroxy-pregnenolone is at least 3 times proportionally higher (Caulfield et al., 2002; Shackleton, 1986). There is also a high excretion of 3α,15β,17α-trihydroxy-5β-pregnan-20-one (15β-triolone) (Joannou, 1981;

Shackleton, 1986). Nowadays, urinary steroid profiling is rarely employed for classical disease due to the easy availability of rapid serum measurements of 17OHP. However, it remains a useful investigation for non-classical or late onset forms of disease.

CYP21A2 and its non-functional but highly homologous pseudogene (CYP21A1P) are located close together in the HLA class III region on the short arm of chromosome 6 (White et al., 1986). Approximately 95% of disease-causing mutations in 21-hydroxylase deficiency are variants, point mutations, or deletions caused by meiotic recombination events between CYP21A2 and CYP21A1P. Approximately 1% of inactivating mutations in CYP21A2 arise de novo. Nearly 300 mutations have been described (www.hgmd.cf.ac.uk), the majority being missense/nonsense type. Most CAH patients (65–75%) are found to have compound heterozygous mutations with the clinical phenotype determined by the more productive allele (Krone & Arlt, 2009). Depending on the degree of loss-of-function caused by a distinct mutation, patients present with either isolated glucocorticoid or combined glucocorticoid and mineralocorticoid deficiency. In childhood, the classical forms of 21-hydroxylase deficiency are accordingly differentiated into “salt-wasting” CAH (75% of cases; glucocorticoid + mineralocorticoid deficiency) and “simple virilizing” (25% of cases; glucocorticoid deficiency only) (Claahsen-van der Grinten et al., 2022). Complete inactivating mutations result in the salt-losing phenotype, whereas mutations which result in an enzyme with some residual function result in the simple virilising phenotype. As little as 1–2% of 21-hydroxylase function is sufficient to manifest with the simple virilizing phenotype (Krone & Arlt, 2009).

In the absence of neonatal screening, the salt-wasting form presents with life-threatening adrenal “salt-losing” crisis within the first 2 weeks of life and ambiguous genitalia in affected 46,XX babies from birth (46,XX DSD). Deficiency of cortisol affects the growth and subsequent function of the adrenal medulla. Reduced production of epinephrine and metanephrine exacerbates the potential for hypoglycaemia during adrenal crisis (Merke et al., 2000). The simple virilizing form presents with symptoms of androgen excess in isolation. Genetic females present with DSD from birth. Genetic male individuals affected by the simple virilizing form may present with premature pubic/axillary hair growth or precocious pseudo-puberty in early childhood and rapid skeletal growth with advanced bone age (El-Maouche et al., 2017).

Treatment requires glucocorticoid replacement, usually with hydrocortisone or alternatively prednisolone in adults. If deficient, treatment also consists of mineralocorticoid replacement with fludrocortisone, and additional sodium supplementation through infancy. Glucocorticoid dosing must be increased during times of physical stress/illness (stress dose cover) to mimic the normal physiological stress response. As with other virilizing forms of CAH, the aim is not only to replace for GC deficiency and prevent adrenal crisis but also to “switch off” adrenal androgen production by reducing ACTH drive. Inadequate dosing results in increased risk of adrenal crisis and inadequate androgen suppression, the former being a leading cause of mortality in CAH and the latter leading to rapid premature skeletal growth and reduced height potential in affected children (Claahsen-van der Grinten et al., 2022; Falhammar et al., 2014; Hargitai et al., 2001; Idkowiak, Lavery, et al., 2011). Women with hyperandrogenism secondary to insufficient glucocorticoid replacement can be affected by menstrual irregularities and chronic anovulation. Male patients can have

compromised fertility due to hypothalamic-pituitary-gonadal axis suppression by chronically raised androgens. Men can also develop testicular adrenal rest tumours, which may be related to chronically raised ACTH levels and can compromise spermatogenesis (Arlt et al., 2010; Engels et al., 2018). However, attempts to completely normalize biochemical marker levels frequently result in overtreatment. Patients run the risk of developing Cushingoid side effects, impacting on growth and metabolic health. In the long term, patients have an increased risk of cardiovascular disease and metabolic syndrome. Fertility problems are frequent in both sexes, related to both under- and overtreatment with glucocorticoids (Arlt et al., 2010; Claahsen-van der Grinten et al., 2022).

1.6.2.2. 11β-hydroxylase deficiency

First technically described as a cause of CAH by Eberlein and Bongiovanni in 1956 (Bongiovanni & Eberlein, 1956)through assessment of plasma and urinary corticosteroid measurements, though the association of CAH and hypertension had been made earlier in the 50s in several case reports (Shepard & Clausen, 1951; Wilkins et al., 1952). 11β-hydroxylase deficiency (11βOHD) is the second most common cause of CAH, accounting for 2–5% of reported cases of European ancestry, with an overall incidence of 1 in 100,000–200,000 live births (Krone & Arlt, 2009). The clinical phenotype consists of glucocorticoid deficiency with concurrent symptoms of mineralocorticoid excess and androgen excess. Most cases of 11βOHD present with a classical phenotype with the hallmark features of virilisation and hypertension (White et al., 1994). It is caused by inactivating mutations in the CYP11B1 gene. Due to founder mutations the condition is more common in Israel among Jewish immigrants

from Morocco (Bulsari & Falhammar, 2017; Rosler et al., 1992). The CYP11B1 gene is located on the long arm of chromosome 8 and in very close proximity to the highly homologous CYP11B2 gene. Recombination events have been reported where the CYP11B1 gene is under control of the CYP11B2 promoter, which responds to angiotensin II and not ACTH, resulting clinically in classic CYP11B1 deficiency (Krone & Arlt, 2009). More than 140 mutations have been reported to date in the CYP11B1 gene (www.hgmd.cf.ac.uk). There is no mutational hotspot, and mutations have been found along the whole gene, most being missense mutations significantly reducing enzymatic activity.

CYP11B1 catalyses the conversion of 11-deoxycortisol to cortisol and of DOC to corticosterone in the zona fasciculata. Reduced cortisol concentrations in CYP11B1 deficiency leads to HPA axis activation via a negative feedback mechanism. Under the influence of ACTH, concentrations of DOC increase markedly, which acts as an agonist to the mineralocorticoid receptor, leading to hypertension, suppression of the RAAS and low aldosterone concentration despite the ability of the zona glomerulosa to produce aldosterone. Hypertension is apparent in 2/3 of patients at the time of presentation (Bulsari & Falhammar, 2017; Rosler et al., 1992). It might not become apparent during the neonatal period due to the renal mineralocorticoid resistance, and some newborns may even present with salt loss (Holcombe et al., 1980). Hypokalaemia is rarely observed at diagnosis. Androgen excess can be severe, presenting with 46,XX DSD from birth and precocious pseudo-puberty in males (Rosler et al., 1982).

Non-classic forms of 11βOHD are rare and have a similar phenotype to non-classic 21OHD, exhibiting signs of androgen excess like precocious pseudo-puberty or symptoms suggestive of polycystic ovary syndrome in post-pubertal females. Arterial hypertension is not commonly found early in the non-classic form but can develop later in life (Joehrer et al., 1997; Peters et al., 2007; Reisch, Hogler, et al., 2013).

Diagnostically, the classical marker steroid is 11-deoxycortisol, which accumulates in great excess prior to the enzymatic block. DOC is also measurably raised, along with progesterone as an upstream precursor steroid from this. Active androgens and androgen precursors are observed to be raised, but characteristically not 11-oxygenated androgens (Caulfield et al., 2002; Shackleton, 1986; Storbeck et al., 2019). As 17OHP can accumulate prior to the more pronounced increase in 11-deoxycortisol, some babies in newborn screening programs have been wrongly diagnosed as suffering from 21OHD (Peter et al., 2008).

The urinary steroid metabolome is dominated by huge concentrations of tetrahydro-11-deoxycortisol (Watson et al.). TetrahydroDOC (THDOC) is classically raised. Androgen and androgen precursor metabolites (androsterone (An), etiocholanolone (Et), DHEA and 16α–OHDHEA) are also significantly raised. Cortisol metabolites are significantly diminished. There is an absence of 11-oxygenated steroids, notably the 11-hydroxyandrostenedione metabolite (11βOHAn) is low in urine. Two diagnostic urinary steroid metabolome ratios have been proposed – androgen metabolites over 11β-hydroxylated metabolites ((An+Et)/(11βOHAn+11βOHEt)) and 11-deoxycortisol over glucocorticoid metabolites (THS/THE+THF+5αTHF). PTONE is low as 11β–hydroxylase is required to catalyse the conversion of 21-deoxycortisol from

17OHP. In neonates, THS and 6α-hydroxytetrahydrotetrahydro-11-deoxycortisol (6α-hydroxyTHS) are significantly raised in babies due to active 6α-hydroxylase in the neonatal period. In healthy newborns, neither of these metabolites is usually measurable due to normally low abundance. Raised 3β-hydroxy-5-ene steroids are also observed (Caulfield et al., 2002; Honour, 2018; Shackleton, 1986; Storbeck et al., 2019).

Treatment Is with glucocorticoid replacement in sufficient doses to suppress the negative feedback from the HPA axis which is driving androgen and mineralocorticoid agonist excess. Hypertension may require a mineralocorticoid receptor antagonist such as spironolactone to treat.

1.6.2.3. P450 Oxidoreductase Deficiency

P450 oxidoreductase (POR) is required as a cofactor for electron donation for all microsomal (type 2) CYP enzymes, including steroidogenic enzymes CYP21A2, CYP17A1 and CYP19A1. The urinary steroid metabolome for P450 oxidoreductase deficiency (PORD) was first described by Peterson R.E. et al in 1985, described as "apparent combined 17α-hydroxylase and 21-hydroxylase deficiency" (Peterson et al., 1985a). The molecular basis for the disease was not fully elucidated until 2004 through the simultaneous efforts of Fluck C et al, Arlt W et al and Adachi et al (Adachi et al., 2004; Fluck et al., 2004; Miller et al., 2004). The exact incidence is unknown, but PORD is a rare CAH variant, with approximately 88 mutations reported to date (www.hgmd.cf.ac.uk). The POR gene is located on the long arm of chromosome 7. There are 2 mutations which are frequently found in distinct ethnic groups: the

p.A287P mutation is most common in White Caucasians, whereas the p.R457H mutation is most commonly found in individuals of Asian ancestry (Krone et al., 2012).

Most affected patients will have some degree of cortisol deficiency, with 90% requiring at least stress related glucocorticoid replacement. Therefore, all affected patients should undergo cosyntropin stimulation testing, regardless of presence or absence of DSD (Krone et al., 2012). Approximately 50% of those who require glucocorticoid treatment will need daily hydrocortisone replacement, with the other 50% requiring replacement only during periods of stress. Mineralocorticoid biosynthesis is unaffected or sometimes observed to be increased with hypertension described in affected adults. 75% of patients will have some degree of DSD, and this has been reported in both sexes: 46,XY babies are often undermasculinized and 46,XX babies are born virilized without progression of virilization postnatally (Krone et al., 2012). This phenomenon is explained by the presence of the alternative pathway to 5α-DHT synthesis, active during foetal life (Reisch et al., 2019). Postnatally, activity of the pathway ceases, so there is sex steroid deficiency in both sexes. In puberty, both sexes can present with delayed development of sexual characteristics, in particular females often develop significant hypergonadotropic hypogonadism and severe ovarian cysts (Fukami et al., 2009; Idkowiak, O'Riordan, et al., 2011). Expectant mothers carrying affected foetuses can experience virilization during pregnancy, which characteristically resolves postpartum; explained by the production of 5α-reduced androgens via the alternative pathway by the foetus, that cross into maternal circulation unable to be aromatized by the placenta (Arlt et al., 2004; Fluck et al., 2004).

PORD is a multi-system disorder, as it affects all microsomal CYP enzymes that require POR for function. A striking clinical finding in many affected POR-deficient patients are skeletal abnormalities, which have been described as part of the Antley-Bixler syndrome phenotype. Typical malformations include large joint synostosis (for example, radio-humeral), congenital bowing of the femurs, and hand and foot malformations such as long palms, camptodactyly, and rocker bottom feet. Craniofacial anomalies are frequent, such as craniosynostosis with midface hypoplasia, most commonly involving the coronal and lambdoid sutures, which can in severe forms be complicated by hydrocephalus requiring surgical intervention (Krone et al., 2012). The number and severity of associated malformation features form the basis of a scoring system proposed to classify the severity of Antley-Bixler syndrome phenotype (Krone et al., 2012). It is thought that the molecular basis of the malformation phenotype is the disruption of the activity CYP enzymes involved in sterol synthesis and retinoic acid metabolism, both of which have been shown to cause a skeletal malformation phenotype in murine knockout models (Ribes et al., 2007). In addition, loss of POR function affects key CYP enzymes involved in hepatic phase 1 metabolism, including CYP1A2, CYP2C9, CYP2C19, CYP2D6, and CYP3A4 (Nebert & Russell, 2002). CYP3A4 metabolizes oestrogens and glucocorticoids, which is important for treatment considerations in patients with PORD as the metabolic clearance of exogenous hormone replacement is expected to be reduced, thus potentially exposing patients to higher glucocorticoid and oestrogen levels than intended. Various studies have assessed the impact of certain POR mutations on POR-dependent hepatic cytochromes in vitro (Burkhard et al., 2017; Pandey & Sproll, 2014), with one detailed in vivo phenotyping study confirming altered drug detoxification for a variety of substances (Tomalik-Scharte et al., 2010).

Phenotypic expression of a distinct genotype can be variable, but mutations causing 46,XX DSD usually present with normal genital phenotype in 46,XY individuals and vice versa; similarly, the severity of the skeletal malformation phenotype is increased in carriers of major loss-of-function mutations. With severe phenotypes, there have been reported cases of stillbirth and early neonatal death (Krone et al., 2012; Reisch, Idkowiak, et al., 2013).

Pregnenolone and Progesterone are characteristically observed to be raised in affected individuals, with 17OHP raised to a lesser extent. The urinary steroid metabolome and diagnosis of PORD can usually be readily established by urinary steroid profiling by gas chromatography-mass spectrometry (Krone et al., 2010; Krone et al., 2012; Shackleton et al., 2004; Storbeck et al., 2019). This reveals the hallmark feature of combined impairment of CYP21A2 and CYP17A1 activities. There is high excretion of 17OHP and 21-deoxycortisol metabolites, 17HP, PT and PTONE, reflecting 21-hydroxylase deficiency. There is also concurrent high excretion of corticosterone metabolites, tetrahydro-11-dehydrocorticosterone, tetrahydrocorticosterone and 5α– reduced counterparts (THA, 5αTHA, THB, 5αTHB) as seen in 17α-hydroxylase deficiency. Upstream steroid precursor metabolites from progesterone and pregnenolone are characteristically markedly raised (pregnanediol (PD) and pregnenediol (5PD) respectively). Cortisol metabolites are usually measurable and may be within normal limits. In the neonatal period, babies affected by PORD will show high excretion of 5PD (though predominantly seen as pregnadienol which is an artefact of pregnanediol disulfate), 5β-pregnane-3α,17α-diol-20-one, its 5α-reduced counterpart and PTONE. The conventional neonatal steroid 16-

hydroxyDHEA and 16-hydroxypregnenolone are virtually absent because of PORD attenuation of 17,20 lyase and 16-hydroxylase. Over the first months of life, excretion of 5PD increases and the emergence of raised corticosterone metabolites begins to be seen. Metabolite ratios proposed to assist in diagnosis include those attributed for diagnosis of 21OHD and 17α-hydroxylase deficiency and also ratios of Pregnenolone and Progesterone over glucocorticoid metabolites and 17OHP over androgen metabolites).

Prenatal testing is possible using maternal urinary steroid profiling from 12 weeks onwards. In mid-pregnancy, diagnosis can also be made by measuring unconjugated serum oestriol, which will be characteristically low due to several enzyme deficiencies which require POR for oestrogen production. This is in stark contrast to the huge concentrations observed in unaffected healthy pregnancies (Reisch, Idkowiak, et al., 2013). Molecular genetic techniques can additionally be employed and will confirm the diagnosis.

1.6.2.4. 17α-hydroxylase deficiency

First described by Biglieri et al in 1966 (Biglieri et al., 1966), CAH due to 17α-hydroxylase deficiency (17αOHD) is rare, accounting for 1% of CAH cases (Krone & Arlt, 2009). Deficiency affects both glucocorticoid and androgen production, with mineralocorticoid production unaffected. Therefore, negative feedback via the HPA axis leads to mineralocorticoid excess. 17αOHD is caused by inactivating mutations in the CYP17A1 gene located on the long arm of chromosome 10. Inactivating mutations have been located along the whole length of the gene, without any distinct

hot spot, although N-terminal mutations are found more frequently. To date, just over 120 mutations have been reported, the majority of which are missense/nonsense mutations (www.hgmd.cf.ac.uk). There are some frequently occurring mutations which correlate with ethnicity. In particular, a small 4bp insertion in exon 8 in Dutch Friedlaenders, a 3 amino acid in-frame deletion of exon 8 in Southeast Asia, the p.F53/54del mutation, and, within the Brazilian population, the missense mutations p.W406R and p.R362C (Auchus, 2017; Costa-Santos et al., 2004; Sun et al., 2021).

Despite affected individuals being profoundly glucocorticoid deficient, they do not classically present with symptoms of adrenal insufficiency or adrenal crisis. This is due to the affinity of mineralocorticoids and precursors to the glucocorticoid receptor acting as glucocorticoid agonists. Partial adrenal insufficiency might become evident with cosyntropin testing. The accumulation of DOC, enhanced by the activation of the hypothalamic-pituitary- adrenal axis with some regulatory effect on the mineralocorticoid pathway, results in suppression of the RAAS and aldosterone production with biochemical findings of hypernatremia, hypokalaemia, hyporeninaemia, and arterial hypertension. 46,XX patients affected by CYP17A1 deficiency tend to present at early pubertal age, due to lack of pubertal development including primary amenorrhea, with marked hypergonadotropic hypogonadism. In contrast, 46,XY patients usually present with DSD, having undervirilized external genital from birth. Low renin hypertension is often an incidental finding at presentation. Serum androgens and androgen precursors are characteristically low on measurement and do not increase with cosyntropin testing (Sun et al., 2021).

Milder forms with some residual enzyme activity have been reported; whilst those disrupt 17,20 lyase activity, the impact on the enzyme's 17α-hydroxylase activity can be milder. Such cases, sometimes labelled as "isolated 17,20 lyase deficiency," are extremely rare, and patients present with isolated androgen deficiency. However, their capacity to produce sufficient amounts of glucocorticoids is attenuated, indicating some degree of 17α-hydroxylase impairment (Geller et al., 1997; Sherbet et al., 2003).

The observed urinary steroid metabolome for 17αOHD was first characterized in 1978 (Honour et al., 1978). Low cortisol metabolite excretion accompanies low excretion of androgen and androgen precursor metabolites, androsterone, etiocholanolone and DHEA. There is significantly raised corticosterone metabolite excretion, namely tetrahydro-11-dehydrocorticosterone, tetrahydrocorticosterone and 5α- reduced counterparts (THA, 5αTHA, THB, 5αTHB). The 11-deoxycorticosterone (DOC) metabolite tetrahydrodeoxycorticosterone (THDOC) is substantially raised. Progesterone and Pregnenolone metabolites, pregnanediol (Lee et al.) and pregnenediol (5PD), are usually mildly raised (Storbeck et al., 2019). Neonatal mineralocorticoid excretion is substantially different from older children and adults. 11-dehydrocorticosterone is a predominant mineralocorticoid which is increased in affected neonates. Metabolites THA, 5αTHA and neonate-specific 6α-hydroxy 11-dehydro-tetrahydrocorticosterone (6αOHTHA), are notably raised. C19 steroids are low, but there is not significant neonatal production of androsterone or etiocholanolone for a difference to be evaluated at this age, whereas 16α-hydroxyDHEA is usually produced in abundance so low excretion makes for a better neonatal marker. It is noteworthy, however, that diagnosis is very rarely made in the neonatal period as

patients rarely present with any symptoms at this age (Caulfield et al., 2002; Shackleton, 1986).

Treatment involves anti-hypertensives, usually mineralocorticoid receptor antagonists such as spironolactone, coupled with conservative glucocorticoid replacement. The aim of glucocorticoid replacement is to attenuate the negative feedback effect which is driving mineralocorticoid excess, and thus improve biochemistry and blood pressure. Replacement is conservative to avoid long term side effects from glucocorticoid replacement. This is a safe strategy as relatively high levels of DOC will protect against adrenal crisis through glucocorticoid receptor agonism, and concurrent mineralocorticoid receptor antagonist treatment will help to treat hypertension. At the time of puberty, hormonal induction with exogenous testosterone for males and oestrogen for females is usually required.

1.6.2.5. 3β-Hydroxysteroid Dehydrogenase Type 2 Deficiency

3β-hydroxysteroid dehydrogenase type 2 deficiency (3βHSD2D) is a very rare form of CAH, affecting <0.5% of CAH cases, and usually presenting in early infancy (Al Alawi et al., 2019; Bongiovanni, 1962). HSD3B2 is at a critical branch point in steroidogenesis, gating entrance to all three steroid pathways (**Figure 1.3)**. Deficiency results in insufficient production of all three types of adrenal steroid hormone and accumulation of upstream steroid precursors 17OHPreg and DHEA in response to negative feedback stimulating the HPA axis (Storbeck et al., 2019). The two isoforms of 3β-hydroxysteroid dehydrogenase are encoded by genes located in close proximity to each other on the short arm of chromosome 1 (HSD3B1 and HSD3B2). HSD3B2 is

expressed in adrenals and gonads, whilst HSD3B1 is expressed in multiple peripheral tissues (Krone & Arlt, 2009). More than 60 mutations have been described (www.hgmd.cf.ac.uk), and good genotype-phenotype correlations exists with regards to mineralocorticoid deficiency, where severe loss-of-function mutations predict neonatal salt-wasting.

Classically, affected patients present in infancy with salt-wasting adrenal crisis and high-renin hypotension, as in 21OHD. However, both sexes can present with DSD: in 46,XY babies, deficiency of adrenal and gonadal androgens result in under masculinization. The observation that 46,XX babies can present with virilized genitalia is explained by the compensatory action of HSD3B1 expressed in placenta and peripheral tissues; HSD3B1 facilitates downstream conversion of accumulating adrenal DHEA to androstenedione in the periphery, causing androgen excess. Apparently, this is sufficient to virilize an 46,XX baby, but would not compensate for the loss of gonadal HSD3B2 activity in 46,XY babies (Al Alawi et al., 2019; Rheaume et al., 1992).

There is broad phenotypic variation, and presentation can vary from severe salt-wasting forms to non-salt wasting with variable degrees of virilization. This may include late-onset forms manifesting with hirsutism and irregular periods resembling a polycystic ovary syndrome phenotype in patients with confirmed pathogenic mutations. In addition, a clinical and biochemical phenotype exists, where androgen excess is present with predominance of Δ^5 steroids, but no genetic abnormalities were detected. The molecular basis of this condition, often referred to as “functional

HSD3B2 deficiency," is currently unknown (Al Alawi et al., 2019; Lutfallah et al., 2002; Mermejo et al., 2005).

The presence of the HSD3B1 isoform can make this diagnosis difficult: HSD3B1 peripherally converts 17OHPreg, which accumulates before the block, into 17OHP, which can be increased in newborns with 3βHSD2D (Balsamo et al., 2020). There have been reports of infants being wrongly diagnosed with 21OHD following initial testing revealing a raised 17OHP. The overall steroid precursor constellation in this condition is a predominance of Δ^5 steroids (i.e. pregnenolone, 17OHPreg, and DHEA) over Δ^4 steroids (progesterone, 17OHP, and androstenedione); this ratio is further exaggerated after ACTH stimulation (Storbeck et al., 2019).

Urinary steroid profiling is similarly accurate and less invasive to establish the diagnosis. The steroid metabolome findings in 3βHSD2D were first described by Bongiovanni (Bongiovanni, 1962) when the disease was characterized. There is increased excretion of 3β-hydroxy-5-ene steroids noted and markedly raised 17-hydroxypregnenolone metabolite, 5-pregnene-3β,17α,20α-triol (5PT), and DHEA. PT and 17HP are raised, similar to 21OHD, but PTONE is low. Cortisol metabolite excretion is low, but the degree to which depends on whether the enzymatic defect is partial or complete (Storbeck et al., 2019). In neonates, the difficulty in diagnosis during the early neonatal period was described affording to the relatively high levels of 3β-hydroxy-5-ene steroid observed in healthy neonates. 5PT is usually a minor component of the healthy neonatal urinary steroid metabolome, so any significant quantitative presence is pathognomonic of 3βHSD2D. In the first months of life

3β,15β,17α-trihydroxy-5-pregnen-20one and it 20- reduced metabolite are valuable analytes (Caulfield et al., 2002; Reeder & Joannou, 1996; Shackleton, 1986, 2008).

Metabolite ratios of DHEA over glucocorticoid metabolites and 17-hydroxypregnenolone (5PT) over glucocorticoid metabolites have been proposed to aid diagnosis. Later, 5PT/PTONE was proposed as an additional ratio useful for the discrimination of 3βHSD2D from 21OHD (Caulfield et al., 2002; Shackleton, 2008)

Treatment involves replacement of mineralocorticoid and glucocorticoid deficiency but with more conservative dosing than used for other causes of CAH where treatment aims to suppress negative feedback within the HPA axis. Conservative dosing would reduce the risk of adverse long-term outcomes associated with high doses of glucocorticoid replacement. Sex hormone replacement may additionally be required at the time of puberty (Al Alawi et al., 2019).

1.6.3. Differences in Sex Development Due to Abnormal Steroidogenesis

1.6.3.1. 5α-Reductase Type 2 Deficiency

The 5α-reductase type 2 (SRD5A2) enzyme catalyses the conversion of testosterone to the much more potent androgen, dihydrotestosterone (DHT). Deficiency of functional SRD5A2 leads to 46,XY DSD. 5α-reductase type 2 deficiency (5αRD) was first described in 1961 by Nowakowski and Lenz and was initially termed “pseudovaginal perineoscrotal hypospadias” (Nowakowski & Lenz, 1961). The condition was better characterised through the 70s after studying numerous affected

individuals, including large kindred from the Dominican Republic (Imperato-McGinley, Gautier, Peterson, et al., 1986; Imperato-McGinley et al., 1974). 5αRD is inherited in an autosomal recessive manner. Genetic males appear to have externally undervirilised genitalia. There is no persistence of internal Müllerian structures and internal urogenital tract is consistent with male phenotype. Testes are present but for the majority located in the inguinal canal. There is wide phenotypic variation observed regards genital ambiguity at birth, with many having a completely female external genital phenotype. Milder degrees of undervirilisation have also been described, such as hypospadias and micropenis. Affected genetic males later present with rapid and profound virilisation at the time of puberty. There is no observable phenotype in genetic females (Maimoun et al., 2011; Mendonca et al., 2016; Odame et al., 1992). There are currently 3 described isozymes of 5α-reductase (type 1, type 2 and type 3) but only type 1 and 2 are responsible for in vivo conversion of testosterone to DHT (Jaeken et al., 2020; Okeigwe & Kuohung, 2014). They both catalyse the same reaction, but they are expressed in different tissues, genes are located on different chromosomes and their biological roles are distinct. Human 5α–reductase-2 gene (*SRD5A2*) is located on the short arm of chromosome 2. It is expressed in genital tissues from foetal life onwards. Human 5α reductase-1 gene (*SRD5A1*) is located on the short arm of chromosome 5. It is expressed in non-genital skin and liver but only following birth (Okeigwe & Kuohung, 2014). Testosterone produced from the testes is responsible for in utero development of the wolffian ducts into epididymis, vas, seminal vesicles, and ejaculatory ducts. Dihydrotestosterone is critical for the development of the prostate and external masculinisation such as midline fusion and development of the penis and scrotum. In the absence of SRD5A2, in addition to normally absent SRD5A1, expression in utero, DHT is not produced, therefore external structures are

not masculinised. During puberty, testosterone production by the testes increases considerably. Although there remains no functional 5α−reductase type 2 enzyme, the type 1 isoform can make the conversion to DHT, and this results in the profound and rapid virilisation observed in these individuals at the time of puberty (Maimoun et al., 2011; Okeigwe & Kuohung, 2014).

Over 120 mutations have been identified, with nearly 70% being missense/nonsense mutations (http://www.hgmd.cf.ac.uk). Affected individuals have been found to have homozygous or less commonly compound heterozygous loss of function mutations. Many mutations are associated with specific geographical and ethnic groups, with most affected patients having been described in populations from the Dominican Republic, Asia, Brazil, Turkey, and Papua New Guinea. There is phenotypic variation observed even in individuals with the same causative genetic mutation (Batista & Mendonca, 2020).

Diagnosis can be made through molecular genetic analysis at any age. Patients display age-appropriate levels of testosterone but the testosterone to DHT ratio is high, particularly apparent after human chorionic gonadotropin (hCG) stimulation testing. It is challenging to establish the diagnosis biochemically in newborns and very young infants owing to physiologically normal very low circulating levels of androgens in this age group. Increased concentrations of testosterone produced during minipuberty might provide an opportunity for diagnosis but owing to the presence of SRD5A1 post birth, there can be peripheral conversion of testosterone to DHT which can lead to misleadingly normal testing (Maimoun et al., 2011; Peterson et al., 1977).

A significant body of work was carried out by Imperato-McGinley and Shackleton in the 1970s to characterise the observed urinary steroid metabolome changes observed in 5αRD by examining a large cohort of affected individuals from the Dominican Republic (Imperato-McGinley, Gautier, Peterson, et al., 1986; Imperato-McGinley et al., 1974; Peterson et al., 1985b). Androsterone is a urinary metabolite of androstenedione, testosterone, and dihydrotestosterone, whereas etiocholanolone is a metabolite from the upstream androgens, androstenedione and testosterone. As in this condition, the block is from conversion of testosterone to dihydrotestosterone, it would be expected that etiocholanolone would be higher than androsterone and a ratio of these metabolites has been proposed to assist with diagnosis. Metabolites of cortisol and corticosterone are usually 5α− reduced by SRD5A2 in normal physiology and ratios of these metabolites over their 5α− reduced counterparts (THB/5aTHB and THF/5aTHF) effectively display the enzymatic defect. The diagnosis is challenging in neonates as healthy babies excrete very little THF or THB for the first months of life. It is generally accepted that urinary steroid profiling for the diagnosis of 5αRD is best employed after 3 months of age (Perry et al., 2011; Shackleton, 2008; Storbeck et al., 2019).

Treatment depends on sex assignment and age of presentation. Antenatal androgen exposure is an important factor to consider with gender-related neural programming and identity. Antenatal androgen exposure correlates with frequency of male gender role and identity amongst causes of 46,XY DSD, more so than genital phenotype (Loch Batista et al., 2019; Wisniewski et al., 2019). Certainly, in populations such as those with complete androgen insensitivity syndrome, where 46,XY individuals have no functional exposure to androgens during prenatal life, female sex is assigned at birth

and almost universally maintained throughout life. Whereas, in 5αRD, the chance of patient driven sex reassignment is significant, with 60% changing social sex and choosing to identify as male. With this evidence as support, the 2016 consensus statement from the Global DSD Update Consortium recommended male sex of rearing for patients diagnosed in infancy with 5αRD (Lee et al., 2016). Patients diagnosed at puberty when virilisation occurs who have been living in the female gender role require a multidisciplinary approach including psychological support to enable them to choose their gender identity. Patients can be treated with GnRH agonists to pause and block further pubertal development, giving them the time to make an informed choice. Patients who are assigned male sex of rearing in infancy can be treated with topical DHT to induce external masculinisation of structures, with the knowledge that further virilisation will occur at the time of puberty. Surgical intervention may be indicated, including hypospadias repair and orchidopexy (Mendonca et al., 2016).

1.6.3.2. Cytochrome b5 deficiency

Cytochrome b5 (CYB5A) deficiency is a very rare monogenic disorder, with only 5 cases reported (Giordano et al., 1994; Hegesh et al., 1986; Idkowiak et al., 2012; Kok et al., 2010). In addition to conveying 17,20 lyase activity of CYP17A1, cytochrome b5 also plays key roles in haemoglobin synthesis and hepatic phase 1 drug metabolism. The first reported case of CYB5A deficiency was a patient with severe methaemoglobinaemia presenting with fatal cyanosis at birth (Hegesh et al., 1986). The baby, who carried a severe splice-site mutation, was also found to have ambiguous genitalia. Only recently, 4 children from 2 different pedigrees with CYB5A deficiency have been reported: all affected children did not have any clinical cyanosis,

but insufficient masculinization in 46,XY individuals (46,XY DSD) and lack of pubertal development in the only affected 46,XX individual (Idkowiak et al., 2012; Kok et al., 2010). Methaemoglobin levels were found to be mildly raised without clinical evidence of cyanosis. Affected children were homozygous for a novel nonsense mutation (p.W27X), likely to abolish protein activity (Kok et al., 2010), and a homozygous missense mutation, p.H44L, which retains minimal protein function on 17,20 lyase activity in vitro (Idkowiak et al., 2012).

Biochemical investigations, particularly urinary steroid profiling, revealed an exclusive impairment of the 17,20 lyase activity of CYP17A1 with intact 17α-hydroxylase activity. This reflects normal glucocorticoid and mineralocorticoid production with isolated androgen deficiency, in essence true isolated 17,20 lyase deficiency (Idkowiak et al., 2012).

1.6.3.3. 17β–Hydroxysteroid Dehydrogenase Type 3 Deficiency

17β–hydroxysteroid dehydrogenase type 3 (HSD17B3), also known as 17β-keto reductase, is expressed exclusively in the Leydig cells of the testes and converts androstenedione to testosterone. HSD17B3 deficiency results in 46,XY DSD. 46,XX individuals are unaffected (Geissler et al., 1994; Mendonca et al., 2017). It is an autosomal recessive disease and caused by homozygous or compound heterozygous mutations in the *HSD17B3* gene, located on the long arm of chromosome 9. Just under 60 mutations have been described in the gene, the vast majority being missense/nonsense mutations (http://www.hgmd.cf.ac.uk). HSD17B3 deficiency was first described in 1971 by Saez et al in a male described as having

pseudohermaphroditism and gynaecomastia (Saez et al., 1971). Registry data would suggest it is one of the most common disorders of androgen synthesis, constituting 4% of 46,XY DSD cases (Wisniewski et al., 2019). However, incidence is ethnicity and geographically dependent, with most cases described in families of Mediterranean, Middle Eastern or African origin (Mendonca et al., 2017).

Affected genetic males usually appear phenotypically female at birth and as such are usually assigned female sex. Internal anatomy is consistent with male development and testes are usually descended to the inguinal canals or labioscrotal folds. Occasionally, children come to medical attention in infancy if there is some degree of virilization, gonads are palpable or if gonads are found at inguinal hernia repair. More commonly, patients present at the time of puberty with primary amenorrhoea and signs of virilisation (Mendonca et al., 2017). While HSD17B3 is the predominant enzyme responsible for production of testicular testosterone, there are other HSD17B isoforms expressed in a variety of tissues, capable of catalysing this conversion, that remain unaffected. This explains the described classical phenotype as testosterone is required for the development of male internal urogenital structures and for virilisation at puberty. The presence of different isoforms may also contribute to the observed phenotypic variability (Storbeck et al., 2019).

Diagnosis can be made biochemically: androstenedione levels are raised with concurrently low testosterone meaning a low ratio of testosterone over androstenedione supports the diagnosis. This finding is exaggerated following hCG stimulation testing (Mendonca et al., 2017). Diagnosis of HSD17B3 deficiency from urine steroid metabolome analysis is challenging. Hypothetically logical urinary steroid

metabolome findings appear to be masked by the function of other isoforms of HSD17B (Storbeck et al., 2019). It is generally accepted that the reference standard method for biochemical diagnosis is serum measurement of hormone levels before and following hCG stimulation testing. Molecular genetic techniques can confirm the diagnosis.

The 2016 consensus statement by the Global DSD Update Consortium (Lee et al., 2016) regarding the management of patients with DSD recommends male sex assignment for patients diagnosed to have HSD17B3 deficiency as over half of affected patients later switch to male gender identity. However, multidisciplinary assessment and support is required to enable each individual and their families (where appropriate) to make an informed decision regarding gender assignment. There are reports of both satisfactory and non-satisfactory psychological and psychosexual outcomes for individuals assigned to either gender, so an individualised approach including a suitably experienced multidisciplinary team is essential. Treatment is tailored to gender assignment with appropriate sex hormonal replacement and surgical intervention, for example orchidopexy, as required.

1.7. Steroid measurement techniques

For many decades serum steroids have been measured with immunoassays. Currently, mass spectrometry has emerged as the reference standard for steroid measurement as it is highly sensitive and specific, does not suffer the same issues with cross reactivity and allows for simultaneous profiling of multiple steroids.

1.7.1. Immunoassay

Immunoassay is a technique where a labelled antibody is used to bind to an analyte of interest in order to indirectly quantify its abundance. The technique of immunoassay and later radioimmunoassay were developed in the late 50s and 60s. Over the following decades they became the predominant technique for steroid hormone measurement due to low cost, technical simplicity, and rapid results. However, modern studies have demonstrated the specificity of immunoassay techniques are not as reliable as once thought due to cross reactivity of the antibody employed with other unintended molecules. They are also not suitable for urinary metabolite measurement, again due to cross reactivity with other constituents in the matrix. Now that our understanding of the various enzyme defects that can result in CAH or DSD, and our understanding of the process of steroidogenesis is increasing, clinicians are more inclined to request several hormones to be measured rather than single hormones in isolation. As a different immunoassay would be required for each hormone required to be measured, this impacts on the amount of sample required which is an important consideration for neonates in particular (Wudy et al., 2018).

1.7.2. Chromatography with Mass Spectrometry

Mass spectrometry (MS) based methods are now considered superior for steroid hormone measurement and are slowly becoming more widely available in routine clinical practice. Equipment costs and expertise required is much higher, however simultaneous measurement of whole steroid hormone panels are possible, from relatively low volume patient samples in a variety of matrices, and sensitivity and

specificity are much improved (Wudy et al., 2018). Current recommendations acknowledge the superiority of MS based analysis but support appropriate use of both MS and Immunoassay at present owing to the wider availability of immunoassay at the current time. Measurement of steroid hormones in serum/plasma and measurement of steroid hormone metabolites in urine are currently supported for routine clinical use in the diagnosis of inborn disorders of steroidogenesis, whereas salivary measurements are not yet recommended for routine clinical practice until further work is done to establish analysis using this matrix (Kulle et al., 2017).

1.7.3. Principles of Chromatography

The basic components are common to each chromatographic technique. There is an inlet, a mobile phase, and a stationary phase. The mobile phase carries the sample through the machine and can be gas (as in gas chromatography, GC), liquid (as in liquid chromatography, LC) or supercritical fluid (as in supercritical fluid chromatography, SFC). The different types of mobile phase have their own intrinsic advantages and disadvantages. Gas chromatography-mass spectrometry (GC-MS), was the first hyphenated separation/MS methodology, dating back to the 1960's. In GC, helium is frequently used as the mobile phase. The technique has excellent chromatographic resolution resulting in excellent specificity, but it is time consuming and labour intensive- primarily due to the fact that chemical derivatisation has to be used, and therefore analyses are expensive (Krone et al., 2010). For many purposes, Liquid Chromatography tandem Mass Spectrometry (LC-MS/MS), developed in 80's and 90's, is much more suited to high throughput measurements as sample work-up is considerably quicker (no derivatisation), cheaper, and can be automated at all

stages. In comparison to GC, resolution is poorer and, particularly for the type of urinary metabolites we encounter, sensitivity is not as high. The method is limited to a lower number of analytes measurable per run, epimeric and isomeric steroid separation more challenging, and LC-MS/MS doesn't have the qualitative "scan" ability (described below) for non-targeted studies. LC-MS/MS is ideal for measuring smaller panels of steroids, particularly, for example, analysis of serum hormonal 3-Oxo-Δ^4 steroids; steroids which can be challenging with GC-MS (Krone et al., 2010; Shackleton et al., 2018).

Supercritical fluid chromatography-mass spectrometry (SFC-MS) is a relatively new concept and uses very high-pressure carbon dioxide, resulting in properties of both gas and liquid states. It is thought that this will confer the benefits of both gas and liquid chromatography, but SFC-MS remains mainly limited to the research setting at present (Storbeck et al., 2018).

The stationary phase is inside the column that the mobile phase carries the sample through. The components of a sample interact with the stationary phase differently depending on their chemistry and this affects the speed at which they travel across it, helping to further separate the different sample components out by time. Column chemistry is complex and beyond the scope of this thesis, briefly, the stationary phase is selected according to the specific analytes of interest.

1.7.4. Principles of Mass Spectrometry

The separated components of a sample exit the column and enter the mass spectrometer, which consists of an ionisation source, a mass analyser, and a detector. The time taken for a given component to traverse the column and reach the detector is known as the retention time. Classically, the mass spectrometer ionizes and fragments the component substances, accelerates the ions, and deflects them through an electromagnetic field before finally hitting the detector plate. Modern mass spectrometers use quadrupole mass analysers to select and separate ions by their mass-to-charge (m/z) ratio. These consist of 4 parallel cylindrical metal rods, charged by direct current and radiofrequency voltage, with opposing rods connected to carry the same charge, to create an oscillating electromagnetic field through which ions are passed. As the voltage varies, ions with differing m/z ratios will preferentially pass through the quadrupole. Once an ion achieves a stable trajectory, it will reach the ion detector. Finally, the abundance is measured as an electrical current from the ion detector. The more abundant the ion, the greater the current produced (Wudy et al., 2018).

1.7.5. Application of GC-MS in Urinary Steroid Metabolite Analysis for Adrenal Disease

GC-MS as a technique for the separation and detection of urinary steroids/metabolites was introduced into clinical practice in the late 60s and 70s (Shackleton, 1986). GC-MS is highly suited to detect the distinct steroid metabolome fingerprints for inborn disorders of steroidogenesis. Multiple steroid metabolites from all 3 steroid hormone

pathways can be quantified simultaneously in a single run, making it an effective way to diagnose steroidogenesis related disorders non-invasively (Wudy et al., 2018). The various steroidogenic deficiencies have mostly been characterised by their urinary steroid metabolome well before the serum/plasma steroid hormone descriptions were made. A summary of the key urinary steroid metabolome changes observed for the causes of CAH, as determined by GC-MS, is shown in **Figure 1.7**. 24-hours of urine collection is preferable to allow for absolute quantification of steroid metabolites and comparison to laboratory specific age- and sex-determined reference ranges (Honour et al., 2018; Malunowicz et al., 1997).

1.7.6. GC- Scan vs SIM, Identification vs Quantification

In “scan” mode, the mass spectrometer records throughout the range of detectable ions (m/z, mass/charge, where charge is 1, in GC-MS) from low to high. The output is a full mass spectrum, usually very definitive for the compound. The “spectrum” of an individual peak or component is structurally informative and such spectra have remained very consistent during decades of analysis throughout the world. This technique provides a detailed, unfiltered picture of all the components in each sample, allowing characterisation of all analytes. However, scanning is mostly used as a qualitative exercise because sensitivity of detection is low, and quantitation of minor components, and compounds within overlapping peaks is challenging (Shackleton & Marcos, 2005).

Routine quantification is much more feasible when the instrument can focus on specific ions within the mass spectrum of an individual compound. This results in an

increase in sensitivity, as well as much higher degree of specificity. In “selected ion monitoring” (SIM) mode data is collected only for an inputted selection of masses specific to the analytes of interest, rather than looking at the whole spectrum, allowing more precise measurement of abundance, and therefore exact quantification. The compounds separated appear as SIM peaks eluting sequentially as oven temperature increases. Broadly speaking, compounds are eluted in order of their mass and structural complexity; androgen metabolites, for example, appear early, and multi-hydroxylated corticosteroid metabolites late. The data system integrates the SIM peaks and relates the area of each peak to that of the designated SIM ion for an internal standard (IS) such as stigmasterol (SS) (m/z 394). This gives the abundance of the analyte of interest relative to the abundance of the known quantity of stigmasterol in the sample. An External Standard (ES) containing all the analytes of interest at known concentrations and the same internal standard is run prior to each batch of samples. The data system calculates the concentration of all analytes within the urine sample by comparing the comparative abundance of analyte and SS SIM peaks in the sample to that of the ES (Shackleton, 2008; Shackleton & Marcos, 2005).

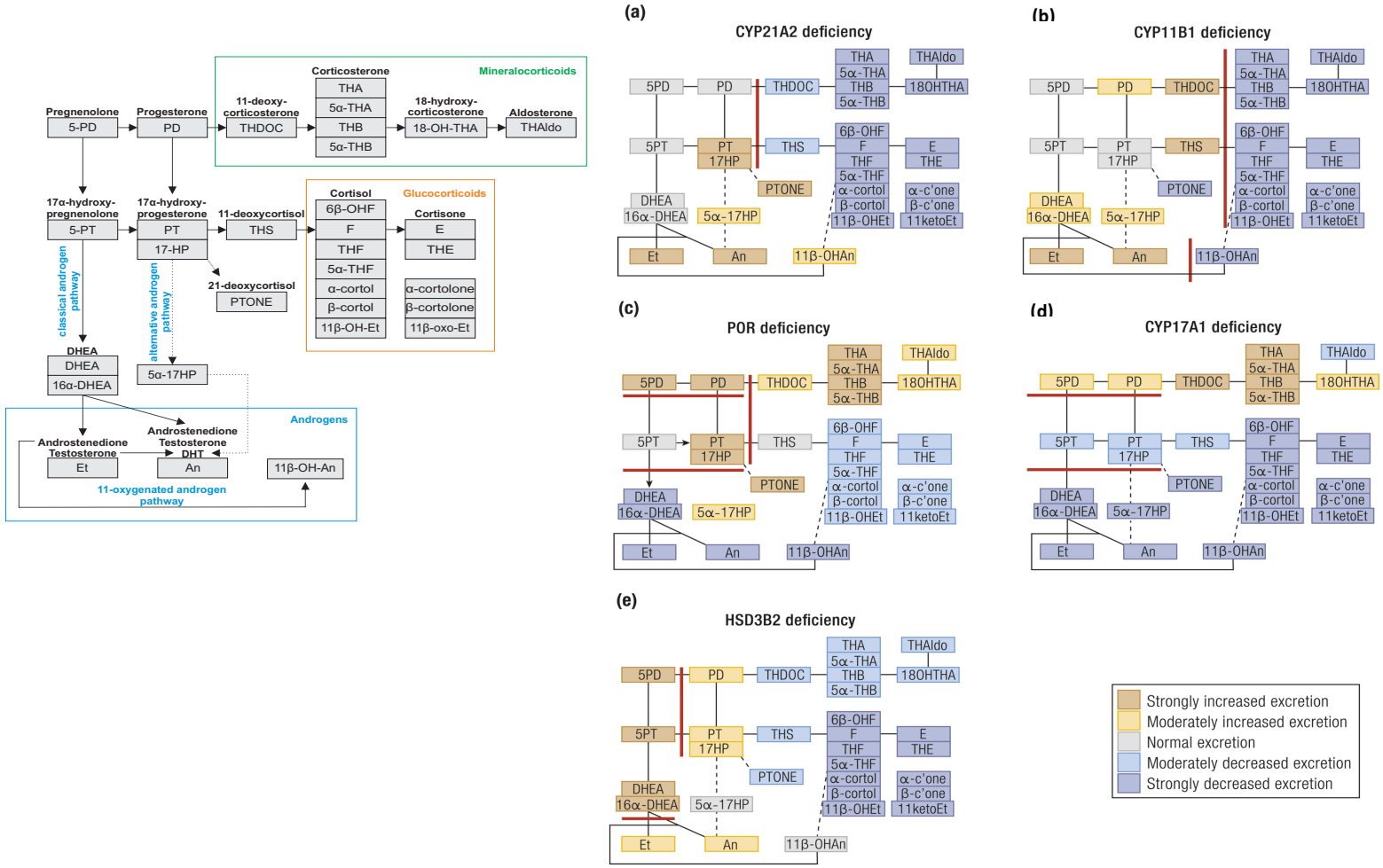


Figure 1.7: Characteristic urinary steroid metabolome findings in CAH

Heat map depicts relative excretion of urinary steroid metabolites for 5 enzyme deficiencies leading to CAH. Corresponding serum steroids are shown in the left panel (Storbeck et al., 2019).

1.8. Machine Learning and its Application to Steroid Metabolome Analysis

1.8.1. Principles of Machine Learning Analysis

From Alan Turing's remarkable description of "Learning Machines" in 1950 (TURING, 1950), and Arthur Samuel's popularisation of the term "Machine Learning" through his publication of a program developed to learn to play checkers in 1959 (Samuel, 1959), the field of machine learning has shown exponential growth.

According to the Oxford English Dictionary, machine learning can be defined as

"The use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyse and draw inferences from patterns in data".

In essence, machine learning algorithms are designed through various statistical, probabilistic and optimisation methods, to learn through "experience" to find natural patterns in data. This "knowledge" can be generalised to a wider cohort to make decisions or predictions. This is a powerful approach which is particularly useful for complex tasks, if the problem involves a large amount of data or if the inherent internal structure of the data is not clear (Schwartz, 1970; Uddin et al., 2019). Recent years have seen a boom in machine learning applications, including spam email recognition, home voice assistants, targeted advertising, smart motorways, self-driving cars, and image recognition.

There are multiple categorisations of machine learning methods proposed. Learning methods are commonly categorised as unsupervised or supervised. Unsupervised learning methods are used to discover inherent patterns or structure within an unlabelled set of data. Common tasks unsupervised learning is employed for include clustering and feature reduction (Sidey-Gibbons & Sidey-Gibbons, 2019). Examples of unsupervised learning techniques include probabilistic principal component analysis (PPCA), Gaussian Mixture Model, and K-means/medoids. Supervised learning methods take a set of labelled data to train and optimise a model according to the required task, which is then evaluated by application to an unseen set of labelled test data. Common tasks supervised learning is employed for include classification and prediction (Deo, 2015).

1.8.2. Machine Learning in Medicine

In 1970, an article in the New England Journal of Medicine predicted computing science would augment or replace the intellectual functions of the physician, achievable only through interaction and collaboration across medicine, information sciences, management sciences and policy makers (Schwartz, 1970). Machine learning techniques are being increasingly used within medical and research domains to meet the demands for analysis of increasingly large amounts of complex data (Hey et al., 2020). As data has become digitised, such as imaging modalities, and databases have grown to include massive amounts of data of increasing complexity such as in the “-omics” fields (genomics, transcriptomics, metabolomics, etc), the need for increasingly sophisticated analysis and prognostication techniques has increased. A study has found that nearly 1/3 of preventable deaths in the UK are attributable to

misdiagnosis, prompting demand for decision support systems/tools in the healthcare setting (Hogan et al., 2012; Watson et al., 2019). This has driven the need for collaborative work between domain knowledge holders, statisticians, mathematicians, and computer scientists to develop strategies to comprehensively analyse “big data” and determine meaningful outputs. Solution strategies have been developed for specific tasks within a wide range of medical and scientific fields leading to an exponential rise in applicable and transferable algorithm methods (Rajkomar et al., 2019). A PubMed search on the 23rd Feb 2021 using the search term “Machine Learning” returned 55,434 results published since the year 2000, including 424 clinical trials and randomised controlled trials published since 2010. Machine learning tools have been created for a variety of conditions for risk stratification, phenotyping, treatment outcome prediction, for example. There has also been recent work in using machine learning algorithms for the interpretation of imaging data such as for retinal photography for diabetic retinopathy (Gulshan et al., 2016) or tissue biopsies for malignancy (Beck et al., 2011; Esteva et al., 2017).

Explicitly data driven models aim to provide an unbiased perspective and potentially allow for knowledge gain. However, some variables may be inherently biased (Vayena et al., 2018; Watson et al., 2019). I include two examples of this. Firstly, the algorithm designed to predict GCSE and A-level grades for students when written exams could not occur in 2020 due to the COVID-19 pandemic which (inadvertently) unfairly discriminated against students from low socioeconomic backgrounds (https://rpubs.com/JeniT/ofqual-algorithm, https://www.bbc.co.uk/news/explainers-53807730). Secondly, an algorithm designed to predict mortality risk in hospitalised patients with pneumonia classified people with asthma as low risk because they were

so frequently cared for in the intensive care unit from the outset having been sent there directly from the emergency department, which despite frequently needing escalating levels of care consistently improved their prognosis (Caruana et al., 2015). These examples highlight the importance for interdisciplinary working, those with domain knowledge to work together with computer scientists, to maximise the accuracy, fairness and generalisability of any model created.

1.8.3. Biological Considerations for the Development of Machine Learning Methods

1.8.3.1. Interpretability and Accountability

Along with the increasing demand for powerful analysis tools, there is also now demand for model transparency and explainability (Barredo Arrieta et al., 2020; Cabitza et al., 2017; Holzinger et al., 2019; Vayena et al., 2018; Watson et al., 2019). The UK Royal Society in a 2017 report (royalsociety.org/machine-learning) highlighted algorithm interpretability, robustness, fairness, inference of causality, and human-machine interactions as priority funding areas in machine learning research (Blacklaws, 2018). Using explainable AI (XAI) models, in the healthcare setting in particular, promotes trust, improves understanding, ensures model accountability and protects against unjust outcomes (Vayena et al., 2018; Watson et al., 2019). Furthermore, considering interpretability as a driver to designing ML methods can aid acceptability and facilitate wider implementation (Holzinger et al., 2019; Ipsos, 2017). Interpretability within a model helps with bias detection and so ensures impartiality in decision making, it provides robustness to the method, and acts as insurance for the

method by being able to ensure model reasoning comes from meaningful variables inferring the output (Barredo Arrieta et al., 2020). The European Union's General Data Protection Regulation (GDPR) has provided a further driver for explainable models as any institution using algorithmic decision tools may be required to justify and explain decisions made to any person for whom they hold data (Vayena et al., 2018).

Carvahlo et al (Carvalho et al., 2019) (2019) describe different classifications of interpretability applicable for machine learning models. Pre-model interpretability techniques describe characteristics of the data, such as descriptive statistical techniques and data visualisation techniques. In-model interpretability describes the working logic of a model and is a characteristic of "intrinsically interpretable models". Post model interpretability is the same as post-hoc explainability and is a characteristic of models which provide a detailed output to explain how results are reached.

1.8.3.2. Heterogeneity

Heterogeneity within medical data is a known challenge. It can arise due to characteristic population or ethnicity differences (Tillin et al., 2014). Most human systems are highly dynamic as they transition through stages of life which has implications for ensuring adequate representation for this additional heterogeneity. Any variation in sample collection or processing, or differences in limits of quantification between centres or modalities can introduce heterogeneity. Gold standard diagnostic tests, responsible for sample labels a model is trained from often carry imprecision. Additional heterogeneity can be introduced through human factors such as physician documentation or interpretation (Cios & Moore, 2002). Metabolomics data is often very skewed with large outliers. Outliers may not represent

typical patterns observed and can mislead or negatively influence how an algorithm makes decisions- for example, nearest neighbour classifiers can be negatively influenced by heavily outlying samples (Kubat, 2015).

1.8.3.3. Missing Values

Medical databases are often fraught with missing values and can be categorised as “missing not at random” (MNAR), “missing at random” (MAR) or “missing completely at random” (MCAR) (Little, 2019). The aetiology of missing values is important to ascertain to avoid introducing bias into a model. Systematic missingness implies structure or a dependent variable underlying the cause. For example, if samples were collected or processed at different centres, or if a method evolved over time such that samples collected more recently may have different metabolites measured than the more historic samples. This systematic missingness is termed “Missing Not at Random” or MNAR. “Missing At Random” (MAR) means there is a relationship between the observed values and the value to be missing, the information about what is missing can be inferred from other variables. For example, in GC-MS, having large amounts of a particular metabolite in your sample can mean that certain other metabolites are more likely to be unquantifiable. Whereas “missing completely at random” (MCAR), there is absolutely no relationship between a missing value and any observed or unobserved other values. For example, presence of a contaminant which co-elutes with a metabolite of interest regarding the metabolite unquantifiable.

It has been suggested by García-Laencina et al (García-Laencina et al., 2010) that strategies to handle missing values fall in to four broad categories. Subjects with

missing values can be deleted and model training informed only by complete profiles. This is appropriate for MCAR, otherwise deleting cases with missing values is likely to introduce bias into the model. Data distribution can be modelled according to the maximum likelihood by using model-based procedures such as an expectation–maximization algorithm. This strategy is also best used when missing values are MCAR, otherwise caution must be exercised because there is the potential for introducing bias or error into the modelling procedure (Little, 1988). Missing data can be estimated or imputed, and then synthetically completed profiles can be included in model training along with the experimentally complete profiles. Finally, missing values can be incorporated into a machine learning classifier capable of training with incomplete datasets such as prototype-based machine learning methods or probabilistic random forests.

1.8.3.4. Cohort Size

Health research is frequently challenged with having small or imbalanced class sizes, that don't necessarily reflect disease importance or prevalence (Cios & Moore, 2002). A dataset needs to be sufficiently sized to be representative of the heterogeneity within the population it is derived from. Insufficiently sized classes can lead to poor training performance, or overfitting and poor model generalization performance (Ghosh et al., 2017; Ying, 2019). If the classes being examined do not contain similar numbers of subjects, then the classes are referred to as imbalanced. The number of healthy controls in a database often far outweigh the number of patients, which potentially introduces a bias in performance measures towards the majority class. In medical datasets it is often the case that the minority class/es are the more important patient

classes, and the majority class is the less relevant healthy class (Chawla et al., 2002; Cios & Moore, 2002; Ghosh et al., 2017). Where there are severe imbalances between classes, certain machine learning techniques will tend to focus training on the majority class. Minority classes are relatively ignored, smaller numbers make it more difficult to create decision boundaries for them, and thus overall performance for those classes will be poor. Therefore, it is important to address the imbalance so that the classes are treated at least equally in training (Chawla et al., 2002). Several techniques have been suggested to address this issue in the pre-processing step. Cost-function weighting can be implemented in training to introduce misclassification costs to improve sensitivity of smaller but important classes (Ghosh et al., 2017). It is possible to under sample the majority class, taking only a random or representative sample of equal size to the minority classes. This is a feasible option if the minority classes still have a substantial enough number to represent the natural variability within the class, but more often it is better to somehow increase the numbers in the minority class up to the majority class. The simplest approach would be to duplicate results, but this doesn't necessarily add any new information for training the model. Alternatively, new sample results can be synthesized based on the existing results. One method for this is called the Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002). SMOTE works by using samples in close proximity in the feature space and creates / adds additional samples at random positions in the space between them. This approach can also be used in combination with under sampling the majority class and can be used to create as many synthetic samples as required.

1.8.3.5. Examples of Translated Use in Steroid Metabolome Analysis

In 2011, Arlt W et al (Arlt et al., 2011) published the details of a urine steroid metabolomics approach, combining comprehensive urinary steroid metabolite quantification by mass spectrometry methods with a machine learning analysis method designed to differentiate malignant from benign adrenal tumours. The machine learning method created used a technique called Generalized Matrix Relevance Learning Vector Quantization (GMLVQ) to differentiate the two groups (Schneider et al., 2009). Over a 3-year period, 147 patients with adrenal tumours, recruited in 6 different specialist centres, collected 24-hour urine samples. Urinary steroid metabolites were quantified using GC-MS. Using all 32 quantified urine steroid metabolites, this method was able to discriminate adrenal cortical carcinoma (ACC = malignant) from adrenal cortical adenoma (ACA = Benign) with 90% sensitivity and specificity. Furthermore, a hierarchy of discriminatory ability for all 32 features was determined by analysis of the relevance matrix obtained by GMLVQ. Using only the 9 most discriminatory features, the method was able to discriminate ACC from ACA with 87.7% sensitivity and specificity. In a large follow-on prospective validation study (Bancos et al., 2020), 2017 patients with newly diagnosed adrenal tumours were included in the study, recruited from 14 specialist centres in 11 countries. 15 urinary steroid metabolites were quantified from 24-hour urine collections using LC-MS/MS. GMLVQ was again employed, the method trained on retrospective data from patients with either confirmed ACC or ACA. The algorithm was designed to predict one of three outcome scores determined by the distance to respective prototype – “low risk”, “medium risk” and “high risk”. It was proposed that urine steroid metabolomics should be used in combination with imaging defined clinical characteristics such as tumour

size and tumour tissue attenuation on an unenhanced CT scan in order to improve detection of ACC and reduce the number of false positive results observed with single test strategies. Positive predictive value if the triple test strategy yielded a result that was unanimously positive or “high risk” was 76.4%. A result that was unanimously negative or “low risk” on the triple test strategy correctly ruled out ACC in 95.7% of patients with a non-ACC tumour.

In 2019, Chortis et al (Chortis et al., 2020) proposed that a urine steroid metabolomics approach could also be used to detect recurrence of ACC post-surgical resection. They recruited patients with fully resected ACC within 2 years of their surgery and compared 32 patients who provided pre and post recurrence 24-hour urine samples with urine from 39 patients who showed no evidence of recurrence. 19 urinary steroid metabolites were quantified using GC-MS. They used a random forest classification machine learning algorithm to differentiate between patients who had recurrent disease and those who did not based on their urine steroid metabolome. The classifier was able to predict recurrence of disease with 81% sensitivity and specificity and 6 metabolites were identified as being most relevant for this discrimination.

Wilkes et al (Wilkes et al., 2018) investigated whether machine learning algorithms could be used to approximate human expert interpretation of urinary steroid profiling, for a range of suspected diagnoses, but without knowledge of the final confirmed clinical diagnosis. They investigated three types of algorithms: random forest, weighted-subspace random forest, and extreme gradient boosted tree algorithms. They investigated 6 groups of profiles where experts had concluded a suspected diagnosis of 1. “Query adrenal suppression”, 2. “Query Adrenal Tumour”, 3. “Query

Cushing's", 4. No significant abnormality, 5. "Query 21OHD CAH" or 6. "query 5α-reductase inhibition". Like many medical databases, the healthy or "no abnormality suspected" class was by far the majority class, so they also investigated the effect of under sampling (or subsampling) this class on model performance. They found similar performance for the 3 algorithms investigated or a binary problem to approximate expert analysis to differentiate normal from abnormal profiles. For the multiclass problem, they found the best performing algorithm for approximating expert opinion on the individual suspected diagnosis was Random Forest with subsampling the majority class. They were also able to use a Boruta algorithm to aid interpretability of their method and ascertain the more important features (or metabolites) for the separation of the classes, though they did not ascertain any disease (suspected) specific information. The authors specifically emphasize

"This study was an evaluation of the system to predict the interpretation of profiles by experienced practitioners and not a model of diagnostic accuracy itself (i.e., not using definitive clinical outcome data)".

The authors aimed to create an algorithm that can think like an expert, however, this does not allow for the possibility for the expert being incorrect, and the authors do state the limitation of their study being the possible subjectivity of their "ground truth".

In 2020, Kamrath C et al (Kamrath et al., 2020), looked at a cohort of 109 patients with 21OHD aiming to identify a metabolomic fingerprint, or "metabotype", to explain patient variability in response to glucocorticoid treatment. They quantified 31 urinary steroid metabolites which they corrected for body surface area then log-transformed and z-score transformed. 11 of the 31 metabolites were excluded from preliminary analysis to reduce bias and one further metabolite was later excluded due to being

non-discriminatory between groups. The authors used a web-based software tool which used a k-means clustering algorithm (unsupervised learning approach) to identify distinct metabotypes. Metabotype clusters were further investigated using statistical techniques including ANOVA, sparse partial least square discriminant analysis and receiver operating characteristic curve analysis to identify discriminative biomarker metabolites and to build and assess a biomarker classification model. They identified 4 distinct metabotypes, defined by differing urinary steroid metabolome patterns, which represented patients who were adequately treated, over treated, undertreated and a final group that did not adequately respond to treatment. Identification of this final group is significant as these patients may either represent non-compliance or those that require a different dosing regimen for adequacy of treatment.

1.9. Project Aims and Objectives

The aim of this thesis was to investigate the use of urinary steroid metabolite ratios as surrogate markers of enzyme activity, to simplify and standardise the interpretation of urinary steroid metabolome data. We aim to create a non-invasive, easily interpretable test method for differential diagnosis of inborn disorders of steroidogenesis, which is suitable for routine clinical use. The method is required to be fit for use with single spot urine samples as well as 24-hour urine sample collections. We design this test to be used in a targeted symptomatic population, not as a whole population screening test.

In order to achieve this aim, we have 3 broad objectives which we set out in the following chapters.

Our first objective was to characterise the intrinsic properties of urinary steroid metabolome data through the use of 14 previously published ratios of precursor to product steroid metabolites as surrogate markers of steroidogenic enzyme activity. Particular attention is paid to determining factors which would impact on the creation of meaningful normative reference ranges. In chapter 3, we define how steroidogenic enzymatic activity, as determined by 14 key urinary steroid metabolite ratios, varies across the life course. We further characterise sex-related variability observed in the urinary steroid metabolome. Finally, we assess how the healthy population compares to the relevant patient populations with respect to these biochemical ratios.

Our second objective was to create a robust and straightforward algorithm for urine steroid metabolome interpretation, built using the 14 biochemical ratios of precursor to product steroid metabolites. In chapter 4 we investigate the specificity and discriminative ability of these ratios to each enzymatic defect. We propose a simple algorithm for interpretation of comprehensive steroid metabolome data as quantified by GC-MS. We investigate two methods for determining reference intervals- one using a percentile as the upper limit of normal, and a second using a multiple of the median approach. We assessed the performance of this diagnostic tool using a database of comprehensive urine steroid metabolome data, quantified by GC-MS, from healthy controls and patients with definitively determined diagnoses.

Our final objective was to take a Urine Steroid Metabolomics approach to create a data-driven, but biologically relevant, diagnostic tool for automated differential diagnosis of inborn disorders of adrenal steroidogenesis. In chapter 5 we describe the

construction and refinement of a fully interpretable machine learning-based model, trained using comprehensive urinary steroid metabolite measurements as quantified by GC-MS. We employ a cross validation approach to approximate expected model performance using our cohort of healthy controls and patients with definitively determined diagnoses. We present methods to graphically display results to end users. Finally, we investigate the urine steroid metabolome patterns the machine learning model has determined as characteristic for each enzymatic defect. Through this strategy, we hope to create a rapid, high-performance method that is informative, acceptable to patients, clinicians and biochemists working in the field, and potentially could increase domain knowledge.

2. Materials and Methods

2.1. Study Design

This was a retrospective study using a database of anonymised urine steroid metabolome data collected through multiple previous studies.

2.1.1. Participants

Urine samples were collected from 829 healthy controls and 178 patients definitively confirmed through genetic testing to harbour an inborn disorder of steroidogenesis. All healthy controls were screened by means of a questionnaire to ensure that they had no medical condition or regular medication that would interfere with steroidogenesis.

There were 6 patient groups. 21-hydroxylase deficiency (21OHD, N= 27), 11β-hydroxylase deficiency (11βOHD, N= 12), P450 Oxidoreductase deficiency (PORD, N= 36), 17α-hydroxylase deficiency (17OHD, N= 30), 3β-hydroxysteroid dehydrogenase type 2 deficiency (3βHSD2D, N= 22) and 5α-reductase type 2 deficiency (5αRD, N= 51). The healthy control group had adequate representation across sex and age ranges with near equal proportions of male and female contributors (Female, N = 440; Male N = 389), and similar group sizes for 3 defined age groups for each sex. A detailed cohort description is included in **Table 2.1.**

Table 2.1: Cohort Description

Total numbers of healthy controls and patient group participants. Sub-group description including breakdown of each group by age and sex are detailed here.

Healthy Control Cohort Total = 829 Age range = 3 days – 81 years	**Male**			**Female**		
	389			440		
	Under 6m	**6m-16yr**	**Over 16yr**	**Under 6m**	**6m-16yr**	**Over 16yr**
	126	125	138	123	95	222
Patient Cohort Total = 178	**Male**			**Female**		
21-hydroxylase deficiency N = 12 Age range = 5 days – 32 years	7			20		
	Under 6m	**6m-16yr**	**Over 16yr**	**Under 6m**	**6m-16yr**	**Over 16yr**
	4	2	1	5	10	5
11β-hydroxylase deficiency N = 12 Age range = 1.8 years – 28years	4			8		
	Under 6m	**6m-16yr**	**Over 16yr**	**Under 6m**	**6m-16yr**	**Over 16yr**
	0	3	1	0	6	2
P450 oxidoreductase deficiency N = 36 Age range = 20 days – 46 years	16			20		
	Under 6m	**6m-16yr**	**Over 16yr**	**Under 6m**	**6m-16yr**	**Over 16yr**
	4	12	0	4	10	6
17α-hydroxylase deficiency N = 30 Age range = 5 years – 57 years	16			14		
	Under 6m	**6m-16yr**	**Over 16yr**	**Under 6m**	**6m-16yr**	**Over 16yr**
	0	2	14	0	1	13
3β-hydroxysteroid dehydrogenase deficiency N = 22 Age range = 5 months – 18years	13			9		
	Under 6m	**6m-16yr**	**Over 16yr**	**Under 6m**	**6m-16yr**	**Over 16yr**
	1	12	0	1	7	1
5α-reductase type 2 deficiency N = 51 Age range = 3 months – 32years	47			4		
	Under 6m	**6m-16yr**	**Over 16yr**	**Under 6m**	**6m-16yr**	**Over 16yr**
	1	23	23	0	0	4

2.1.2. Sample Collection, Processing and Storage

Urine samples collected were a mixture of random single “spot” and 24-hour collection urine samples. Single spot urine samples were predominantly collected from children under 6 months of age. All samples from patients were collected pre-treatment or off steroid hormone replacement treatment for a minimum of 48 hours. Anonymised,

aliquoted samples, free from sediment, were stored at -20^{o}C. Samples were extracted and quantified at the Institute of Metabolism and Systems Research, University of Birmingham, UK or at the Children's Hospital Oakland Research Institute, USA. Urinary steroid metabolite concentrations for 32 individual metabolites were quantified using GC-MS (as described below). The two centre GC-MS platforms have undergone regular cross validation. Subjects with less than 80% of the 32 steroid metabolites we aimed to quantify were not included in the study database.

2.1.3. Study Ethics

Urine samples were collected as part of the Euro-DSD study (the European Community's Seventh Framework Program (Collaborative Research Project EuroDSD GA-2008-201444), donated to the Birmingham Human Bioresource Centre and their approved research studies (HTA Licence 12358, North West - Haydock Research Ethics Committee; Ref 20/NW/0001) or were collected with informed consent as part of a local research protocol, with approval of local ethical review boards (approved by the committee on ethics in human research from Escola Paulista de Medicina (n.1703/98), the Ethics Committee of Marmara University, Faculty of Medicine, Istanbul, Turkey (10840098-604.01.01-E0.4621)).

Healthy controls were recruited via a local, University of Birmingham ethically approved study, approved by the Science, Technology, Engineering and Mathematics Ethical Review Committee of the University of Birmingham (ERN_17–0494).

2.2. Urinary Steroid Metabolite Analysis by Gas Chromatography-Mass Spectrometry

2.2.1. Steroid Extraction Procedure

The methods used for steroid extraction from urine are long established and largely based on those described by Shackleton and co-workers (Shackleton, 1986; Shackleton & Marcos, 2005). In the studies reported here, 1ml of urine (or 2ml if the sample volume was greater than 2.5L) is passed across a C18 solid phase extraction column (disposable cartridge, Sep-Pak®) after the column is primed with 4ml methanol followed by 4ml distilled water. C18 columns are used to adsorb analytes from aqueous solutions and have been shown to aid efficiency of recovery of free and conjugated steroids from urine (Shackleton & Whitney, 1980). The column is washed with a further 4ml of distilled water and then the sample is eluted off the column with 4ml of methanol. The methanol is evaporated under nitrogen in a heat block at 55°C. Next, the steroids are hydrolysed in 0.1M acetate buffer with sulphatase/glucuronidase enzyme (commercial powder from the snail *Helix Pomatia*) for 3 hours at 55°C. The hydrolysis mixture is passed through the re-primed C18 solid phase extraction columns with the same procedure as before, and the hydrolysed steroids are recovered with 4ml of Methanol, this time into silanised glass tubes. 100μL of an internal standard (IS) mixture containing stigmasterol (SS) at a concentration of 25μg/mL, cholesterol butyrate (CB) at a concentration of 25μg/mL and 3β,5β-tetrahydroaldosterone (3β5βTHAldo) at a concentration of 2.5μg/mL is added to each sample before the methanol is evaporated under nitrogen at 55°C. Steroids are fragile molecules, and “chemical derivatisation” must be carried out to protect them from the

high temperatures used for gas chromatography. Carbonyl (ketone) groups are protected as “methyloxime” derivatives and hydroxyl groups protected by “silylation”. Methoximation is carried out with 100µL of methoxyamine hydrochloride in pyridine (2% w/v) for 1 hour at 55^{o}C, then silylation achieved with 75µL TMSI (N-trimethylsilylimidazole) at 120 ^{o}C overnight. Derivatisation reagents are involatile which is unsuitable for gas chromatography, therefore the derivatised steroids are purified and recovered by liquid-liquid extraction using 2ml of cyclohexane and 2ml distilled water. After vortexing and centrifugation, the cyclohexane phase is recovered and evaporated under nitrogen at 55^{o}C. Finally, the samples are re-dissolved in 400µL cyclohexane and transferred to autosampler vials ready for GC-MS analysis.

2.2.2. Sample Analysis by Gas Chromatography-Mass Spectrometry

The auto-sampler syringe takes 2µL of sample and injects it into the heated inlet (260^{o}C) ahead of the column where the sample is vapourised. The column temperature is initially held at 100^{o}C for each injection, so the sample mixture (less solvent vapour) recondenses at the start of column. After holding for 1 minute the column temperature is ramped up to 230^{o}C at a rate of 30^{o}C per minute. Once 230^{o}C is achieved there is a slower second temperature increase up to 300^{o}C at a rate of 3^{o}C per minute which is subsequently held for 8 minutes. The instrument then cools down to 100^{o}C again prior the next injection. Typically, the column type used for steroid analysis is termed “non-polar” based on the type of stationary phase coating the interior. The “column” used for the analysis presented in this study is an open silica capillary (Nominal length: 30.0 m, nominal diameter: 250.00 µm. Model

Number: Agilent 222-0132LTM DB-1ms (G3900-63017)). The mobile phase used is helium and it flows at a rate of 1mL/min.

After chromatographic separation, the samples go into the mass spectrometer detector. The compounds exiting the column are ionised, generally with electron beam (electron impact, EI) in an instrument component called the "ion-source." In our quantitative SIM method, the quadrupole separates and quantifies the selected "molecular" and "fragment" ions designated for individual compounds by the operator. The steroid metabolites which have been selected to be included in the SIM mode method have been chosen to comprehensively represent the breadth of gonadal and adrenal steroidogenesis.

We quantified 32 selected urinary steroid metabolites, providing comprehensive representation from all steroid and steroid precursor groups in the process of mineralocorticoid, glucocorticoid, and androgen biosynthesis (steroidogenesis). **Table 2.2** lists these steroid metabolites routinely quantified for the urine samples included in this study, the specific ions used for each compound and their abbreviations and derivations. We describe the use of 14 steroid metabolite precursor to product ratios as surrogate markers of steroidogenic enzyme activity (Storbeck et al., 2019). Included ratios have been previously published as part of a visual interpretation tool (Krone et al., 2010). These ratios are listed in **Table 2.3.**

Table 2.2: List of steroid metabolites quantified *Abbreviation, common name, chemical name and derivative.*

Abbreviation	Common name	Chemical name	Metabolite of
An	Androsterone	5α-androstane- 3α-ol-17-one	Androstenedione, testosterone, 5α-dihydrotestosterone
Et	Etiocholanolone	5β-androstane- 3α-ol-17-one	Androstenedione, testosterone
11β-OH-An	11β-hydroxy-androsterone	5α-androstane- 3α,11β -diol- 17-one	11β-hydroxyandrostenedione
DHEA	Dehydroepiandrosterone	5-androstene- 3β-ol-17-one	DHEA, DHEAS
16α-OH-DHEA	16α-hydroxy- dehydroepiandrosterone	5-androstene- 3β,16α-diol- 17-one	DHEA, DHEAS
5-PT	Pregnenetriol	5-pregnene- 3β,17-20α- triol	17-hydroxypregnenolone
5-PD	Pregnenediol	5-pregnene- 3β,20α-diol and 5,17,(20)- pregnadien- 3β-ol	Pregnenolone
THDOC	Tetrahydro- 11-deoxycorticosterone	5β-pregnane- 3α,21-diol- 20-one	11-deoxycorticosterone
THA	Tetrahydro- 11-dehydrocorticosterone	5β-pregnane- 3α,21-diol- 11,20-dione	Corticosterone, 11-dehydrocorticosterone
5α-THA	5α-tetrahydro- 11-dehydrocorticosterone	5α-pregnane- 3α,21-diol- 11,20-dione	Corticosterone, 11-dehydrocorticosterone
THB	Tetrahydrocorticosterone	5β-pregnane- 3α,11β,21- triol-20-one	Corticosterone
5α-THB	5α-tetrahydro-corticosterone	5α-pregnane- 3α,11β,21- triol-20-one	Corticosterone
18-OH-THA	18-hydroxytetrahydro-11- dehydrocorticosterone	5β-pregnane- 3α,18,21- triol-11,20- dione	18-hydroxycorticosterone
THAldo	3α,5β- tetrahydroaldosterone	5β-pregnane- 3α,11β,21- triol-20-one- 18-al	Aldosterone
PD	Pregnanediol	5β-pregnane- 3α,20α-diol	Progesterone
3α-5α- 17HP	5α-17- hydroxypregnanolone	5α-pregnane- 3α,17α-diol- 20-one	17-hydroxyprogesterone
17HP	17-hydroxypregnanolone	5β-pregnane- 3α,17α-diol- 20-one	17-hydroxyprogesterone
PT	Pregnanetriol	5β-pregnane- 3α,17α,20α- triol	17-hydroxyprogesterone
PT'ONE	Pregnanetriolone	5β-pregnane- 3α,17α,20α- triol-11-one	21- deoxycortisol
THS	Tetrahydro- 11-deoxycortisol	5β-pregnane- 3α,17α,21- triol-20-one	11- deoxycortisol
F	Cortisol	4-pregnene- 11β,17,21- triol-3,20- dione	Cortisol
6β-OH-F	6β-hydroxycortisol	4-pregnene- 6β,11β,17,21- tetrol-3,20-dione	Cortisol
THF	Tetrahydrocortisol	5β-pregnane- 3α,11β,17,21- tetrol-20-one	Cortisol
5α-THF	5α-tetrahydrocortisol	5α-pregnane- 3α,11β,17,21- tetrol-20-one	Cortisol
α-cortol	α-cortol	5β-pregnane- 3α,11β,17,20 α,21-pentol	Cortisol
β-cortol	β-cortol	5β-pregnane- 3α,11β,17,20 β,21-pentol	Cortisol
11β-OH-Et	11β-hydroxyetiocholanolone	5β-androstane- 3α,11β-diol- 17-one	Cortisol
E	Cortisone	4-pregnene- 17α,21-diol- 3,11,20-trione	Cortisone
THE	Tetrahydrocortisone	5β-pregnene- 3α,17,21- triol-11,20-dione	Cortisone
α-cortolone	α-cortolone	5β-pregnane- 3α,17,20α,21- tetrol-11-one	Cortisone
β-cortolone	β-cortolone	5β-pregnane- 3α,17,20β,21- tetrol-11-one	Cortisone
11-oxo-Etio	11-oxo-etiocholanolone	5β-androstane- 3α-ol-11,17- dione	Cortisol, Cortisone

Table 2.3: Urinary steroid metabolite ratios proposed to support the diagnosis of inborn disorders of steroidogenesis

14 urinary steroid metabolite ratios of precursor steroids to product steroids, which enzyme, or co-factor they are surrogate markers of activity for, and the inborn steroidogenesis disorder an elevated ratio indicates.

No.	Urinary Steroid Metabolite Ratio	Enzyme / Co-Factor Activity	Steroidogenesis Disorder Disruption is Indicative of
1	100*PTONE/(THE+THF+5αTHF)	CYP21A2 + POR	21-hydroxylase deficiency or P450 oxidoreductase deficiency
2	(17HP+PT)/(THE+THF+5αTHF)	CYP21A2 + POR	21-hydroxylase deficiency or P450 oxidoreductase deficiency
3	PD/(THE+THF+5αTHF)	CYP17A1 + POR	P450 oxidoreductase deficiency
4	5PD/(THE+THF+5αTHF)	CYP17A1 + POR	P450 oxidoreductase deficiency
5	(17HP+PT)/(An+Et)	CYP17A1 + POR	P450 oxidoreductase deficiency
6	(THA+5αTHA+THB+5αTHB)/ (THE+THF+5αTHF)	CYP17A1 + POR	17α-hydroxylase deficiency
7	(THA+5αTHA+THB+5αTHB)/ (An+Et)	CYP17A1 + POR	17α-hydroxylase deficiency
8	100*THS/(THE+THF+5αTHF)	CYP11B1	11β-hydroxylase deficiency
9	DHEA/(THE+THF+5αTHF)	HSD3B2	3β-hydroxysteroid dehydrogenase deficiency
10	5PT/(THE+THF+5αTHF)	HSD3B2	3β-hydroxysteroid dehydrogenase deficiency
11	5PT/PTONE	HSD3B2	3β-hydroxysteroid dehydrogenase deficiency
12	Et/An	SRD5A2	5α-reductase type 2 deficiency
13	THB/5αTHB	SRD5A2	5α-reductase type 2 deficiency
14	THF/5αTHF	SRD5A2	5α-reductase type 2 deficiency

2.2.3. Quality Control Procedures

Quality control (QC) samples are extracted and run alongside experimental samples. One is the first sample in the batch extracted and analysed, another the last sample in the batch. The QC "urine" is a pooled volume obtained from collection of urine from an equal number of male and female adults. Running identical samples at the beginning and end of each batch demonstrates if there is any analyst-, derivatisation- or instrument-dependent variability in the analysis. Not only does this measure variability within a run but comparing QC results obtained weeks apart gives an observation of instrument functional degradation over time and can indicate when maintenance needs to occur. Variability is calculated as relative standard deviation (RSD), or also known as coefficient of variation (CV), which is defined as the ratio of the standard deviation to the mean. An RSD of less than 20% is an acceptable degree of variation.

An internal standard (IS) is added to every sample to be analysed, including ES, QC and experimental samples. The internal standard consists of three components. Stigmasterol (SS) is added for the quantification of steroid metabolites to enable adjustment for any for any issues during the steroid extraction procedure such as potential sample losses. Cholesterol butyrate (CB) is incorporated for evaluating sample derivatisation and instrument performance. Finally, 3β,5β-tetrahydroaldosterone (3β5βTHALDO) is added for the specific measurement of 3α,5β–tetrahydroaldosterone, as derivatisation of this compound is not comparable to SS. The IS mix used for the sample analysis in this study contains SS at a

concentration of 25μg/mL, CB at a concentration of 25μg/mL and 3β5βTHALDO at a concentration of 2.5μg/mL. 100μL of the IS mixture is added to each sample.

While uncommon, the most vulnerable item in the analysis is the stability of the silylated steroid derivatives. Silyl ethers can hydrolyse, particularly on storage and samples should be run as soon as possible after preparation. We quantify derivative in all samples, including QCs and experimental samples, by observing the abundance of the silylated SS (m/z 394) compared to the ion (m/z 368) of CB, an internal standard without a silyl group. In addition, we semi-quantify non-derivatised SS with SIM ion at 412. This is generally negligible <1%) but we accept up to 5% hydrolysis.

2.3. Machine Learning Analysis

2.3.1. Unsupervised Learning: T-distributed Stochastic Neighbour Embedding

Unsupervised learning is an exploratory analysis. An algorithm seeks to find intrinsic patterns, groups or structure within unlabelled data. We used t-distributed stochastic neighbour embedding, t-SNE (Van der Maaten & Hinton, 2008) for unsupervised exploratory work and to create visualisations. T-distributed stochastic neighbour embedding, or t-SNE, is a non-linear embedding method useful for representing high dimensional data in a lower dimensional visualisation space as it is able to preserve object relationships observed in the high dimensional space. The t-SNE algorithm calculates a similarity measure between object pairs in the high dimensional space and in the low dimensional space. The algorithm uses a cost function, which is minimised through an optimisation procedure so that the difference between the two

similarity measures is as low as possible. Pairs of high dimensional objects are each assigned a probability depending on how similar (i.e., how close) the objects are from each other (the more similar the objects, the higher the assigned probability). It then aims to construct probability distributions in the lower dimensional space that are as close as possible to the corresponding probability distributions assigned to pairs of objects in the high dimensional space. Visually, samples / data points will be seen to position according to their similarity as they would in the higher dimensional space. The t-SNE methodology follows the philosophy of multidimensional scaling. However, using t-distributions instead of distances allows us to preferentially preserve the local distance structures with the highest precision, while less importance is given to minor differences in larger distances (Van der Maaten & Hinton, 2008).

2.3.2. Supervised Learning: Learning Vector Quantization

Popular dissimilarity-based classification algorithms, such as *k*-Nearest Neighbour (*k*NN, (Altman, 1992)), compare the features of an unknown sample to the features of known and labelled samples to make classification decisions based on similarity. This can be a complex task if there are a lot of samples, can be computationally quite expensive, and nearest neighbour classifiers can be highly affected by outliers. Learning Vector Quantization (LVQ) (Biehl et al., 2016; Kohonen, 1990) builds on the concept of a nearest neighbour classifier but simplifies the classification task as it is a prototype-based classification algorithm. For each defined group a typical representative example is derived, known as a prototype. A prototype is a manufactured sample vector constructed from what the algorithm has learnt from training to be the typical pattern of features for a given class. A supervised training

procedure aims to minimise a cost function so that constructed class prototypes best represent their given class (i.e. is in a position in the high dimensional space where the majority of class samples in training are closest and therefore correctly classified) (Kubat, 2015). Instead of comparing an unknown sample to numerous other samples, it only compares to these representative class points or class prototypes, which is a much simpler task, is computationally much less expensive and is less affected by outliers (Biehl et al., 2016; Kaden et al., 2014).

Classification algorithms compare similarity (or dissimilarity) of an unknown sample with known samples (real or constructed prototypes) for classification decisions. This can be implicit, such as in neural networks, or explicit, such as distance measures or dot products. LVQs use an explicit dissimilarity measure to compare sample features with prototype features and classify samples in the input feature space (Biehl et al., 2016; Kaden et al., 2014).

In chapter 5 we propose an extension to Learning Vector Quantization, generalized relevance learning vector quantization and generalised matrix learning vector quantisation methods, termed Angle Learning Vector Quantization (ALVQ) (Bunte et al., 2016). It uses a parameterised angle-based dissimilarity measure, adapting a global metric tensor, which is automatically refined through the training procedure. This enables weighting of individual dimensions according to discriminative power and rotation of the co-ordinate system towards the more discriminant directions in order to optimise the classification performance. Using this information, we can derive the classical “fingerprint” for each classification task – what is the typically observed pattern of features and which features are most relevant to the classification task. This

informs the construction of the prototypes. We further inspect prototypical classification terms for dimensionality reduction, by selecting and excluding the least relevant features then retraining using only the reduced feature set.

2.3.3. Strategies Employed to Handle Specific Biological and Intrinsic Factors

Our dataset contains some challenging intrinsic properties which will need to be addressed in order to create an unbiased machine learning method for automated diagnosis. These are small and imbalanced classes, heterogenous measurements, and missing data.

2.3.3.1. Small and imbalanced Classes

In our dataset we have highly imbalanced classes, with the healthy majority class having 829 subjects and our minority class (11βOHD) having only 12 subjects. Furthermore, class sizes for the patient classes do not represent relative population prevalence of each enzymatic defect, given 21OHD is by far the most prevalent. To maximise training performance for the minority classes we present two approaches. First, we oversample the minority classes to the approximate size of the majority class using geodesic SMOTE (Ghosh et al., 2017). Geodesic SMOTE was proposed as a variant of SMOTE (Chawla et al., 2002) for generating synthetic samples on a hypersphere. This is particularly relevant for our problem because ALVQ classifies on the unit hypersphere.

The second approach we present uses a weighted cost function to artificially put emphasis on the smaller classes by assigning a higher penalty for misclassification. Within the minority classes, we implement a cost function matrix which was refined by incorporating domain knowledge to allow for medical priorities. The highest misclassification cost was assigned to patients misdiagnosed as healthy as this clinical outcome is potentially devastating to the patient. Subsequent misclassification costs were assigned based on detriment to patient- patients with CAH being misclassified as 5αRD being penalised greater than patients with CAH being assigned as CAH but due to an incorrect enzyme defect. Finally, misclassification costs were given according to prevalence of disease, with correct classification of 21OHD given priority over the rarer disorders. Integration of expert driven priorities for the classification task enabled construction of an algorithm which is highly relevant for our classification task.

2.3.3.2. Heterogeneity

Because the urine samples that were collected from our participants are a combination of 24-hour urine collections and single spot sample collections, absolute quantification of individual metabolites are not comparable for the different sample collection methods. Within each individual class, we would expect to see some heterogeneity because of age- and sex-specific differences. The whole steroid output of a 1-year old's adrenal gland is very different to the whole steroid output of a 20-year-old, so the order of magnitude observed for the metabolites will be expected to vary considerably. Because inborn disorders of steroidogenesis result in an "enzymatic block" in the steroidogenesis pathway, we would expect to see an increase in precursor metabolites and a reduction in downstream product metabolites regardless of method

of collection, age-specific or sex-specific variability. Proportional relationships between metabolite concentrations will be preserved despite these intrinsic variabilities, and thus more revealing than comparing absolute metabolite quantifications. Therefore, we constructed features which use ratios of metabolites in order to create a robust classifier for use with any collected urine sample, regardless of collection method, participant age or participant sex. By using an angle-based dissimilarity measure we further prioritise the relationship between features as important rather than the absolute magnitude.

2.3.3.3. Missing Data

Figure 2.1 gives an overview of the missing metabolite values in our database. Subjects were only included if they had greater than 80% of the 32 measured urinary steroid metabolites quantified. From the figure we can see that there is more than one type of missingness affecting this database. The most frequently observed aetiology of missing values in our database is systematic or MNAR missingness. This arises primarily from the inclusion of a greater number of metabolites in the GC-MS quantification method over time as knowledge increased and the availability of synthetic steroid standards improved. Observed systematic missingness further arises from samples having been quantified in two different centres. MAR is likely to be observed within the dataset. Babies excrete high amounts of neonatal specific steroids as addressed in the introductory chapter which may interfere with the quantification of other metabolites. Within the patient groups, classically precursors and their metabolites are observed to be in very high concentrations which could also interfere with the quantification of other metabolites. It is unlikely that we will have any

significant MCAR observed in our database. This can arise from contaminants in a sample co-eluting with steroid metabolites of interest, for example.

ALVQ uses an angle-based dissimilarity measure, which is more robust with respect to missing values as the total angles are less effected by removing single dimensions when compared to a Euclidean based distance measure. This means that the model intrinsically handled missing values observed in our dataset, without the need for specific handling or pre-processing such as imputation. The effect of different types of missingness on the robustness of our proposed LVQ method has been systematically studied and is presented in (Ghosh, 2021).

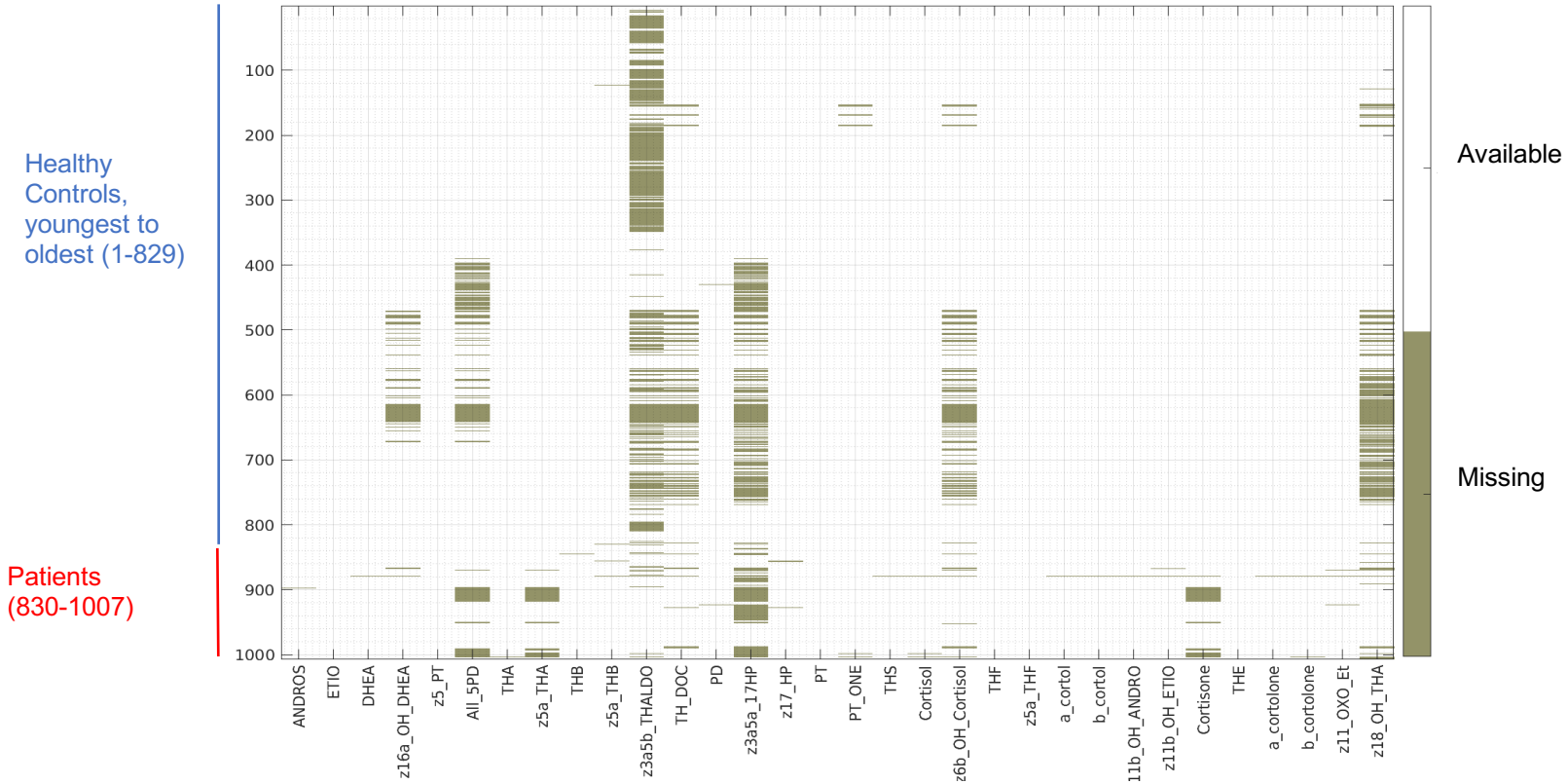


Figure 2.1: Graphical representation of urinary steroid metabolites that have not been quantified

Each row represents an individual participant numbered 1 to 1007 on the y-axis. Healthy controls are labelled numbers 1 to 829 in ascending order of age. Patient values are represented from numbers 830 to 1007. Columns represent each of the 32 metabolites routinely measured in this study. Gold bands display metabolites which have not been quantified for a given individual and are therefore missing values within the dataset

2.3.3.4. Conveying Generalizability Without a Specific Generalisation Cohort

With an appropriately sized cohort of data, the model is trained on an isolated subset of the total, termed the “training set”. Once trained, the model can be applied to a second subset of the total data, called a “validation set”. Performance when using the validation set can help to make an assessment for overfitting, can highlight the need for model refinement and can assist with model selection where different variants or experimental settings have been explored. Finally, a third subset of data is examined using the trained, and optimally refined model (as determined from validation set performance). This third set of data is termed the “generalisation set” and would ideally be a larger, prospectively collected cohort to convey the most accurate estimate of future performance.

In the context of small class sizes, as in our case, compromises to model performance analysis process must be made as there is not enough data to separate into training, validation and generalisation sets and no possibility of adequately powered prospective sample collection. Therefore, we employ a cross validation strategy to make realistic approximations of performance estimates (Berrar, 2019; Ferdinandy et al., 2020). The cohort is split into a larger subset used for model training and a smaller subset for testing model performance, with multiple iterations of this sub-grouping. For classification tasks, different splits are made such that each group or training set has roughly the same data distribution. This will result in multiple slightly different training sets of data to train the model and corresponding unseen groups in order to demonstrate performance. Results are averaged over the different folds of cross

validation as the average is statistically more reliable than looking at individual test fold performance. By using this approach, validation set performance metrics are more likely to be representative of generalisation performance.

The absence of a prospective validation cohort and the unfeasibility to collect one with such rare disorders hinders the ability to ensure robustness of any created model. To show evidence of the stability of the model to perform under unfavourable conditions, a much larger synthetic dataset, representative of the characteristics of our real-life dataset, was also created. By using a synthetic dataset, we systematically varied the degree of missingness and noise within the dataset to evaluate the effect on performance and compared to other classification techniques. This work is beyond the scope of this thesis but is presented in (Ghosh, 2021).

2.4. Statistical Analysis

Throughout this study we use non-parametric statistical tests. Descriptive statistics are shown in detail in **Chapter 3.** To assess for statistically significant group differences, Mann-Whitney *U* test was used for comparison between two groups and Kruskal-Wallis test was used for multiple group comparisons. To determine confidence intervals, the Wilson Method was used. Differences were considered statistically significant at $p<0.05$.

Diagnostic accuracy measures compared in this study are sensitivity to diagnose any patient with any disease, sensitivity to diagnose any patient with the correct disease, class-wise sensitivity to diagnose a patient of a given class with the correct disease,

and specificity. Performance metrics are shown as absolute proportions and percentages with 95% confidence intervals (CI).

Statistical analyses have been conducted and figures created using GraphPad Prism (version 9.1.0), StataCorp Stata (version 16), Microsoft Excel (version 16.48) or MathWorks MATLAB (R2020a). Machine learning based analysis and figures from this approach were performed/ created using MathWorks MATLAB (R2020a) or R Statistical software (V4.0.1; R Core Team 2020) using the package ggplot2 (WIckham, 2016).

For the machine learning work, we use cross validation. Performance metrics are stated as an average and standard deviation over 5 folds of cross validation. For the biochemical ratio approach, we use a whole cohort analysis which aims to provide the most accurate thresholds by using the most amount of data. For some aspects of the analysis for the biochemical approach, cross validation has been used for comparability with the machine learning approach. The exact same distribution and allocation of subjects in the 5 folds of training and testing as are used for both approaches.

The Standards for Reporting Diagnostic Accuracy (Bossuyt et al.) 2015 statement (Cohen et al., 2016), includes 30 essential items that should be included for diagnostic accuracy studies. We follow this framework in the reported work.

3. Biochemical Steroid Ratios as Surrogate Markers of Physiological Steroidogenic Enzyme Activity Across the Life Course

3.1. Introduction

The observed changes in adrenal cortex function, steroid hormone production, metabolism, and excretion over the course of life have been long studied and well described (Auchus & Rainey, 2004; Becker & Hesse, 2020; Bocian-Sobkowska, 2000; Davio et al., 2020; Dhayat et al., 2015; Eisenhofer et al., 2017; Garagorri et al., 2008; Moffat et al., 2020; Reiter et al., 1977; Suzuki et al., 2000). Previous authors have described the histological changes observed at various life stages (Suzuki et al., 2000), others have looked at macroscopic changes (Hata, Nagata, et al., 1988; Lam et al., 2001), changes in enzyme expression (Martel et al., 1994; Naccache et al., 2016; O'Shaughnessy et al., 2013), steroid hormone measurements in blood and metabolite excretion in urine (Dhayat et al., 2015; Eisenhofer et al., 2017; Garagorri et al., 2008; Homma et al., 2003; Sulcova et al., 1997; Zalas et al., 2018). The early neonatal period has already been identified as distinct regards steroidogenesis function and steroid hormone metabolism, with several “neonatal-specific” metabolites having been identified as important for diagnosis of congenital adrenal hyperplasia exclusively within this age group (Shackleton et al., 1979; Shackleton et al., 1980). Biochemical ratios of precursor to product steroid metabolites have been proposed as surrogate markers of steroidogenic enzyme activity and an aid to the differential diagnosis of congenital adrenal hyperplasia and differences in sex development (Caulfield et al., 2002; Honour & Brook, 1997; Shackleton, 1986; Shackleton et al., 1980). The further benefit of using a ratio approach is that it allows diagnosis from single spot urine samples as well as timed collections and 24-hour collections of urine, making this a much more feasible approach for young infants and much more convenient for patients and their parents.

Using mass spectrometry techniques for quantification, various authors have published methods to identify and quantify panels of serum steroids or urine steroid metabolites and have quoted appropriate reference ranges for these (Dhayat et al., 2015; Eisenhofer et al., 2017). However, methods of establishing appropriate reference intervals continue to be extensively discussed (Bertholf, 2006; Katayev et al., 2010; Ozarda et al., 2018; Siest et al., 2013). The most common reference intervals in use in clinical laboratories quote a central percentage, such as 95%, accepting that the 2.5% of the healthy population in the extreme tails will be defined as having an abnormal test (Jones & Barker, 2008; Wright & Royston, 1997). Using percentiles is particularly suitable for non-normally distributed data. This method approximates using a mean and standard deviation approach as 95.4% of data in a normal distribution is covered by using mean +/- 2 standard deviations either side of the mean. Rather than laboratory-specific reference intervals, use of common reference intervals where possible has been recommended (Jones & Barker, 2008; Ozarda et al., 2018; Yamamoto et al., 2013). The prerequisite for implementation is a similarity in population and accuracy between the laboratory which developed a reference range and the laboratory implementing it. Intra- and inter-laboratory variability in mass spectrometry quantification is well documented and highlights the importance for uniform protocols for method validation, performance reporting and monitoring (Fanelli et al., 2022). External quality assurance (EQA) programmes and standardised commercial calibration materials are recommended to harmonise cross laboratory testing, with inter-laboratory comparisons being required for accreditation (Greaves, 2017; Lynch, 2018; Vogeser & Stone, 2020). Though, inter-laboratory imprecision and variability is documented despite the use of EQA and standardised calibration materials (Fanelli et al., 2022).

Jones et al (Jones & Barker, 2008) recommended considerations for the development of reference ranges. Elements they discussed included consideration of the analyte, such as the clinical utility and biological variations. They discussed accuracy of the method employed for measurement and potential for interference. It is suggested that a minimum of 120 data measurements for each interval to be created should be obtained, based on the required number to calculate 90% confidence limits using non-parametric statistics. It is important to consider how to partition reference intervals, for example according to age, sex, ethnicity, or co-morbidity, which will impact on the number of data points required. Statistical methods to determine appropriate partitions based on data distribution have been published (Lahti et al., 2002; Lahti et al., 2004). Ultimately, reference intervals should be informative and relevant for informing clinical decision making.

3.2. Objective

The objective which we present in this chapter is threefold. Firstly, to characterise age- and sex-specific properties of urinary steroid metabolome data using 14 previously published ratios of precursor to product steroid metabolites as surrogate markers of steroidogenic enzyme activity. Secondly, to identify intrinsic factors which need to be considered for the determination of appropriate reference intervals. Finally, to assess how these biochemical ratios observed in the healthy population compare to the relevant patient populations for discrimination of the target disease.

3.3. Methods

3.3.1. Cohort Description

The study design and detailed cohort description are included in chapter 2, section 1. Please refer to **Table 2.1** for a detailed description of the cohort included in this study. For all samples, urinary steroid metabolites were quantified using gas chromatography-mass spectrometry as described in chapter 2, section 2.

3.3.2. Sub-Group Definitions

For investigation of age related changes observed in enzymatic function, our healthy control cohort was initially split into 10 informative age groupings: Less than 1 month (n=115), 1 month to 3 months (n=72), 3 months to 6 months (n=76), 6 months to 12 months (n=42), 1 year to 4 years (n=96), 4 years to 10 years (n=44), 10 years to 18 years (n=24), 18 years to 30 years (n=67), 30 years to 50 years (n=165) and >50 years of age (n=94). 34 samples were excluded from this analysis as they were collected from adults over 18 years of age, but details of exact age were not collected. We selected these groups to capture changes in the urinary steroid metabolome for key adrenal life stages as described in the introduction to this chapter. Male and Female data were combined for this investigation.

For more in-depth statistical analyses, including assessment of sex- as well as age-specific observed variability in enzyme function within healthy controls, the healthy cohort was split into 3 age groups. The 3 age groups were chosen to enable

comparative analysis of critical life stages (postnatal adaption during infancy, adrenarche, and adulthood) while allowing for the maintenance of substantial sex-specific class sizes for comparison. These groupings were as follows: under 6 months (n=249, 126 male and 123 female), 6 months to 16 years (n=220, 125 male and 95 female) and over 16 years (n=360, 138 male and 222 female).

Patient groups were used for comparative analysis alongside the healthy controls. Patient groups were not subdivided by age or sex for this analysis.

3.3.3. Steroid Analysis and Calculation of Defined steroid Ratios

Urinary metabolites were analysed by GC-MS, see Chapter 2 for details. 14 individual ratios of urinary metabolites of precursor steroids to product steroids were employed as representative of enzymatic function. They have been constructed so that ratio values have an inverse relationship with enzymatic activity. These ratios have been proposed for diagnosis of inborn disorders of steroidogenic function and are shown in **Table 2.3.**

3.4. Statistical Analysis

Analysis was carried out using GraphPad Prism (version 9.1.0) and Microsoft Excel (version 16.48). For statistical analysis, non-parametric tests have been used. Box plots demonstrate distribution of data, showing median and interquartile range. For the 10 age groups comparison, whiskers represent interdecile (10^{th} – 90^{th} percentiles) range. For the 3 age group comparisons, whiskers show 2.5^{th} and 97.5^{th} percentiles.

Scatter plots show median and interquartile range for healthy controls and the median alone for patient groups. For measures of spread, interquartile range and interdecile range are quoted. To assess for normal distribution formally, frequency histograms were constructed and Shapiro-Wilk test for normality was also used. For pair wise group comparisons Mann-Whitney *U* test was employed. For 3 group comparisons, Kruskal Wallis test was used. Statistical significance was determined by a p-value threshold of 0.05.

3.5. Results

3.5.1. Descriptive Statistics and Normality Testing

Using data from the 829 healthy controls, skewness, kurtosis, and interquartile range (IQR) were measured for all biochemical ratios examined in this work when the cohort was split by age (using 3 age groupings of less than 6 months, 6 months to 16 years and over 16 years) and split by age and sex. These results are presented in **Table 3.1**. Data distribution was highly skewed with a positive skew distribution. For all but one age- and sex-specific group for one ratio, the distribution was leptokurtic, wide with heavy tails implying significant outliers. Each of the ratios showed a variable distribution, and a high degree of variability was observed even between the 3 age groups for a given sex-specific ratio. For the female cohort, skewness ranged from 0.57 to 9.86, kurtosis ranged from 0.17 to 115.8 and interquartile range varied from 0.006 to 22.87. For the male cohort, skewness ranged from 0.68 to 9.66, kurtosis ranged from -0.12 to 100.8 and interquartile range varied from 0.003 to 34.32. With both male and female data were combined, skewness ranged from 0.72 to 11.5,

kurtosis ranged from 0.28 to 145 and interquartile range varied from 0.004 to 31.34. The significantly high values observed in these tables for skewness and kurtosis imply a non-normal distribution for most of our data.

To formally assess for data distribution, frequency histograms were constructed and Shapiro-Wilk test to assess for a normal and lognormal distribution was applied to healthy control data. For this analysis, the same age 3 age groups were used. Distribution was evaluated for male and female subject data individually, and with sex combined for each age group. With female and male data combined, no groups for any of the ratios displayed a normal distribution. Corresponding frequency histograms are shown in **Figure 3.1**. 13 of 42 groups showed a log normal distribution. For male and female data analysed individually, no groups for any of the ratios showed a normal distribution. 20 of 42 male groups and 15 of 42 female groups showed a lognormal distribution. Data for each ratio was log transformed and Shapiro-Wilk test for normality was reconducted to confirm these results. It was not possible to normalise all calculated ratios for all subgroups using log transformation. These results are shown in **Table 3.2.**

Table 3.1: Descriptive statistics for 14 biochemical ratios

Ratio	21OHD/PORD Ratio 1: 100*PT'ONE/(THE+THF+5aTHF)								
Sex	Female			Male			Female and Male		
Age	<6m	6m-16yr	>16yr	<6m	6m-16yr	>16yr	<6m	6m-16yr	>16yr
Number of values	123	95	222	118	125	138	241	217	363
IQ range	0.99	0.20	0.17	0.53	0.18	0.13	0.75	0.18	0.15
Skewness	8.37	3.88	6.05	4.96	5.47	2.99	12	5.5	7.3
Kurtosis	75.00	17.94	43.71	27.73	34.92	11.39	145	38	66

Ratio	21OHD/PORD Ratio 2: (17HP+PT)/(THE+THF+5aTHF)								
Sex	Female			Male			Female and Male		
Age	<6m	6m-16yr	>16yr	<6m	6m-16yr	>16yr	<6m	6m-16yr	>16yr
Number of values	123	95	222	126	125	138	249	217	363
IQ range	0.02	0.03	0.09	0.03	0.02	0.08	0.02	0.02	0.09
Skewness	8.00	3.23	2.39	4.29	3.68	2.05	11	3.6	2.4
Kurtosis	70.38	12.32	7.90	25.27	14.95	7.11	133	16	8.9

Ratio	PORD Ratio 3: 5PD/(THE+THF+5aTHF)								
Sex	Female			Male			Female and Male		
Age	<6m	6m-16yr	>16yr	<6m	6m-16yr	>16yr	<6m	6m-16yr	>16yr
Number of values	123	62	188	126	102	112	249	165	300
IQ range	0.27	0.04	0.05	0.22	0.04	0.05	0.26	0.04	0.06
Skewness	4.09	0.90	4.87	3.20	2.32	1.86	3.6	2	4.1
Kurtosis	23.35	0.78	33.76	12.35	5.98	3.86	17	5.4	28

Ratio	PORD Ratio 4: PD/(THE+THF+5aTHF)								
Sex	Female			Male			Female and Male		
Age	<6m	6m-16yr	>16yr	<6m	6m-16yr	>16yr	<6m	6m-16yr	>16yr
Number of values	123	94	222	126	125	138	249	216	363
IQ range	0.01	0.03	0.08	0.01	0.02	0.03	0.01	0.02	0.05
Skewness	8.85	5.85	5.31	9.66	6.13	2.71	11	6	6.6
Kurtosis	89.13	41.17	39.19	100.80	43.67	10.66	142	41	62

Ratio	PORD Ratio 5: (17HP+PT)/(An +Et)								
Sex	Female			Male			Female and Male		
Age	<6m	6m-16yr	>16yr	<6m	6m-16yr	>16yr	<6m	6m-16yr	>16yr
Number of values	123	95	222	126	125	138	249	217	363
IQ range	2.49	2.21	0.14	1.52	2.06	0.09	2.00	2.07	0.11
Skewness	7.25	1.07	2.36	1.67	2.52	0.76	10	2.1	2.7
Kurtosis	54.50	0.76	8.53	4.64	11.24	1.65	109	9	13

Ratio	17αOHD Ratio 6: (THA+5aTHA+THB+5aTHB)/(THE+THF+5aTHF)								
Sex	Female			Male			Female and Male		
Age	<6m	6m-16yr	>16yr	<6m	6m-16yr	>16yr	<6m	6m-16yr	>16yr
Number of values	123	95	222	126	125	138	249	217	363
IQ range	0.13	0.13	0.05	0.14	0.13	0.05	0.14	0.13	0.05
Skewness	3.28	0.84	2.13	0.81	0.68	1.09	3.2	0.72	2
Kurtosis	14.26	1.08	6.79	1.03	-0.12	0.89	17	0.28	6.5

Ratio	17αOHD Ratio 7: (THA+5aTHA+THB+5aTHB)/(An+Et)								
Sex	Female			Male			Female and Male		
Age	<6m	6m-16yr	>16yr	<6m	6m-16yr	>16yr	<6m	6m-16yr	>16yr
Number of values	123	95	222	126	125	138	249	217	363
IQ range	20.84	22.87	0.19	10.70	34.32	0.14	15.50	31.80	0.17
Skewness	2.37	1.51	2.51	2.23	2.45	1.00	2.6	2.5	2.8
Kurtosis	8.65	2.82	9.41	6.51	11.09	1.17	11	12	13

Table 3.1: Descriptive statistics for 14 biochemical ratios

Ratio	**11βOHD Ratio 8: 100*(THS)/(THE+THF+5aTHF)**								
Sex	Female			Male			Female and Male		
Age	<6m	6m-16yr	>16yr	<6m	6m-16yr	>16yr	<6m	6m-16yr	>16yr
Number of values	123	95	222	126	125	138	249	217	363
IQ range	0.74	1.01	0.85	0.68	0.92	0.54	0.75	1.00	0.68
Skewness	9.54	1.27	2.02	5.56	1.12	2.10	11	1.2	2.1
Kurtosis	99.38	1.67	6.37	39.92	4.20	8.51	141	2.7	7.5

Ratio	**3βHSD2D Ratio 9: DHEA/(THE+THF+5aTHF)**								
Sex	Female			Male			Female and Male		
Age	<6m	6m-16yr	>16yr	<6m	6m-16yr	>16yr	<6m	6m-16yr	>16yr
Number of values	123	95	222	126	125	138	249	217	363
IQ range	0.01	0.01	0.09	0.01	0.00	0.19	0.01	0.00	0.13
Skewness	4.76	6.96	9.86	3.82	3.81	2.48	5.3	7.7	10
Kurtosis	25.95	57.58	115.80	18.41	15.73	6.66	35	78	141

Ratio	**3βHSD2D Ratio 10: 5PT/(THE+THF+5aTHF)**								
Sex	Female			Male			Female and Male		
Age	<6m	6m-16yr	>16yr	<6m	6m-16yr	>16yr	<6m	6m-16yr	>16yr
Number of values	123	95	222	126	125	138	249	217	363
IQ range	0.02	0.01	0.03	0.03	0.01	0.04	0.02	0.01	0.04
Skewness	5.61	1.21	5.35	4.38	1.90	1.41	5.1	1.6	4.7
Kurtosis	39.53	1.84	40.56	23.02	4.11	2.25	33	3.2	38

Ratio	**3βHSD2D Ratio 11: 5PT/PT'ONE**								
Sex	Female			Male			Female and Male		
Age	<6m	6m-16yr	>16yr	<6m	6m-16yr	>16yr	<6m	6m-16yr	>16yr
Number of values	123	95	222	118	124	137	241	217	362
IQ range	6.53	5.90	21.34	6.86	6.07	27.10	7.10	5.80	23.00
Skewness	1.80	0.78	3.45	4.19	1.55	1.88	4.4	1.4	2.9
Kurtosis	3.61	0.17	21.21	21.58	2.42	4.47	28	2.2	15

Ratio	**5αRD Ratio 12: Et/An**								
Sex	Female			Male			Female and Male		
Age	<6m	6m-16yr	>16yr	<6m	6m-16yr	>16yr	<6m	6m-16yr	>16yr
Number of values	123	95	222	126	125	138	249	217	363
IQ range	0.62	0.55	0.68	0.33	0.67	0.50	0.79	0.65	0.68
Skewness	0.57	1.06	2.71	1.24	2.40	0.72	0.87	2.3	2.7
Kurtosis	0.88	1.39	14.97	0.48	10.64	0.51	0.65	11	17

Ratio	**5αRD Ratio 13: THB/5aTHB**								
Sex	Female			Male			Female and Male		
Age	<6m	6m-16yr	>16yr	<6m	6m-16yr	>16yr	<6m	6m-16yr	>16yr
Number of values	123	95	222	126	125	138	249	217	363
IQ range	0.78	0.15	0.28	1.00	0.11	0.27	0.78	0.12	0.27
Skewness	3.71	2.35	4.94	3.35	1.59	1.90	3.5	2.5	5.1
Kurtosis	19.22	8.12	38.88	14.35	3.30	5.76	17	10	46

Ratio	**5αRD Ratio 14: THF/5aTHF**								
Sex	Female			Male			Female and Male		
Age	<6m	6m-16yr	>16yr	<6m	6m-16yr	>16yr	<6m	6m-16yr	>16yr
Number of values	123	95	222	126	125	138	249	217	363
IQ range	0.75	0.60	0.96	0.81	0.37	0.71	0.81	0.44	0.87
Skewness	7.87	1.82	3.01	6.72	1.27	3.74	7.2	2	3.4
Kurtosis	73.49	3.22	11.45	55.92	1.26	21.98	62	5	15

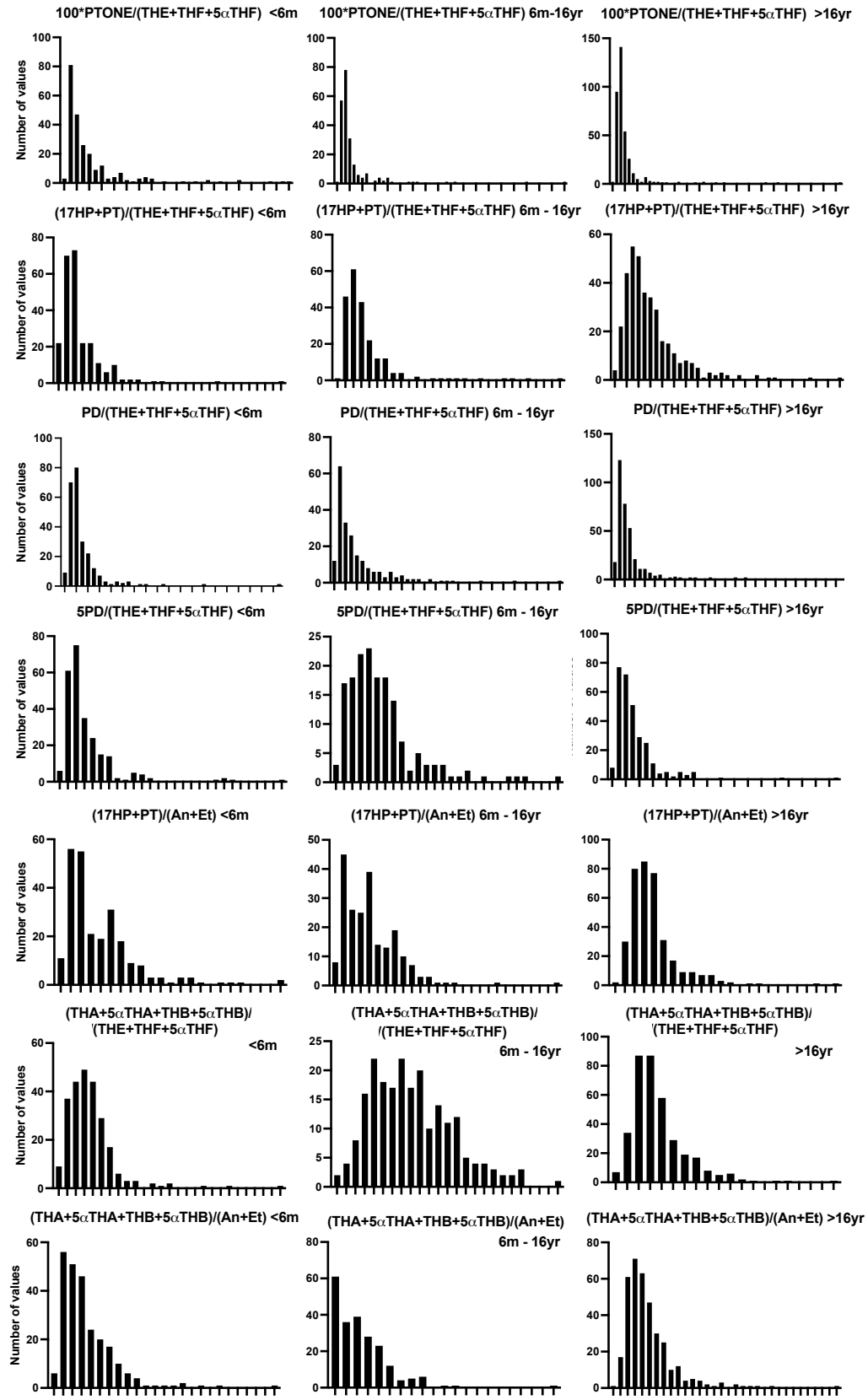


Figure 3.1: Frequency histograms for 14 biochemical ratios

Healthy control data is shown split by age. Bars correspond to number of subjects.

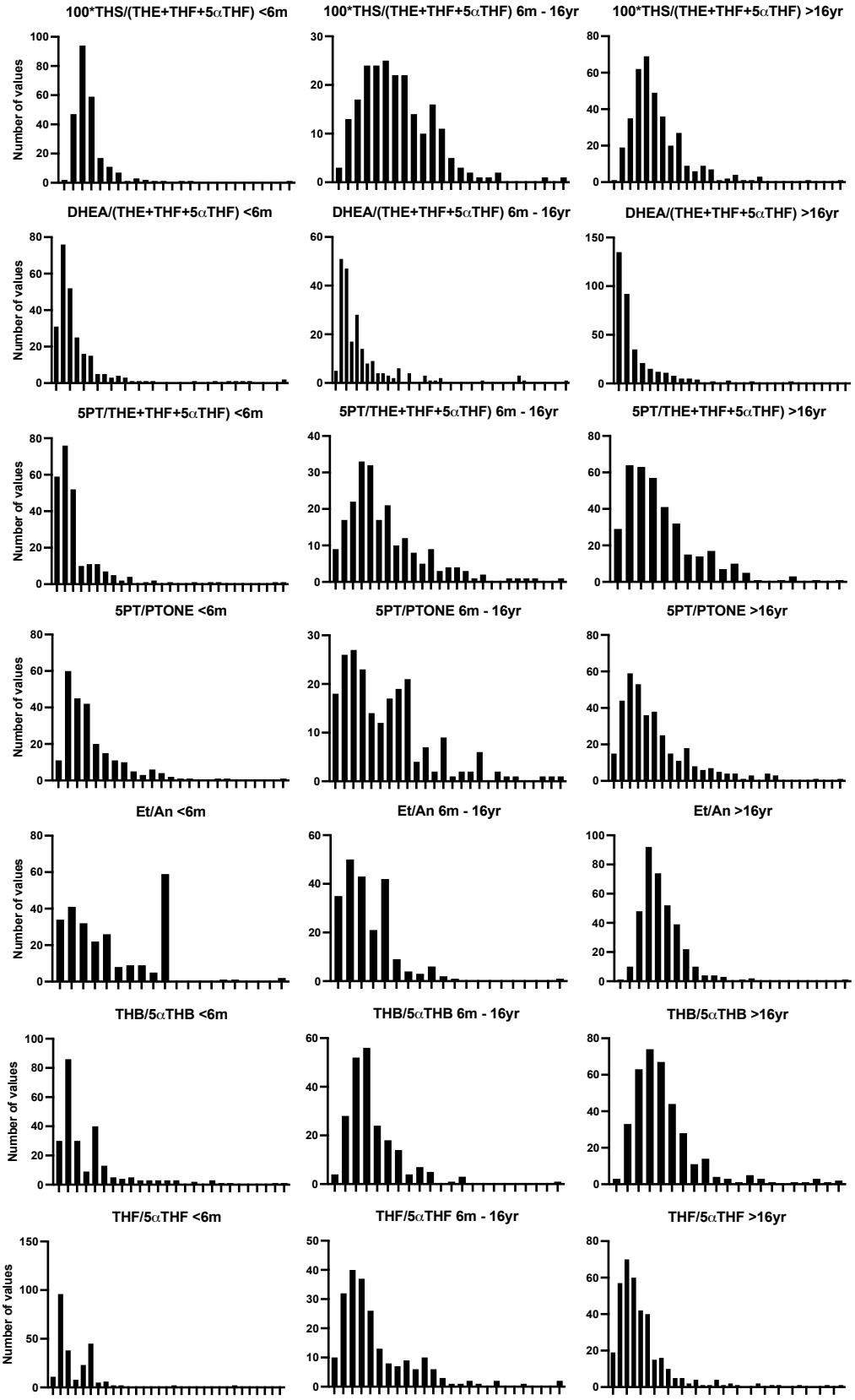


Figure 3.1: Frequency histograms for 14 biochemical ratios

Healthy control data is shown split by age. Bars correspond to number of subjects.

Table 3.2: Log-normality testing results

Shapiro-Wilk test was applied to log transformed data for each of the 14 biochemical ratios. Displayed are results for the 3 defined age groups- less than 6 months (<6m), 6 months to 16 years (16m-16yr) and over 16 years (>16yr). Comparison is made between male subjects (M), female subjects (F), and data for both sexes combined (M+F) for each of the defined age groups. Groups which show a normal distribution following log-transformation are indicated in red text.

Ratio	**21OHD/PORD Ratio 1: 100*PT'ONE/(THE+THF+5aTHF)**								
Sex + Age	**M < 6m**	**F < 6m**	**M+F < 6m**	**M 6m-16yr**	**F 6m-16yr**	**M+F 6m-16yr**	**M >16yr**	**F >16yr**	**M+F >16yr**
P value	<0.0001	<0.0001	<0.0001	<0.0001	0.0033	<0.0001	<0.0001	<0.0001	<0.0001
Lognormal distribution?	No	No	No	No	No	No	No	No	No
Ratio	**21OHD/PORD Ratio 2: (17HP+PT)/(THE+THF+5aTHF)**								
P value	0.0363	<0.0001	<0.0001	0.0002	0.0523	0.0011	0.2536	0.3910	0.8255
Lognormal distribution?	No	No	No	No	Yes	No	Yes	Yes	Yes
Ratio	**PORD Ratio 3: PD/(THE+THF+5aTHF)**								
P value	<0.0001	0.0032	<0.0001	0.0044	0.2330	0.0019	0.9912	<0.0001	<0.0001
Lognormal distribution?	No	No	No	No	Yes	No	Yes	No	No
Ratio	**PORD Ratio 4: 5PD/(THE+THF+5aTHF)**								
P value	0.2710	0.3911	0.1831	0.5597	<0.0001	<0.0001	0.4058	0.6777	0.8572
Lognormal distribution?	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes
Ratio	**PORD Ratio 5: (17HP+PT)/(An +Et)**								
P value	0.4877	<0.0001	<0.0001	0.0007	0.0081	<0.0001	0.0151	0.0034	0.0001
Lognormal distribution?	Yes	No	No	No	No	No	No	No	No
Ratio	**17αOHD Ratio 6: (THA+5aTHA+THB+5aTHB)/(THE+THF+5aTHF)**								
P value	0.0511	0.0104	0.0122	0.2437	0.3215	0.0899	0.1540	0.0017	0.0009
Lognormal distribution?	Yes	No	No	Yes	Yes	Yes	Yes	No	No
Ratio	**17αOHD Ratio 7: (THA+5aTHA+THB+5aTHB)/(An+Et)**								
P value	0.4731	0.0565	0.8926	<0.0001	<0.0001	<0.0001	0.2533	0.4042	0.3035
Lognormal distribution?	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes
Ratio	**11βOHD Ratio 8: 100*(THS)/(THE+THF+5aTHF)**								
P value	0.0001	<0.0001	<0.0001	0.0051	0.9234	0.2927	0.8267	0.2802	0.4086
Lognormal distribution?	No	No	No	No	Yes	Yes	Yes	Yes	Yes
Ratio	**3βHSD2D Ratio 9: DHEA/(THE+THF+5aTHF)**								
P value	0.0240	0.0023	0.0034	0.0075	0.4253	0.0241	0.0247	0.1318	0.3741
Lognormal distribution?	No	No	No	No	Yes	No	No	Yes	Yes
Ratio	**3βHSD2D Ratio 10: 5PT/(THE+THF+5aTHF)**								
P value	<0.0001	0.0003	<0.0001	<0.0001	<0.0001	<0.0001	0.9047	0.0555	0.1856
Lognormal distribution?	No	No	No	No	No	No	Yes	Yes	Yes
Ratio	**3βHSD2D Ratio 11: 5PT/PT'ONE**								
Lognormal distribution?	Yes	No	No	No	No	No	Yes	No	No
Ratio	**5αRD Ratio 12: Et/An**								
P value	0.0011	<0.0001	<0.0001	0.1584	0.5411	0.0723	0.0999	0.0048	0.0266
Lognormal distribution?	No	No	No	Yes	Yes	Yes	Yes	No	No
Ratio	**5αRD Ratio 13: THB/5aTHB**								
P value	0.0184	0.0029	0.0014	0.3473	0.2883	0.1008	0.3358	<0.0001	<0.0001
Lognormal distribution?	No	No	No	Yes	Yes	Yes	Yes	No	No
Ratio	**5αRD Ratio 14: THF/5aTHF**								
P value	0.0001	0.0003	<0.0001	0.1311	0.0489	0.0638	0.0201	<0.0001	<0.0001
Lognormal distribution?	No	No	No	Yes	No	Yes	No	No	No

3.5.2. Steroidogenesis over the life course

To examine the changes in the steroid metabolome observed with age we firstly looked specifically at urine samples from infants in their first week of life, with steroid metabolites extracted and identified by GC-MS in scan mode. We examined in detail the Total Ion Chromatogram (TIC) from 3 infants, selecting each peak in turn to identify the steroid metabolites from their mass spectra. Secondly, to examine the changes observed across the life course, we used the 14 stated biochemical ratios as surrogate markers of enzyme activity and examined how these vary over life. These results are presented in sequence by enzyme activity.

3.5.2.1. Early neonatal life

For the 3 infants examined, abundant 3β-hydroxy-5-ene steroids were observed, along with highly polar glucocorticoid and mineralocorticoid metabolites, mirroring previous findings by Shackleton (Shackleton et al., 1979). A table of steroids characterised from our samples is shown in **Table 3.3**.The TIC from a 3 day old baby, mass spectra for 16α–hydroxypregnenolone, one of the most abundantly seen steroid metabolites in neonatal urine samples, and corresponding underivatized and derivatised chemical structures are shown in **Figure 3.2.**

Table 3.3: Urinary steroid metabolites quantified from infants in the first week of life

1. The Delta-5 Steroids			2. Corticosteroid metabolites			3. Other steroid metabolites		
Excreted as sulphate Conjugates			**Origin Cortisol or corticosterone**			**Origin foeto-placental remnants and newborn**		
			Excreted free or as glucuronides			**Excreted as glucuronides**		
Trivial name *		**Systematic name**	**Trivial Name**		**Systematic name**	**Trivial Name**		**Systematic name**
Origin Pregnenolone and DHEA			***Origin cortisol***			***Origin androgens***		
Androstenediols		5-Androstene-3β,17α (and β)-diol	cortisol		4-Pregnene-11β,17α,21-triol-3,20-dione	Androsterone		5α-Androstane-3α-ol-17-one
16α-OH-DHEA		5-Androstene-3β,16α-ol-17-one	cortisone		4-Pregnene-17α,21-diol-11,20-dione	Etiocholanolone		5β-Androstane-3α-ol-17-one
16β−OH-DHEA	✓	5-Androstene-3β,16β-ol-17-one	6β-OH cortisol		4-Pregnene-6β,11β,17α,21-tetrol-3,20-dione	11-Oxo-androsterone		5β-Androstane-3α-ol 11,17-dione
16-Oxo-andostenediol	✓	5-Androstene-3β,17β-diol-16-one	20α-(or β)THE		4-Pregnane-17α,20α (or β),21-triol-3,11-dione	11β-hydroxyandrosterone		5α-Androstane-3α,11β-diol-17-one
15α,16α-DihydroxyDHEA	✓	5-Androstene-3β,15α,16α-ol-17-one	THE		5β-Pregnane-3α,17α,21-triol-11,20-dione	11β-hydroxyetiocholanolone		5β-Androstane-3α,11β-diol-17-one
Androstenetriol		5-Androstene-3β,16α,17β-triol	THF	✗	5β-Pregnane 3α,11β,17α,21-tetrol-20-one	***Origin progesterone, 17-OHP and 21-deoxy cortisol***		
16α,18-Di-OH DHEA	✓	5-Andostene-3β,16α,18-triol-17-one	5αTHF	✗	5β-Pregnane-3α,11β,17α,21-tetrol-20-one	Pregnanediol	✻	5β-Pregnane-3α,20α-diol
Androtenetetrol (15α)	✓	5-Androstene-3β,15α,16α,17β-tetrol	6α-hydroxyTHE	✓	5β-pregnane-3α,6α,17α,21-tetrol-11,20-dione	Pregnanetriol		5β-Pregnane-3α,17α,20α-triol
Androstenetetrol (18)	✓	5-Androstene-3β,16α,17β,18-tetrol	α-cortolone		5β-Pregnane-3α,17α,20α,21 tetrol-11-one	Pregnanetriolone		5β-pregnane-3α,17α,20α-triol-11-one
Pregnenediol(5PD)		5-Pregnene-3β,20α-diol	β-Cortolone		5β-Pregnane-3α,17α,20β,21-tetrol-11-one	6α-hydroxypregnanolone	✻	3α,6α-dihydroxy-5β-pregnane-20-one
16α-hydroxypregnenolone		5-Pregnene-3β,16α-ol-20-one	α-cortol	✗	5β-Pregnane-3α,11β,17α,20α,21-pentol	16α-hydroxypregnanolone	✻	3α,16α-dihydroxy-5β-pregnane-20-one
21-Hydroxypregnenolone	✓	5-Pregnene-3β,21-diol-20-one	β-Cortol	✗	5B-Pregnane-3α,11β,17α,20β,21-pentol	***Origin foetal 16α-hydroxyDHEA***		
5P-3,20,21-triol	✓	5-Pregnene-3β,20α,21-triol	6α-OH-α-cortolone	✓	5β-Pregnane-3α,6α,17α,20α,21-pentol-11-one	Oestriol	✻	
5P-3,16,20-triol	✓	5-Pregnene-3β,16α,20α-triol	6α-OH-β-cortolone	✓	5β-Pregnane-3α,6α,17α,20β,21-pentol-11-one			
15β,17α-OH-pregnenolone	✓	5 Pregnene-3β,15β,17α-triol-20-one	1β-OH-cortolone	✓	5β-Pregnane-1β,17α,20α,21-pentol-11-one			
			Origin Corticosterone					
			THA		5β-Pregnane-3α,21-diol-11,20-dione			
			5αTHA		5α-Pregnane-3α,21-diol-11,20-dione			
			THB	✗	5β-Pregnane-3α,11β,21-triol-20-one			
			5αTHB	✗	5α-Pregnane-3α,11β,21-triol-20-one			
			6α-OH-THA	✓	5α-Pregnane-3α,6α,21-triol-20-one			

	quantitatively major steroids
*	Steroids listed in approximate order of structural complexity, closely related to increasing GC retention time
✻	Early excretion of foeto-placental metabolites
✓	Unique to neonates, decrease over first months
✗	Low excretion of 11β-hydroxy steroids, increases during first months

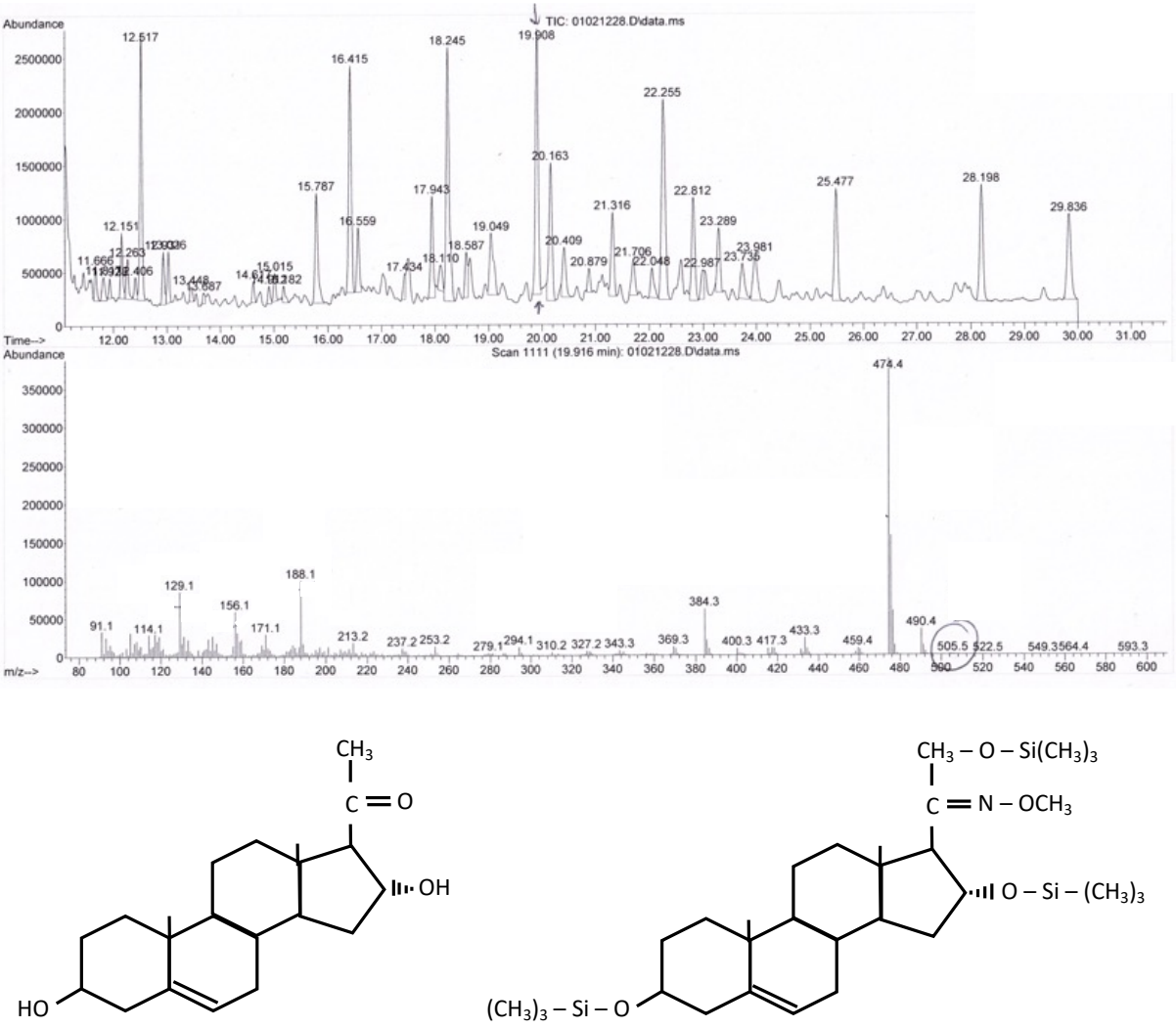


Figure 3.2: 16α-hydroxypregnenolone

Top panel shows Total Ion Chromatogram (TIC) from the urine sample of a 3-day old baby as analysed by GC-MS in Scan mode. TIC displays the summed intensities of all mass spectral peaks across one scan. Middle panel shows mass spectrum for 16α-hydroxypregnenolone, peak indicated by the arrow in the TIC panel. Bottom left panel shows the chemical structure for 16α-hydroxypregnenolone. Bottom right panel shows the chemical structure for the derivatised form of 16α-hydroxypregnenolone.

3.5.2.2. 21-Hydroxylase Activity

21-hydroxylase deficiency is indicated by derangement of two steroid metabolite ratios. Firstly, pregnanetriolone over the sum of tetrahydrocortisone, tetrahydrocortisol and 5α-tetrahydrocortisol (100*PTONE/THE+THF+5αTHF), which is a ratio of 21-deoxycortisol metabolite over downstream glucocorticoid metabolites. 21-deoxycortisol is produced through the action of 11β–hydroxylase on 17α-hydroxyprogesterone. It is a very sensitive biomarker for 21-hydroxylase deficiency but only indirectly represents 21-hydroxylase function. Secondly, the sum of 17-hydroxypregnanolone and pregnanetriol over the sum of tetrahydrocortisone, tetrahydrocortisol and 5α-tetrahydrocortisol, which is a ratio of 17α-hydroxyprogesterone metabolites over downstream glucocorticoid metabolites. These two ratios showed slightly different patterns over life. The ratio of 21-deoxycortisol over glucocorticoid metabolites appeared highest in early infancy (birth to 3 months), indicating lowest enzyme activity, then remained relatively static over life, with a median ranging from 0.18 (over 50 years group) to 0.66 (less than 1 month group). The widest degree of variability was observed in the under 1 month age group (interquartile range = 1.04, interdecile range 3.02). The most homogenous age group with the lowest observed variability was the 18-30 years group (interquartile range 0.18, interdecile range 0.32). The ratio of 17α-hydroxyprogesterone over glucocorticoid metabolites was of a 10-fold lower scale than the 21-deoxycortisol ratio (note most values are less than 0.3, compared to less than 3 for the 21-deoxycortisol over glucocorticoid metabolites ratio). Ratio values were low during infancy and early childhood but seemed to nadir at 1-4 years of age, began to rise during the adrenarche age range (4-10 years), indicating a decrease in enzyme activity, continued to rise into adulthood, peaking in the 18-30 age range then showing a decline thereafter, but

remaining above the values seen in infancy in the over 50 years age group. This suggests a slight increase in enzyme activity post adrenopause. Median varied from 0.02 (3 months to 4 years all groups inclusive) to 0.15 (18-30 years). The highest variability was seen in the 10-18 years and 30-50 years age groups (interquartile range 0.13 and 0.12 respectively, interdecile range 0.22 and 0.23 respectively). The most homogenous group with the lowest observed variability was the 1-4 years age group (interquartile range 0.01, interdecile range 0.03), though 1 month to 4 years inclusive all demonstrated homogenously low ratio values. These results are summarised in **Table 3.4** and graphically displayed in **Figure 3.3.**

Table 3.4: Changes observed in two ratios used for the diagnosis of 21-hydroxylase deficiency with age

*Male and female data are combined for each age group. Median, interquartile range (25th – 75th percentile) and interdecile range (10th – 90th percentile) are displayed as measures of spread for assumed non-normally distributed data. Top table: Ratio of 21-deoxycortisol metabolites over downstream glucocorticoid metabolites (100*PTONE/THE+THF+5αTHF). Bottom table: Ratio of 17α-hydroxyprogesterone metabolites over downstream glucocorticoid metabolites*

100*PTONE/ THE+THF+5aTHF	0-1m	1-3m	3-6m	6-12m	1-4yr	4-10yr	10-18yr	18-30yr	30-50yr	>50yr	min	max
Number of values	115	64	76	42	96	44	24	67	165	94	24	165
Median	0.66	0.41	0.23	0.18	0.20	0.23	0.22	0.21	0.20	0.18	0.18	0.66
Interquartile range	1.04	0.77	0.17	0.16	0.23	0.18	0.38	0.18	0.17	0.14	0.14	1.04
Interdecile range	3.02	2.54	0.56	0.55	0.85	0.51	0.97	0.32	0.40	0.67	0.32	3.02

17HP+PT/ THE+THF+5aTHF	0-1m	1-3m	3-6m	6-12m	1-4yr	4-10yr	10-18yr	18-30yr	30-50yr	>50yr	min	max
Number of values	115	72	76	42	96	44	24	67	165	94	24	165
Median	0.03	0.03	0.02	0.02	0.02	0.04	0.12	0.15	0.13	0.07	0.02	0.15
Interquartile range	0.03	0.01	0.02	0.02	0.01	0.03	0.13	0.09	0.12	0.05	0.01	0.13
Interdecile range	0.07	0.05	0.04	0.03	0.03	0.05	0.22	0.19	0.23	0.10	0.03	0.23

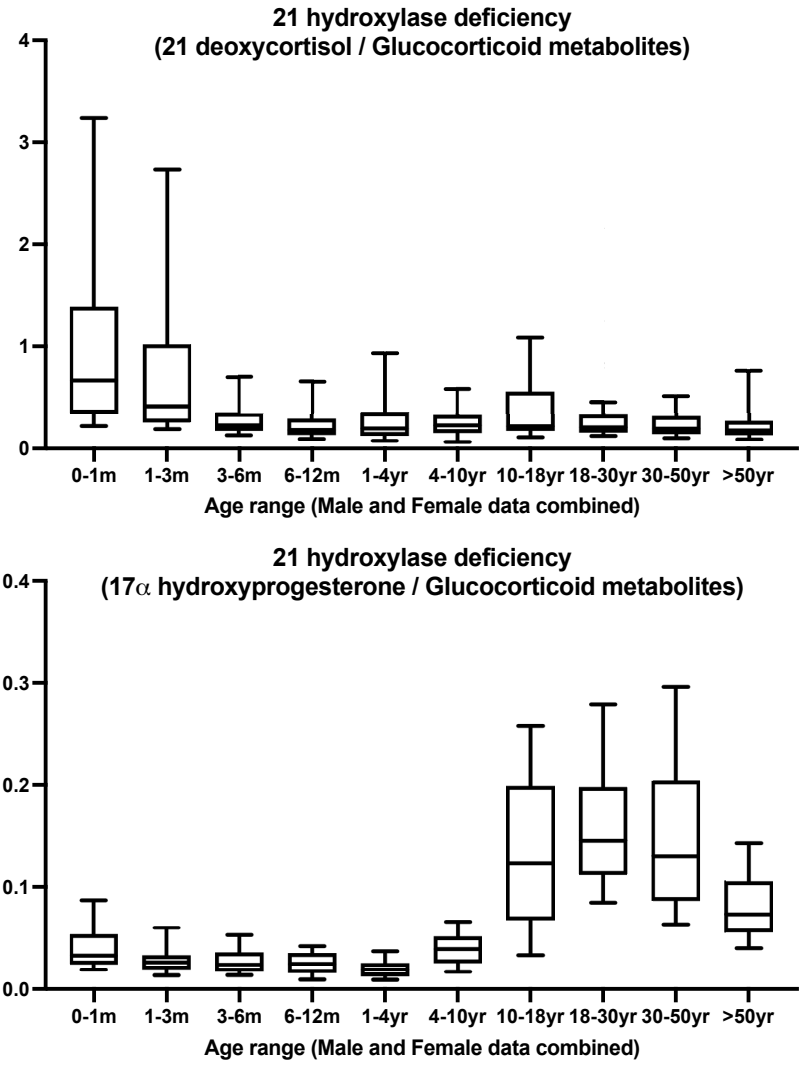


Figure 3.3: Changes observed in two ratios used for the diagnosis of 21-hydroxylase deficiency over the life course

*Male and female data are combined for each age group. Box plots display median and interquartile range (25th – 75th percentile), whiskers show interdecile range (10th - 90th percentile). Top panel: Ratio of 21-deoxycortisol metabolites over downstream glucocorticoid metabolites (100*PTONE/THE+THF+5αTHF). Bottom panel: Ratio of 17α-hydroxyprogesterone metabolites over downstream glucocorticoid metabolites (17HP+PT/THE+THF+5αTHF).*

3.5.2.3. 11β–Hydroxylase Activity

11β-hydroxylase activity is represented by the ratio of 11-deoxycortisol metabolite, tetrahydro-11-deoxycortisol (Watson et al.), over the sum of downstream glucocorticoid metabolites, tetrahydrocortisone, tetrahydrocortisol and 5α-tetrahydrocortisol (THE+THF+5αTHF). Ratio values remained around the range of 1 - 2 throughout life, with a median range of 0.86 (1-3 months of age) to 1.84 (1-4 years of age), indicating static enzyme activity throughout life. A small decline in enzyme activity, seen as the peak ratio value, was observed in the 1-4 years age group. The highest degree of variability was observed in the 0-1-month age group (interquartile range 0.83, interdecile range 2.3). The most homogenous groups with the lowest variability were the 3-6 month and 10-18 years groups (interquartile range 0.56 and 0.68 respectively, interdecile range 1.07 and 1.05 respectively). These results are summarised in **Table 3.5** and graphically represented in **Figure 3.4.**

Table 3.5: Changes observed in the ratio used for the diagnosis for 11β-hydroxylase deficiency over the life course

Ratio of 11-deoxycortisol metabolite over downstream glucocorticoid metabolites. Male and female data are combined for each age group. Median, interquartile range (25th – 75th percentile) and interdecile range (10th – 90th percentile) are displayed as measures of spread for assumed non-normally distributed data.

100*THS/ THE+THF+5aTHF	0-1m	1-3m	3-6m	6-12m	1-4yr	4-10yr	10-18yr	18-30yr	30-50yr	>50yr	min	max
Number of values	115	72	76	42	96	44	24	67	165	94	24	165
Median	1.36	0.86	1.05	1.50	1.84	1.18	1.13	1.01	1.05	1.22	0.86	1.84
Interquartile range	0.83	0.64	0.56	0.98	0.99	0.88	0.68	0.67	0.69	0.62	0.56	0.99
Interdecile range	2.30	1.26	1.07	1.62	1.86	1.53	1.05	1.42	1.55	1.46	1.05	2.30

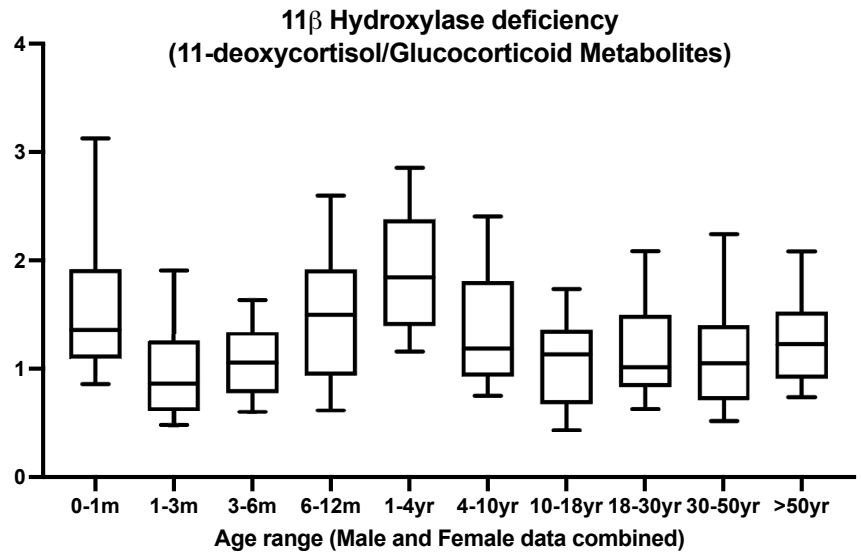


Figure 3.4: Changes observed in the ratio used for the diagnosis for 11β-hydroxylase deficiency over the life course

Ratio of 11-deoxycortisol metabolite over downstream glucocorticoid metabolites. Male and female data are combined for each age group. Box plots display median and interquartile range (25th – 75th percentile), whiskers show interdecile range (10th - 90th percentile).

3.5.3. P450 Oxidoreductase Activity

P450 oxidoreductase is not an enzyme but is a co-factor which enables function of cytochrome P450 enzymes such as 21-hydroxylase and 17α-hydroxylase, therefore diagnostic ratios employed will ideally represent combined function of these enzymes to hint at the activity of P450 oxidoreductase. Three ratios are used in our method to assist with the diagnosis of P450 oxidoreductase deficiency. 1) Ratio of pregnenolone metabolite, pregnenediol (5PD) over the sum of downstream glucocorticoid metabolites, tetrahydrocortisone, tetrahydrocortisol and 5α-tetrahydrocortisol

(THE+THF+5αTHF). 2) Ratio of progesterone metabolite, pregnanediol (Lee et al.) over the sum of downstream glucocorticoid metabolites (THE+THF+5αTHF). 3) The sum of 17α–hydroxylase metabolites, 17-hydroxypregnanolone and pregnanetriol (17HP+PT), over the sum of androgen metabolites, androsterone and etiocholanolone (An+Et). For investigation of the ratio of pregnenolone metabolite 5PD over glucocorticoid metabolites, the 4-10 years and 10-18 years age groups were combined to increase the class size.

The ratio of pregnenolone over glucocorticoid metabolites was highest (indicating lowest activity) from birth to 3 months, declined to nadir (indicating peak activity) between 4-18 years, rising to a lesser degree at 18-30 years, then steadily declined again thereafter. Median varies from 0.02 (4-18 years) to 0.27 (0-1 month and 1-3 months age groups). The highest degree of variability was observed in the 1-3 months age group (interquartile range 0.27, interdecile range 0.61). The most homogenous groups with the least variability were 1-4 years, 4-18 years and over 50 years (interquartile range 0.03, interdecile range 0.04 for all three groups). The progesterone over glucocorticoid metabolites ratio values were low throughout infancy, rising steadily from 1 year of age to a peak (indicating lowest activity) at 18-30 years then ratio values steadily decline thereafter. The median varied from 0.01 (birth to 4 years, all groups inclusive) to 0.06 (18-30 years). The highest variability was observed in the 18-30 years and 30-50 years age group (interquartile range 0.06 and 0.07 respectively, interdecile range 0.2 for both groups). The infant groups were relatively homogenous, with the least variability observed in the 1-3 months and 3-6 months age groups (interquartile and interdecile range 0.01 for both groups). The ratio of metabolites of 17α-hydroxyprogesterone over androgens showed values highest

(indicting lowest activity) in infancy, with a small dip in the 1-3 months age group and peak at 6-12 months. Ratio values then steadily declined and remain low throughout the adult years. Median varied from 0.07 (18-30 years age group) to 3.17 (6 – 12 months age group). The highest variability was observed in the infant groups, particularly the 0–1-month age group (interquartile range 2.08, interdecile range 3.62). The adult groups showed the least variability with the most homogenous group being 18-30 years (interquartile range 0.11, interdecile range 0.22). These results are summarised in **Table 3.6** and graphically represented in **Figure 3.5.**

Table 3.6: Changes observed in the ratios used for the diagnosis of P450 oxidoreductase deficiency over the life course

*Male and female data are combined for each age group. Median, interquartile range (25th – 75th percentile) and interdecile range (10th – 90th percentile) are displayed as measures of spread for assumed non-normally distributed data. Top table: Ratio of pregnenolone metabolite, (5PD) over the sum of downstream glucocorticoid metabolites, (THE+THF+5αTHF). Middle table: Ratio of progesterone metabolite (*Lee et al.*) over the sum of downstream glucocorticoid metabolites, (THE+THF+5αTHF). Bottom table: Ratio of the sum of 17α–hydroxylase metabolites (17HP+PT) over the sum of androgen metabolites (An+Et).*

5PD/THE+THF+5aTHF	0-1m	1-3m	3-6m	6-12m	1-4yr	4-18yr	18-30yr	30-50yr	>50yr	min	max
Number of values	115	72	76	42	90	18	50	122	94	18	122
Median	0.27	0.27	0.11	0.06	0.04	0.02	0.08	0.06	0.03	0.02	0.27
Interquartile range	0.26	0.27	0.10	0.06	0.03	0.03	0.06	0.06	0.03	0.03	0.27
Interdecile range	0.38	0.61	0.14	0.09	0.04	0.04	0.15	0.11	0.04	0.04	0.61

PD/THE+THF+5aTHF	0-1m	1-3m	3-6m	6-12m	1-4yr	4-10yr	10-18yr	18-30yr	30-50yr	>50yr	min	max
Number of values	115	72	76	42	96	43	24	67	165	94	24	165
Median	0.01	0.01	0.01	0.01	0.01	0.02	0.03	0.06	0.04	0.02	0.01	0.06
Interquartile range	0.01	0.01	0.01	0.01	0.02	0.03	0.04	0.06	0.07	0.02	0.01	0.07
Interdecile range	0.03	0.01	0.01	0.02	0.04	0.03	0.06	0.20	0.20	0.03	0.01	0.20

17HP+PT/An+Et	0-1m	1-3m	3-6m	6-12m	1-4yr	4-10yr	10-18yr	18-30yr	30-50yr	>50yr	min	max
Number of values	115	72	76	42	96	44	24	67	165	94	24	165
Median	2.00	1.43	2.80	3.17	1.89	0.60	0.28	0.17	0.21	0.24	0.17	3.17
Interquartile range	2.08	1.13	1.94	2.30	1.24	0.47	0.13	0.11	0.12	0.11	0.11	2.30
Interdecile range	3.62	2.17	2.76	3.17	2.42	0.79	0.37	0.22	0.33	0.28	0.22	3.62

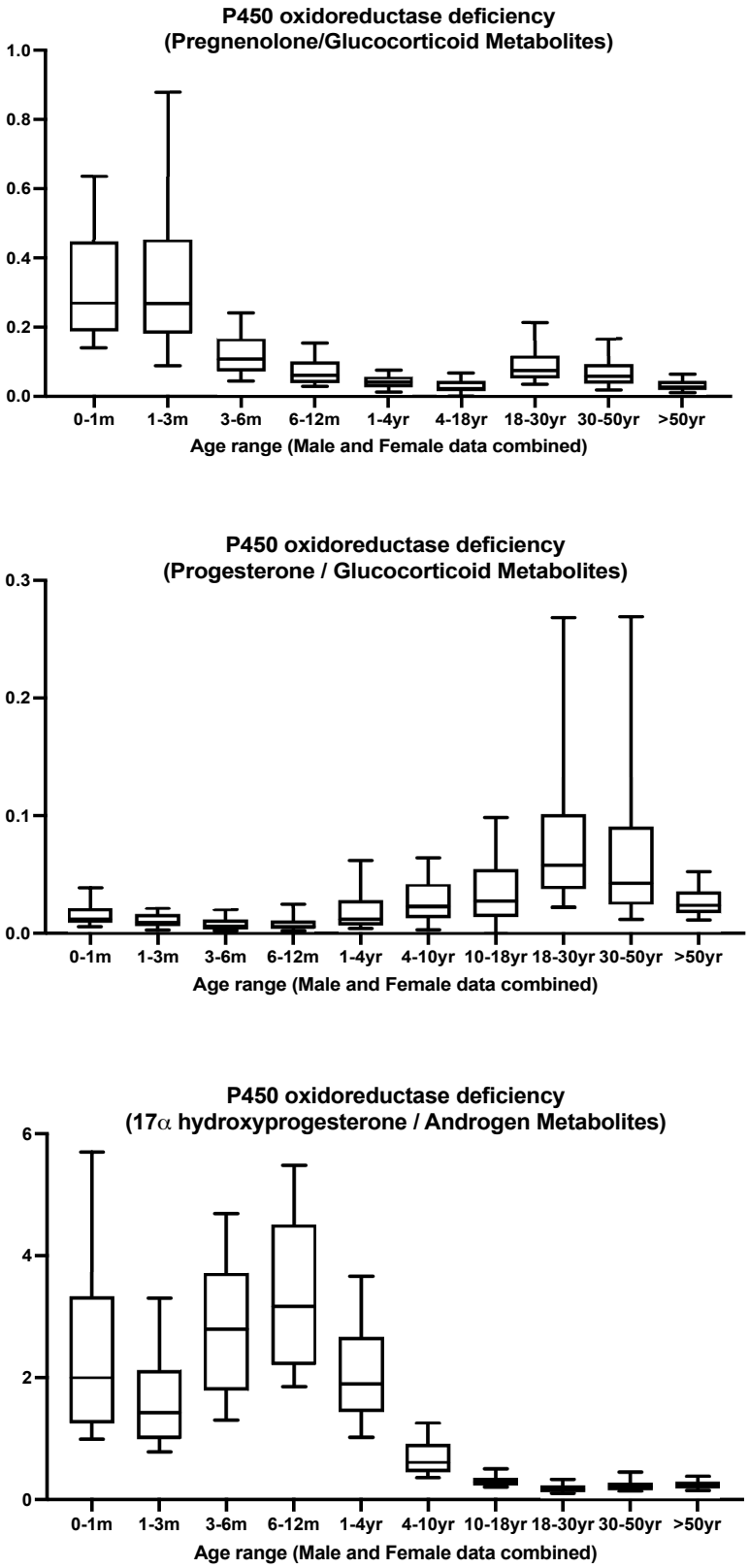


Figure 3.5: Changes observed in the ratios used for the diagnosis of P450 oxidoreductase deficiency over the life course

Figure 3.5: *Male and female data are combined for each age group. Box plots show median, and interquartile range (25th – 75th percentile), whiskers represent interdecile range (10th – 90th percentile). Top panel: Ratio of pregnenolone metabolite, (5PD) over the sum of downstream glucocorticoid metabolites, (THE+THF+5αTHF). Middle panel: Ratio of progesterone metabolite (*Lee et al.*) over the sum of downstream glucocorticoid metabolites, (THE+THF+5αTHF). Bottom panel: Ratio of the sum of 17α-hydroxylase metabolites (17HP+PT) over the sum of androgen metabolites (An+Et).*

3.5.3.1. 17α-Hydroxylase Activity

We have investigated two ratios used for the diagnosis of 17α-hydroxylase deficiency. The first ratio is of the sum of corticosterone and 11-deoxycorticosterone metabolites; tetrahydro-11-dehydrocorticosterone, tetrahydrocorticosterone and their 5α– reduced counterparts (THA+5αTHA+THB+5αTHB) over the sum of downstream glucocorticoid metabolites (THE+THF+5αTHF). The second ratio is the sum of corticosterone and 11-deoxycorticosterone metabolites (THA+5αTHA+THB+5αTHB) over the sum of androgen metabolites, androsterone and etiocholanolone (An+Et). The first ratio of mineralocorticoids over glucocorticoids was relatively static through life. The median ranged from 0.09 (over 50 years) to 0.23 (1-3 months). The highest variability was observed pre-adrenarche, particularly in the 1-3 months age group (interquartile range 0.15, interdecile range 0.25). The age groups were more homogenous with increasing age, with lowest variability observed in the over 50 years group (interquartile range 0.04, interdecile range 0.11). The second ratio of mineralocorticoid metabolites over androgen metabolites showed the highest values pre-adrenarche, with a steady rise

to a peak (suggesting lowest enzyme function) at 6-12 months of age. Post-adrenarche, ratio values reduced. Of note, the scale for this ratio was much higher than for most of the ratios we examined. The median ranged from 0.14 (18-30 years age group) to 30.8 (6-12 months age group). The highest variability was observed pre-adrenarche, maximally in the 6-12 months age group (interquartile range 19.1, interdecile range 55.4). The adult groups were comparatively more homogenous, with the least variability observed in the 18-30 years age group (interquartile range 0.11, interdecile range 0.19). These results are summarised in **Table 3.7** and shown in **Figure 3.6.**

Table 3.7: Changes observed in the ratios used for the diagnosis of 17α-hydroxylase deficiency with age

Male and female data are combined for each age group. Median, interquartile range (25th – 75th percentile) and interdecile range (10th – 90th percentile) are displayed as measures of spread for assumed non-normally distributed data. Top table: Ratio of sum of corticosterone and 11-deoxycorticosterone metabolites over the sum of downstream glucocorticoid metabolites. Bottom table: Ratio of sum of corticosterone and 11-deoxycorticosterone metabolites over sum of androgen metabolites.

THA+5aTHA+THB+5aTHB/ THE+THF+5aTHF	0-1m	1-3m	3-6m	6-12m	1-4yr	4-10yr	10-18yr	18-30yr	30-50yr	>50yr	min	max
Number of values	115	72	76	42	96	44	24	67	165	94	24	165
Median	0.18	0.23	0.23	0.22	0.22	0.14	0.13	0.11	0.11	0.09	0.09	0.23
Interquartile range	0.15	0.15	0.12	0.10	0.13	0.07	0.06	0.07	0.05	0.04	0.04	0.15
Interdecile range	0.19	0.25	0.22	0.26	0.21	0.21	0.13	0.15	0.13	0.11	0.11	0.26

THA+5aTHA+THB+5aTHB/ An+Et	0-1m	1-3m	3-6m	6-12m	1-4yr	4-10yr	10-18yr	18-30yr	30-50yr	>50yr	min	max
Number of values	115	72	76	42	96	44	24	67	165	94	24	165
Median	9.84	12.71	24.70	30.77	22.87	2.61	0.27	0.14	0.19	0.27	0.14	30.77
Interquartile range	13.83	8.61	19.68	19.09	25.87	3.64	0.55	0.11	0.14	0.21	0.11	25.87
Interdecile range	20.57	19.17	28.07	55.42	29.97	6.13	0.75	0.19	0.20	0.51	0.19	55.42

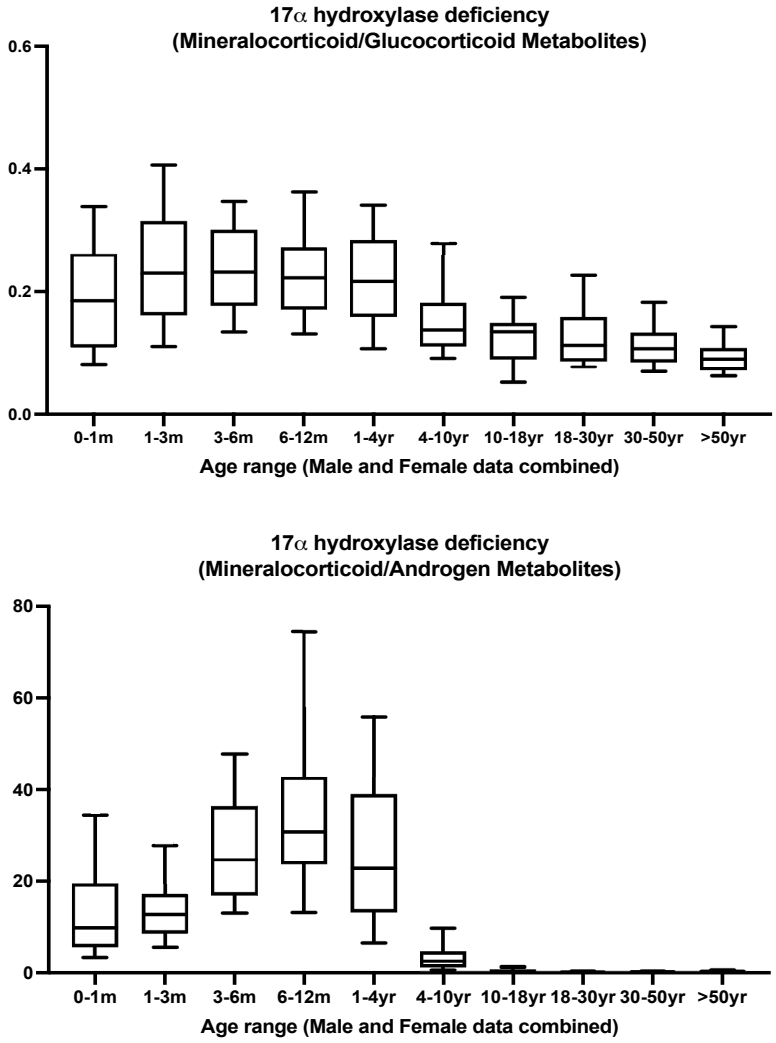


Figure 3.6: Changes observed in the ratios used for the diagnosis of 17α-deficiency over the life course

Male and female data are combined for each age group. Box plots show median and interquartile range (25th – 75th percentile), whiskers represent interdecile range (10th – 90th percentile). Top panel: Ratio of the sum of corticosterone and 11-deoxycorticosterone metabolites over the sum of downstream glucocorticoid metabolites ((THA+5αTHA+THB+5αTHB)/(THE+THF+5αTHF)). Bottom panel: Ratio of the sum of corticosterone and 11-deoxycorticosterone metabolites over the sum of androgen metabolites ((THA+5αTHA+THB+5αTHB)/(An+Et)).

3.5.3.2. 3β-Hydroxysteroid Dehydrogenase Activity

Three ratios are investigated here as proxies for the activity of 3β-hydroxysteroid dehydrogenase type 2. 1) A ratio of DHEA over the sum of downstream glucocorticoid metabolites, tetrahydrocortisone, tetrahydrocortisol and 5α–tetrahydrocortisol (THE+THF+5αTHF). 2) A ratio of the metabolite of 17-hydroxypregnenolone, pregnenetriol (5PT), over the sum of downstream glucocorticoid metabolites (THE+THF+5αTHF). 3) A ratio of the metabolite of 17-hydroxypregnenolone, pregnenetriol (5PT), over the metabolite of 21-deoxycortisol, pregnanetriolone (PTONE). The first ratio of DHEA over glucocorticoid metabolites was suppressed during childhood, with a small decline from birth to 6-12 months of age, increasing post puberty 40-fold by 18-30 years of age (suggesting reduced enzyme activity), then showed a steady decline (suggesting increasing activity) to reach a quarter of the peak value over the age of 50 years. The median ranged from 0.002 (6-12 months of age) to 0.081 (18-30 years). The highest variability was observed in the adult age ranges, with the most heterogeneous groups being 18-30 and 30-50 years (interquartile range 0.275 and 0.154 respectively, interdecile range 0.272 and 0.286 respectively). The least variability was observed in childhood from mid infancy to adrenarche (3 months-10 years), with the most homogenous group being 6-12 months (interquartile range 0.001, interdecile range 0.003). The second ratio of 17-hydroxypregnenolone over glucocorticoid metabolites showed a biphasic pattern over the life course with two value peaks at 1-3 months and 18-30 years, with a value nadir (suggesting peak activity) around the time of adrenarche and pubarche and a later steady decline in older adulthood. The lowest median of 0.01 was seen from 1 year - 18 years, the two peak medians observed were 0.03 at 1-3 months of age and 0.06 at 18-30 years of

age. The most heterogenous group with the highest variability was the 1-3 months age group (interquartile range 9.9, interdecile range 11.8). The third ratio of metabolites of 17-hydroxyprenenolone over 21-deoxycortisol shows a similar biphasic pattern to the other two ratios used here, but with values notably up to 1000 times greater (minimum median of 1.97 compared to 0.002). There was a small rise in values observed from 1 month to 12 months, declining to a nadir between 4 and 18 years. A steep rise, suggesting reducing enzyme activity, was then observed to a second much larger peak observed in the 18 – 30 years age range, followed by a steady decline in later adulthood. The highest variability was again observed in the 18-30 and 30-50 years age groups (interquartile range 26.7 and 25.1 respectively, interdecile range 51.9 and 35.9 respectively). The most homogenous groups displaying the least variability were 0-1 months of age and 1-4 years of age (interquartile range 4.15 and 4 respectively, interdecile range 4.4 and 3.5 respectively). These results consistently suggest reduced enzyme activity from adrenarche onwards. Results are summarised in **Table 3.8** and graphically displayed in **Figure 3.7.**

Table 3.8: Changes observed in the ratios used for the diagnosis of 3β-hydroxysteroid dehydrogenase type 2 deficiency over the life course

Male and female data are combined for each age group. Median, interquartile range (25th – 75th percentile) and interdecile range (10th – 90th percentile) are displayed as measures of spread for assumed non-normally distributed data. Top table: Ratio of DHEA over the sum of downstream glucocorticoid metabolites over the sum of downstream glucocorticoid metabolites. (DHEA/THE+THF+5αTHF). Middle table: Ratio of 17-hydroxypregnenolone metabolite over the sum of downstream glucocorticoid metabolites. (5PT/THE+THF+5αTHF). Bottom table: Ratio of 17-hydroxypregnenolone metabolite over 21-deoxycortisol metabolite (5PT/PTONE).

DHEA/THE+THF+5aTHF	0-1m	1-3m	3-6m	6-12m	1-4yr	4-10yr	10-18yr	18-30yr	30-50yr	>50yr	min	max
Number of values	115	72	76	42	96	44	24	67	165	94	24	165
Median	0.012	0.010	0.003	0.002	0.003	0.005	0.013	0.081	0.055	0.020	0.002	0.081
Interquartile range	0.015	0.016	0.002	0.001	0.003	0.006	0.029	0.275	0.154	0.051	0.001	0.275
Interdecile range	0.023	0.042	0.007	0.003	0.005	0.011	0.023	0.272	0.286	0.079	0.003	0.286

5PT/THE+THF+5aTHF	0-1m	1-3m	3-6m	6-12m	1-4yr	4-10yr	10-18yr	18-30yr	30-50yr	>50yr	min	max
Number of values	115	72	76	42	96	44	24	67	165	94	24	165
Median	0.02	0.03	0.02	0.02	0.01	0.01	0.01	0.06	0.04	0.02	0.01	0.06
Interquartile range	0.01	0.05	0.01	0.01	0.01	0.01	0.02	0.04	0.04	0.02	0.01	0.05
Interdecile range	0.04	0.08	0.02	0.02	0.01	0.01	0.01	0.07	0.06	0.03	0.01	0.08

5PT/PTONE	0-1m	1-3m	3-6m	6-12m	1-4yr	4-10yr	10-18yr	18-30yr	30-50yr	>50yr	min	max
Number of values	115	64	76	42	96	44	24	67	165	94	24	165
Median	3.00	6.08	7.87	7.59	4.00	2.84	1.97	24.86	21.35	12.68	1.97	24.86
Interquartile range	4.15	9.93	6.84	7.48	5.00	5.20	8.46	26.71	25.14	13.49	4.15	26.71
Interdecile range	4.44	11.78	9.56	9.68	3.53	3.80	8.50	51.89	35.87	25.80	3.53	51.89

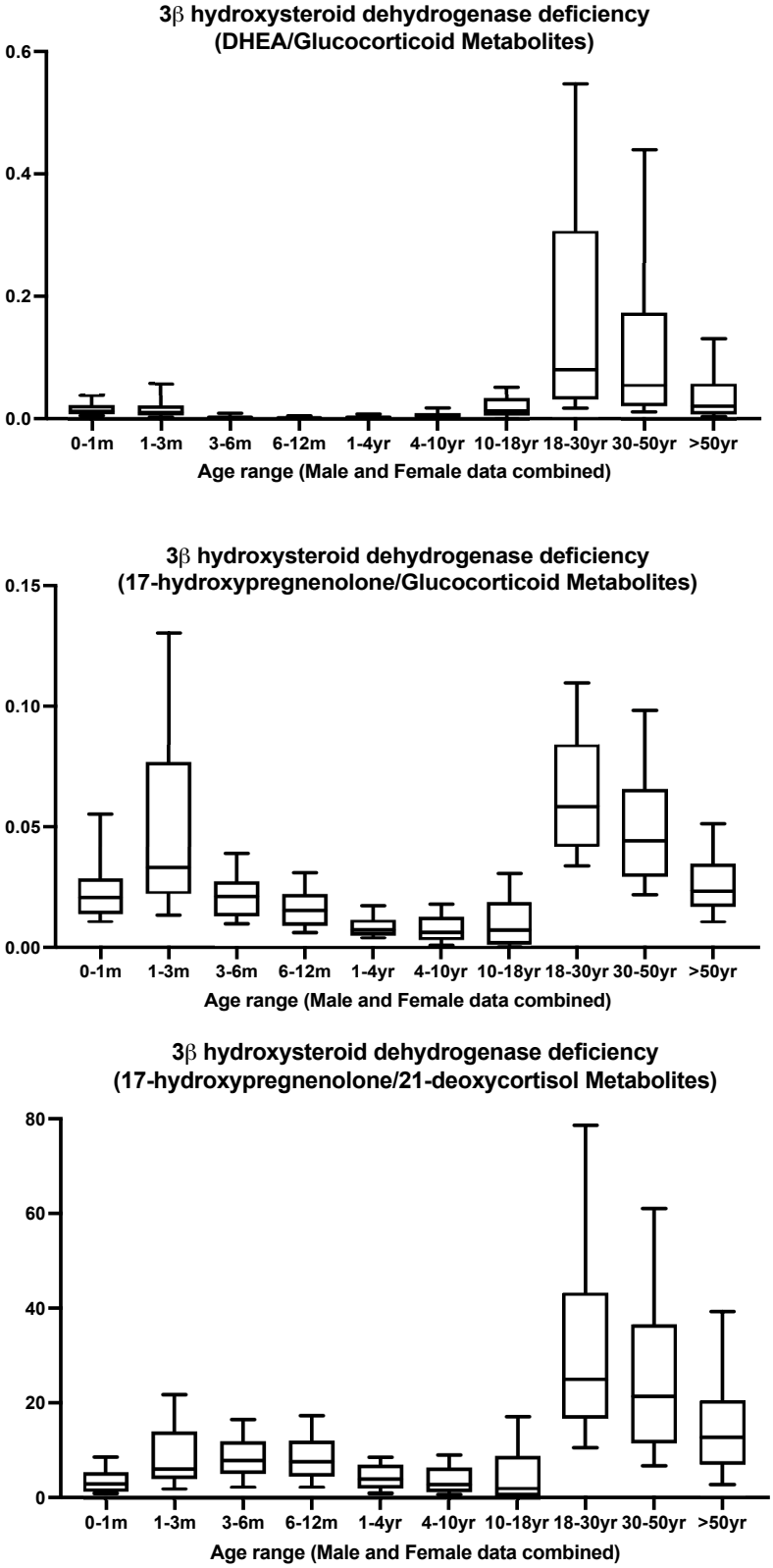


Figure 3.7 Changes observed in the ratios used for the diagnosis of 3β-hydroxysteroid dehydrogenase type 2 deficiency over the life course

Figure 3.7: *Male and female data are combined for each age group. Box plots show median and interquartile range (25th – 75th percentile), whiskers represent interdecile range (10th – 90th percentile). Top panel: Ratio of DHEA over the sum of downstream glucocorticoid metabolites over the sum of downstream glucocorticoid metabolites. (DHEA/THE+THF+5αTHF). Middle panel: Ratio of 17-hydroxypregnenolone metabolite over the sum of downstream glucocorticoid metabolites. (5PT/THE+THF+5αTHF). Bottom panel: Ratio of 17-hydroxypregnenolone metabolite over 21-deoxycortisol metabolite (5PT/PTONE).*

3.5.3.3. 5α-Reductase Type 2 Activity

5α-reductase type 2 converts testosterone to dihydrotestosterone. The urinary metabolite of dihydrotestosterone is androsterone, whereas other androgens (androstenedione, and testosterone) are metabolised and excreted in urine as both androsterone (An) and etiocholanolone (Et). Therefore, looking at a ratio of etiocholanolone over androsterone can be informative of this condition. We have also examined two other ratios which include a corticosterone metabolite and cortisol metabolite as these are both metabolised and excreted in urine also in a 5α–reduced form: tetrahydrocorticosterone and tetrahydrocortisol, and their 5α– reduced counterparts 5α-tetrahydrocorticosterone and 5α-tetrahydrocortisol (THB/5αTHB and THF/5αTHF). The ratio of Et/An dropped after birth to a nadir at 1-3 months then steadily continued to rise from 12 months onwards, ongoing through adulthood. The median ranged from 0.24 in the 1-3 months age group to 1.04 in the over 50 years group. The most variability was exhibited in the over 50 years age group (interquartile

range 0.66, interdecile range 1.32). All age groups displayed similar degrees of heterogeneity, with the lowest variability observed in the 3-6 months age group (interquartile range 0.33, interdecile range 0.5). The second ratio examined (THB/5αTHB), showed ratio values highest in the early neonatal period (0-3 months inclusive, median 1.0 and 0.5 respectively). These groups were also the most heterogeneous, interquartile range 1.03 and 0.67 respectively, interdecile range 1.77 and 2.81 respectively. This is likely a feature of normal neonatal and infant physiology as THB and 5αTHB are rarely detectable in the early neonatal period, and a ratio of two metabolites below the limit of quantification will be 1. Ratio values continued to be raised into the 1-3 months age group. The median for all other age groups was consistently less than 0.5, with a range of 0.18 (6-12 months) to 0.49 (over 50yr age group). The least variability was observed in the 6-12 month and 1-4 years age groups (interquartile range 0.09 and 0.1 respectively, interdecile range 0.26 and 0.25 respectively). The third ratio of THF/5αTHF showed a u-shaped distribution throughout childhood. The peak in the early neonatal period is similarly likely to be due to low excretion noted in early infancy but this isn't sustained beyond one month. Ratio values nadir at 3-6 months of age, suggesting peak activity, then steadily rise into adulthood when values reach a plateau which is sustained throughout adulthood. Median nadir was 0.18 with the peak median at 0-1 month of 1 and later median peak at 30-50 years of 1.27. The highest variability was seen in the over 50 years age group (interquartile range 0.9, interdecile range 2.02). These results suggest low enzyme activity in early infancy, increasing in childhood and activity declines in adulthood. Results are summarised in **Table 3.9** and graphically displayed in **Figure 3.8.**

Table 3.9: Changes observed in the ratios used for the diagnosis of 5α-reductase type 2 deficiency over the life course

Male and female data are combined for each age group. Median, interquartile range (25th – 75th percentile) and interdecile range (10th – 90th percentile) are displayed as measures of spread for assumed non-normally distributed data. Top table: Ratio of etiocholanolone over androsterone, a proxy of the ratio of DHT over the sum of Androstenedione + Testosterone + DHT. (Et/An). Middle table: Ratio of tetrahydrocorticosterone over 5α-tetrahydrocorticosterone (THB/5αTHB). Bottom table: Ratio of tetrahydrocortisol over 5α-tetrahydrocortisol (THF/5αTHF)

Et/An	0-1m	1-3m	3-6m	6-12m	1-4yr	4-10yr	10-18yr	18-30yr	30-50yr	>50yr	min	max
Number of values	115	72	76	42	96	44	24	67	165	94	24	165
Median	0.67	0.24	0.33	0.32	0.64	0.81	0.84	0.94	0.95	1.04	0.24	1.04
Interquartile range	0.64	0.49	0.33	0.43	0.59	0.52	0.64	0.64	0.60	0.66	0.33	0.66
Interdecile range	0.36	0.51	0.50	0.57	0.75	0.86	0.90	1.10	1.00	1.32	0.36	1.32

THB/5aTHB	0-1m	1-3m	3-6m	6-12m	1-4yr	4-10yr	10-18yr	18-30yr	30-50yr	>50yr	min	max
Number of values	115	72	76	42	96	44	24	67	165	94	24	165
Median	1.00	0.51	0.21	0.18	0.19	0.23	0.27	0.48	0.44	0.49	0.18	1.00
Interquartile range	1.03	0.67	0.11	0.09	0.10	0.18	0.22	0.23	0.26	0.34	0.09	1.03
Interdecile range	1.77	2.81	0.25	0.26	0.25	0.25	0.44	0.69	0.58	0.66	0.25	2.81

THF/5aTHF	0-1m	1-3m	3-6m	6-12m	1-4yr	4-10yr	10-18yr	18-30yr	30-50yr	>50yr	min	max
Number of values	115	72	76	42	96	44	24	67	165	94	24	165
Median	1.00	0.28	0.18	0.25	0.39	0.84	1.08	1.26	1.27	1.24	0.18	1.27
Interquartile range	0.49	0.21	0.09	0.18	0.18	0.52	0.47	0.82	0.89	0.90	0.09	0.90
Interdecile range	1.04	1.06	0.24	0.24	0.56	0.86	1.61	1.65	1.44	2.02	0.24	2.02

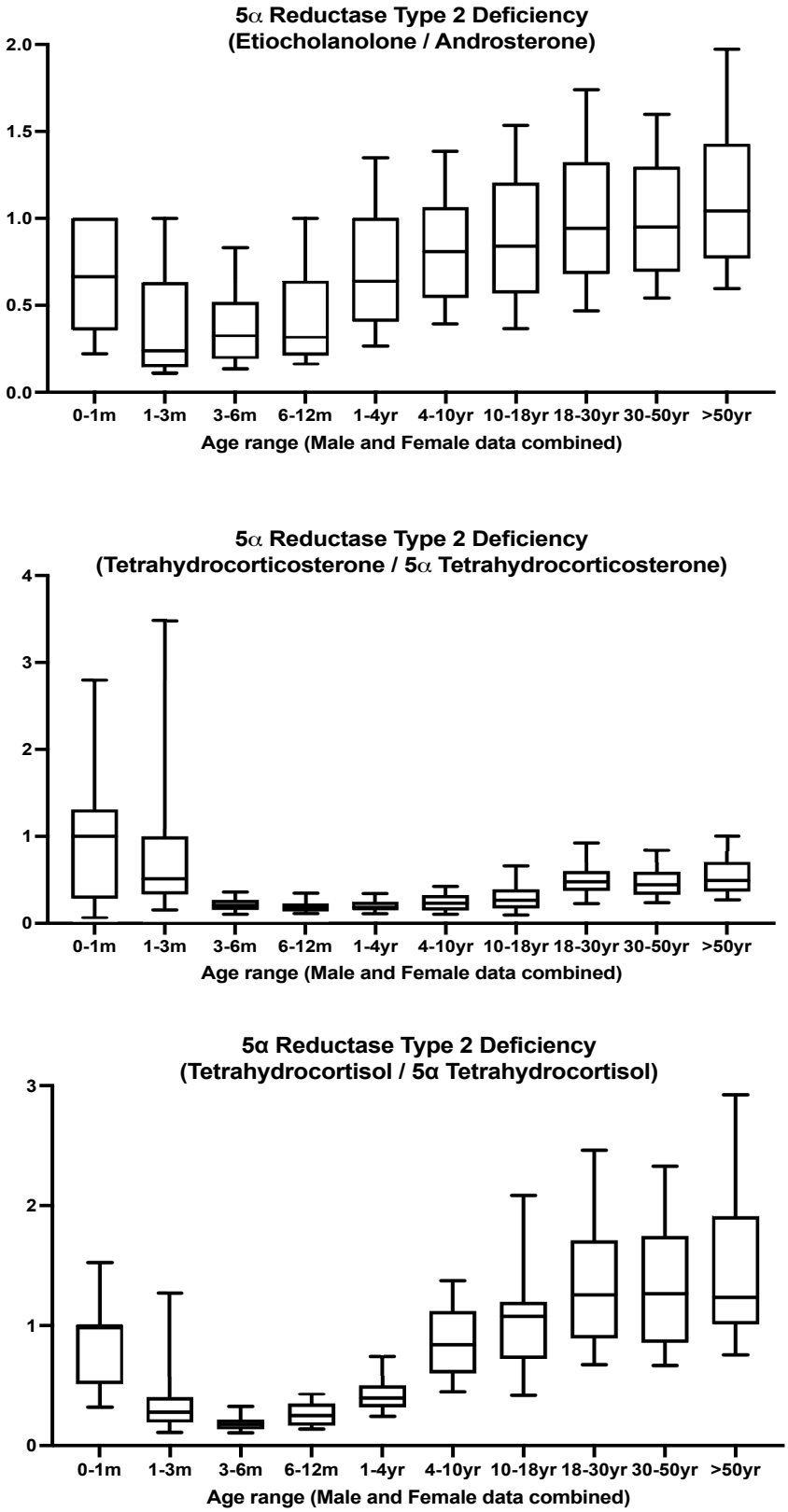


Figure 3.8: Changes observed in the ratios used for the diagnosis of 5α-reductase type 2 deficiency over the life course

Figure 3.8: *Male and female data are combined for each age group. Box plots show median and interquartile range (25th – 75th percentile), whiskers represent interdecile range (10th – 90th percentile). Top table: Ratio of etiocholanolone over androsterone, a proxy of the ratio of DHT over the sum of Androstenedione + Testosterone + DHT. (Et/An). Middle table: Ratio of tetrahydrocorticosterone over 5α-tetrahydrocorticosterone (THB/5αTHB). Bottom table: Ratio of tetrahydrocortisol over 5α-tetrahydrocortisol (THF/5αTHF)*

3.5.4. Effect of Sex and Age on Enzyme Activity over the Life Course

Figure 3.9 shows how each ratio varied across the 3 age groups. Results are shown as scatter plots, each dot representing a different subject, presented on a logarithmic scale with the median and interquartile range displayed. The order of magnitude for the calculated ratios varied considerably between the different ratios. The ratios that contain androgen metabolites (An+Et) as the denominator showed the biggest difference between the under 6 month and over 16 years age groups with the post pubertal values being much lower than the prepubertal values. There was a subsequent wide degree of variation in the corresponding 6 months to 16 years age group with the distribution in this group representing the transition between these two life stages. This was most pronounced for the mineralocorticoid metabolites over androgen metabolites ratio (THA+5αTHA+THB+5aTHB/An+Et), seen in the figures in light pink, where the median for the under 6 months group was 13.4 and the median for the over 16 years group was 0.2, resulting in a 6632% magnitude of change, suggesting a substantial increase in 17α-hydroxylase activity. This was also observed

to a lesser degree for the 17α-hydroxyprogesterone metabolites over androgen metabolites ratio (17HP+PT/An+Et), seen in orange in the figures, where the median for the under 6 months group was 1.85 and the median for the over 16 years group was 0.22, a 752% magnitude of change. The DHEA over glucocorticoid metabolites ratio (DHEA/THE+THF+5αTHF) shown in dark pink, showed a large percentage magnitude of change between the 6 months to 16 years group and the over 16 years group, albeit not a large numerical change, with the median ranging from 0.002 in the 6 months to 16 years group to 0.04 in the over 16 years group, suggesting a decline in HSD3B2 activity. The 17-hydroxypregnenolone over 21-deoxycortisol ratio (5PT/PTONE), dark green in the figure, was observed to have a large numerical increase in the over 16 years age group, with a median of 18.7 compared to 5 and 4.6 in the under 6 months and 6 months to 16 years age group. This, again, suggests a decline in HSD3B2 activity with age. The THB/5αTHB ratio varied widely for the under 6 months group which is likely due to the delayed emergence of these metabolites in reliably measurable quantities in early infancy. A similar observation is made for the THF/5αTHF ratio but to a much lesser degree. The least variation over life and within each age group was seen for the 11-deoxycortisol metabolite over glucocorticoid metabolites ratio (100*THS/THE+THF+5αTHF) and the mineralocorticoids over glucocorticoids ratio (THA+5αTHA+THB+5αTHB/THE+THF+5αTHF).

Figure 3.10 shows how the ratios vary over age for female and male healthy subjects separately. Results are shown as scatter plots, each dot representing a different subject, presented on a logarithmic scale with the median and interquartile range displayed. This analysis shows a very similar distribution and pattern of change across each of the age groups, observed for both males and females for all ratios. There were

minimal visibly observable variations between the magnitude of calculated ratio values seen in male and female subjects.

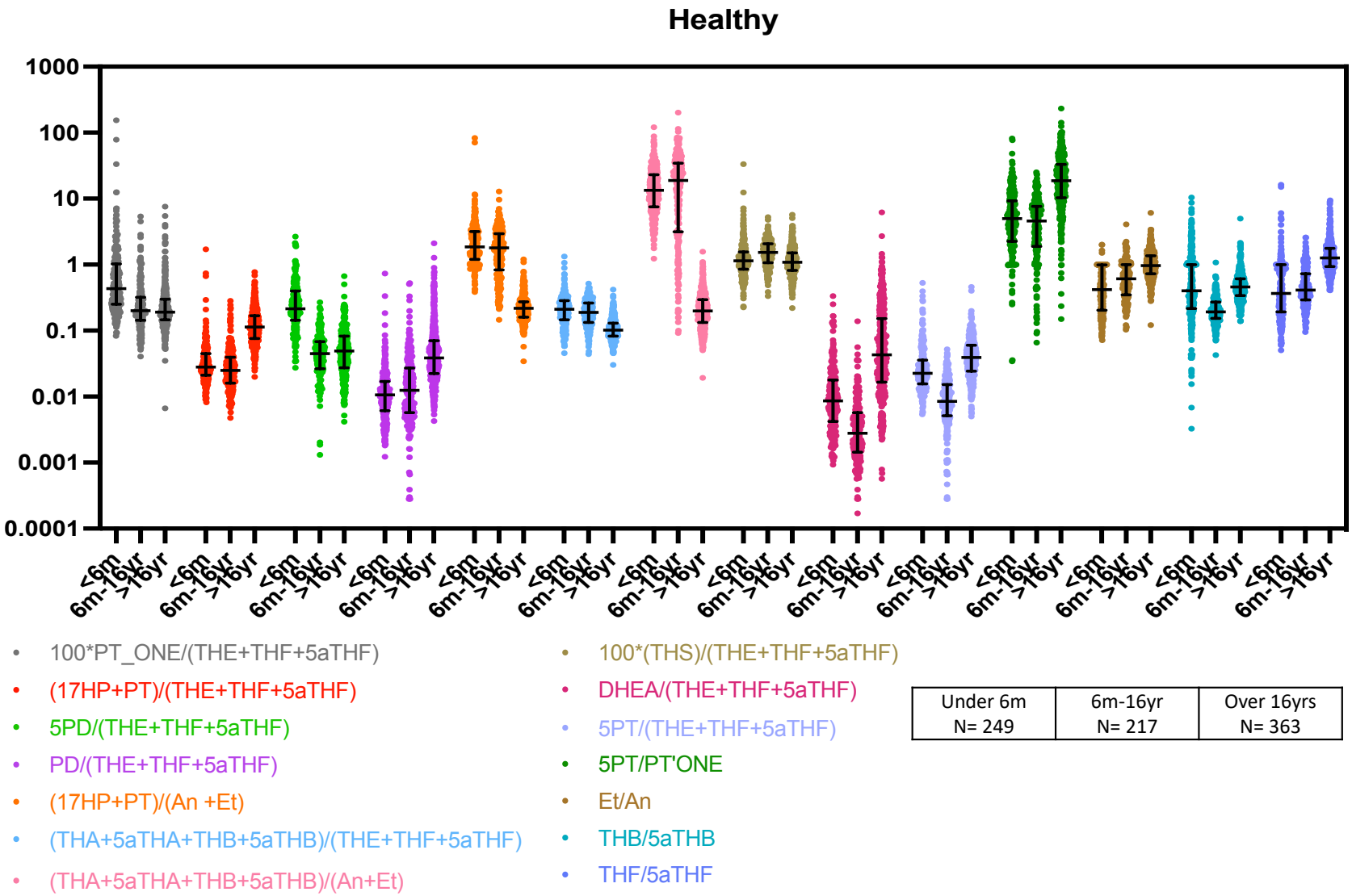
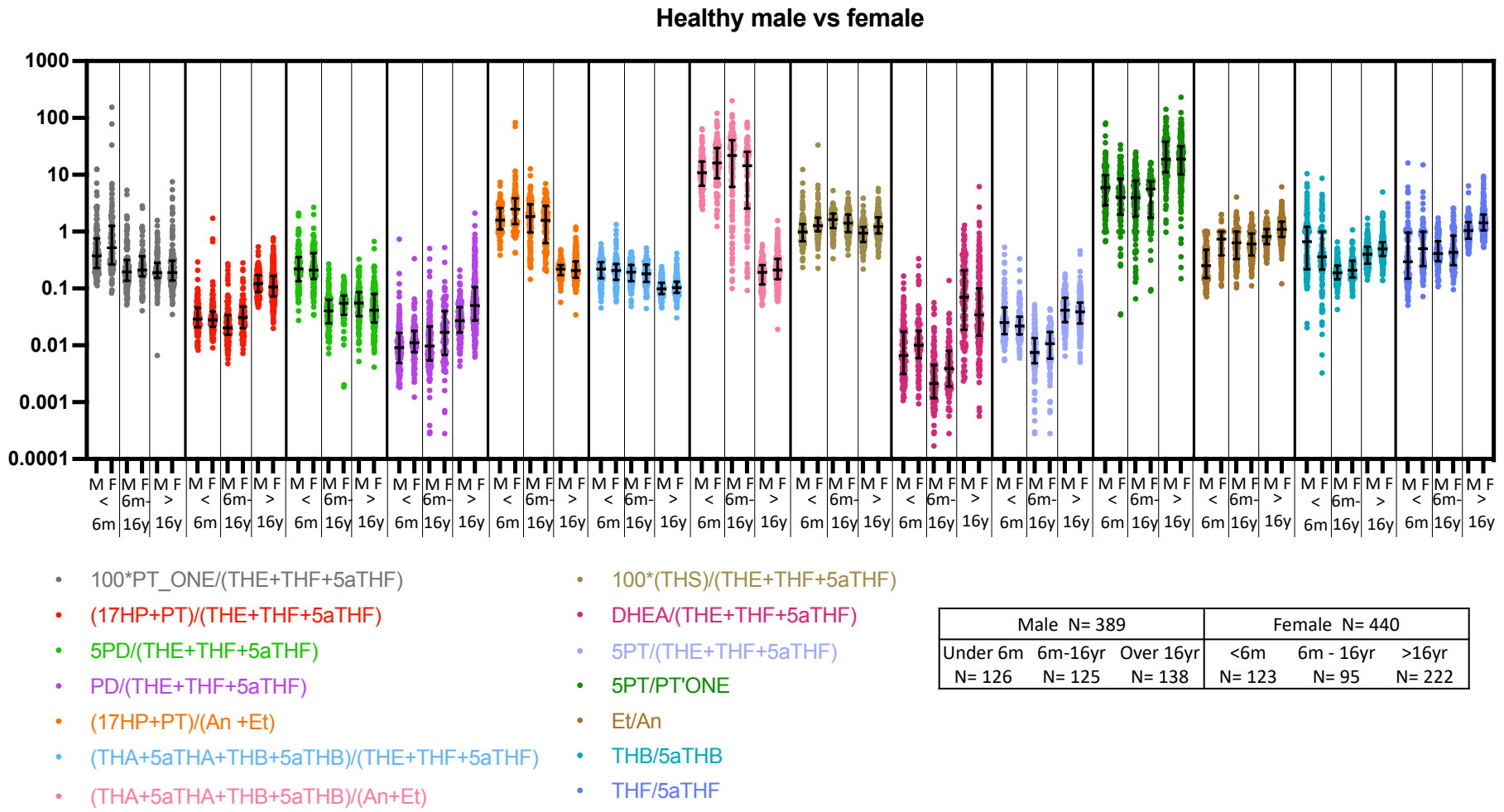


Figure 3.9: Healthy control data for the 14 biochemical ratios split by age

Figure 3.9: Healthy control data for the 14 biochemical ratios split by age

Healthy control data is split by 3 defined age groups with male and female subject data combined. Values for all subjects are presented as a scatter plot with median and interquartile range highlighted for each group of data. All subject data is displayed on a logarithmic scale. Colour corresponds to the ratio presented.



Male N= 389			Female N= 440		
Under 6m	6m-16yr	Over 16yr	<6m	6m - 16yr	>16yr
N= 126	N= 125	N= 138	N= 123	N= 95	N= 222

Figure 3.10: Healthy control data, split by sex and age, for 14 biochemical ratios

Figure 3.10: Healthy control data, split by sex and age, for 14 biochemical ratios

Values for all subjects are presented as a scatter plot with median and interquartile range highlighted for each group of data. All subject data is displayed on a logarithmic scale. Colour corresponds to the ratio presented. For each ratio, Male then female data is shown for each defined age group- less than 6 months, 6 months to 16 years and then over 16 years.

There were more statistically significant group differences comparing age-specific groups than sex- and age-specific groups. Using Kruskal-Wallis test for comparing the 3 age groups, statistically significant differences were seen for all 14 biochemical ratios with sex data combined ($p<0.0001$). A total of 42 age-related pairwise comparisons were made using Mann-Whitney *U* test*:* 3 pairwise age combinations for each of the 3 age groups for each of the 14 biochemical ratios. 34 of the 42 groups showed statistically significant differences. This is shown in **Table 3.10**. To compare male and female data within each of the defined age groups, Mann-Whitney *U* test was employed. This showed statistically significant differences between groups for 24 of the 42 groups compared. Two biochemical ratios showed no sex-specific differences for any of the 3 age groups – the ratio of mineralocorticoid metabolites over glucocorticoid metabolites used for the diagnosis of 17α-hydroxylase deficiency (THA+5αTHA+THB+5αTHB/THE+THF+5αTHF) and the ratio of 17-hydroxypregnenolone metabolite over glucocorticoid metabolites (5PT/THE+THF+5αTHF), used for the diagnosis of 3β-hydroxysteroid dehydrogenase type 2 deficiency. 3 biochemical ratios showed statistically significant sex-specific differences for all 3 age groups examined: Progesterone over glucocorticoid metabolites (PD/THE+THF+5αTHF), mineralocorticoid over androgen metabolites (THA+5αTHA+THB+5αTHB/An+Et), and DHEA over glucocorticoid metabolites (DHEA/THE+THF+5αTHF). These 3 ratios are used for the diagnosis of P450 oxidoreductase deficiency, 17α-hydroxylase deficiency and 3β-hydroxysteroid dehydrogenase type 2 deficiency, respectively. This data is shown in **Table 3.11.**

The calculated median values are shown in **Table 3.12.**

Table 3.10: Comparative analysis of age-specific groups, sex combined

*For each ratio, differences in the distribution observed between pair wise comparisons (group 1 and group 2) were made using Mann-Whitney U test. A p-value of less than 0.05 was indicative of statistical significance (indicated by *)*

Enzyme/Ratio	Group 1	Median	IQR	Group 2	Median	IQR	P-value
21-hydroxylase							
100*PTONE/ (THE+THF+5αTHF)	<6m	0.43	0.75	6m-16yr	0.20	0.18	<0.0001*
	6m-16yr	0.20	0.18	>16yr	0.19	0.15	0.5887
	<6m	0.43	0.75	>16yr	0.19	0.15	<0.0001*
(17HP+PT)/ (THE+THF+5αTHF)	<6m	0.03	0.02	6m-16yr	0.03	0.02	0.0045*
	6m-16yr	0.03	0.02	>16yr	0.11	0.09	<0.0001*
	<6m	0.03	0.02	>16yr	0.11	0.09	<0.0001*
P450 oxidoreductase							
PD/(THE+THF+5αTHF)	<6m	0.01	0.01	6m-16yr	0.01	0.02	0.0381*
	6m-16yr	0.01	0.02	>16yr	0.04	0.05	<0.0001*
	<6m	0.01	0.01	>16yr	0.04	0.05	<0.0001*
5PD/(THE+THF+5αTHF)	<6m	0.21	0.26	6m-16yr	0.04	0.04	<0.0001*
	6m-16yr	0.04	0.04	>16yr	0.05	0.06	0.3412
	<6m	0.21	0.26	>16yr	0.05	0.06	<0.0001*
(17HP+PT)/(An+Et)	<6m	1.85	2.00	6m-16yr	1.80	2.07	0.0084*
	6m-16yr	1.80	2.07	>16yr	0.22	0.11	<0.0001*
	<6m	1.85	2.00	>16yr	0.22	0.11	<0.0001*
17α-hydroxylase							
(THA+5αTHA+THB+5αTHB)/ (THE+THF+5αTHF)	<6m	0.21	0.14	6m-16yr	0.19	0.13	0.0544
	6m-16yr	0.19	0.13	>16yr	0.10	0.05	<0.0001*
	<6m	0.21	0.14	>16yr	0.10	0.05	<0.0001*
(THA+5αTHA+THB+5αTHB)/ (An+Et)	<6m	13.4	15.5	6m-16yr	18.8	31.8	0.2103
	6m-16yr	18.8	31.8	>16yr	0.20	0.17	<0.0001*

	<6m	13.4	15.5	>16yr	0.20	0.17	<0.0001*
11β-hydroxylase							
100*THS/ (THE+THF+5αTHF)	<6m	1.14	0.75	6m-16yr	1.53	1.00	<0.0001*
	6m-16yr	1.53	1.00	>16yr	1.09	0.68	<0.0001*
	<6m	1.14	0.75	>16yr	1.09	0.68	0.4163
3β-hydroxysteroid dehydrogenase type 2							
DHEA/(THE+THF+5αTHF)	<6m	0.01	0.01	6m-16yr	0.003	0.00	<0.0001*
	6m-16yr	0.003	0.00	>16yr	0.04	0.13	<0.0001*
	<6m	0.01	0.01	>16yr	0.04	0.13	<0.0001*
5PT/(THE+THF+5αTHF)	<6m	0.02	0.02	6m-16yr	0.01	0.01	<0.0001*
	6m-16yr	0.01	0.01	>16yr	0.04	0.04	<0.0001*
	<6m	0.02	0.02	>16yr	0.04	0.04	<0.0001*
5PT/PTONE	<6m	5.00	7.10	6m-16yr	4.59	5.80	0.0784
	6m-16yr	4.59	5.80	>16yr	18.7	23.0	<0.0001*
	<6m	5.00	7.10	>16yr	18.7	23.0	<0.0001*
5α-reductase type 2							
Et/An	<6m	0.42	0.19	6m-16yr	0.61	0.65	<0.0001*
	6m-16yr	0.61	0.65	>16yr	0.97	0.68	<0.0001*
	<6m	0.42	0.19	>16yr	0.97	0.68	<0.0001*
THB/5αTHB	<6m	0.40	0.78	6m-16yr	0.19	0.12	<0.0001*
	6m-16yr	0.19	0.12	>16yr	0.46	0.27	<0.0001*
	<6m	0.40	0.78	>16yr	0.46	0.27	0.8261
THF/5αTHF	<6m	0.37	0.81	6m-16yr	0.41	0.44	0.1985
	6m-16yr	0.41	0.44	>16yr	1.26	0.87	<0.0001*
	<6m	0.37	0.81	>16yr	1.26	0.87	<0.0001*

Table 3.11: Comparative analysis of sex-specific differences for each age group

For each ratio, differences in the distribution observed between male and female data for each of the 3 age groups were analysed using Mann-Whitney U test. A p-value of less than 0.05 was indicative of statistical significance (*)

Enzyme / Ratio	Group Comparison	Male		Female		P-value
		Median	IQR	Median	IQR	
21-hydroxylase						
100*PTONE/(THE+THF+5αTHF)	<6m	0.38	0.53	0.52	0.99	0.0499*
	6m-16yr	0.20	0.18	0.21	0.20	0.2534
	>16yr	0.19	0.13	0.19	0.17	0.8566
(17HP+PT)/(THE+THF+5αTHF)	<6m	0.03	0.03	0.03	0.02	0.8129
	6m-16yr	0.02	0.02	0.03	0.03	0.0007*
	>16yr	0.12	0.08	0.11	0.09	0.0429*
P450 oxidoreductase						
PD/(THE+THF+5αTHF)	<6m	0.01	0.01	0.01	0.01	0.0373*
	6m-16yr	0.01	0.03	0.02	0.02	0.0108*
	>16yr	0.03	0.08	0.05	0.03	<0.0001*
5PD/(THE+THF+5αTHF)	<6m	0.22	0.22	0.21	0.27	0.8562
	6m-16yr	0.04	0.04	0.05	0.04	0.0468*
	>16yr	0.06	0.05	0.04	0.05	0.0161*
(17HP+PT)/(An+Et)	<6m	1.59	1.52	2.49	2.49	0.0001*
	6m-16yr	1.84	2.16	1.59	2.21	0.2451
	>16yr	0.22	0.09	0.21	0.14	0.8948
17α–hydroxylase						
(THA+5αTHA+THB+5αTHB) /(THE+THF+5αTHF)	<6m	0.22	0.14	0.21	0.13	0.7993
	6m-16yr	0.19	0.13	0.18	0.13	0.755
	>16yr	0.10	0.05	0.10	0.05	0.421
(THA+5αTHA+THB+5αTHB) /(An+Et)	<6m	10.9	10.7	16.1	20.8	<0.0001*
	6m-16yr	21.9	34.3	14.4	22.9	0.007*
	>16yr	0.19	0.14	0.21	0.19	0.006*
11β-hydroxylase						
100*THS/(THE+THF+5αTHF)	<6m	0.98	0.68	1.27	0.74	<0.0001*
	6m-16yr	1.62	0.92	1.41	1.01	0.1634
	>16yr	0.94	0.54	1.23	0.85	<0.0001*
3β-hydroxysteroid dehydrogenase type 2						
DHEA/(THE+THF+5αTHF)	<6m	0.01	0.01	0.01	0.01	0.0142*
	6m-16yr	0.002	0.00	0.004	0.01	0.0006*
	>16yr	0.07	0.19	0.03	0.09	0.0058*
5PT/(THE+THF+5αTHF)	<6m	0.03	0.03	0.02	0.02	0.1382
	6m-16yr	0.01	0.01	0.01	0.01	0.0581
	>16yr	0.04	0.04	0.04	0.03	0.189
5PT/PTONE	<6m	5.96	21.6	4.00	6.53	0.0173*
	6m-16yr	3.96	2.42	5.61	5.90	0.5224
	>16yr	18.8	4.47	19.0	21.2	0.3786
5α-reductase type 2						
Et/An	<6m	0.25	0.33	0.74	0.62	<0.0001*
	6m-16yr	0.63	0.67	0.61	0.55	0.8397
	>16yr	0.82	0.50	1.10	0.68	<0.0001*
THB/5αTHB	<6m	0.66	1.00	0.36	0.78	0.0215*
	6m-16yr	0.19	0.11	0.21	0.15	0.0584
	>16yr	0.40	0.27	0.50	0.28	<0.0001*
THF/5αTHF	<6m	0.30	0.81	0.50	0.75	0.001*
	6m-16yr	0.41	0.37	0.43	0.60	0.5699
	>16yr	1.05	0.71	1.43	0.96	<0.0001*

3.5.5. Enzyme Activity in Healthy Controls in Comparison to Patients

The calculated medians for the healthy control cohort split by age and split by age and sex are shown in **Table 3.12**. For comparison, the median for each relevant patient group, not split by age or sex, for each ratio is also shown in this table. From this data we can appreciate that while our analysis revealed statistically significant changes between age- and sex-specific groups, they were insignificant in comparison to the difference between healthy controls and patients. For example, for ratio 1, the median for our patients with 21OHD was 860.4, compared to the highest median observed in a healthy group which was 0.52 in the female under 6 months group. The magnitude of the observed difference varied between conditions and ratios, but all patient group calculated medians were at least 4 times greater than the highest median observed in a healthy group. **Figure 3.11** graphically displays the difference in observed ratio values between healthy controls and relevant patient groups, showing that most patient values were far higher than the values observed in the healthy control cohort regardless of age or sex subdivision.

Table 3.12: Calculated medians for each subdivision of the healthy cohort compared to patient groups

Blue panel = median values calculated from the healthy control cohort split into 3 age groups. Green, yellow and orange panels = median values from the healthy cohort split by age and sex. Grey panel= median values calculated from the relevant patient group/s. Ratios 1 and 2 are relevant to two patient groups, 21OHD and PORD, so an additional column is shown.

Condition	No	Ratio	<6m	6m-16yr	>16yr	<6m M	<6m F	6m-16yr M	6m-16yr F	>16yr M	>16yr F	Patients	POR
21OHD + PORD	1	100*PTONE/ (THE+THF+5αTHF)	0.43	0.20	0.19	0.38	0.52	0.20	0.21	0.19	0.19	860.41	15.31
	2	(17HP+PT)/ (THE+THF+5αTHF)	0.03	0.03	0.11	0.03	0.03	0.02	0.03	0.12	0.11	31.63	0.60
PORD	3	PD/ (THE+THF+5αTHF)	0.01	0.01	0.04	0.01	0.01	0.01	0.02	0.03	0.05	0.39	
	4	5PD/ (THE+THF+5αTHF)	0.21	0.04	0.05	0.22	0.21	0.04	0.05	0.06	0.04	1.38	
	5	(17HP+PT)/ (An +Et)	1.85	1.80	0.22	1.59	2.49	1.84	1.59	0.22	0.21	12.49	
17αOHD	6	(THA+5αTHA+THB+5αTHB)/ (THE+THF+5αTHF)	0.21	0.19	0.10	0.22	0.21	0.19	0.18	0.10	0.10	72.12	
	7	(THA+5αTHA+THB+5αTHB)/ (An+Et)	13.36	18.82	0.20	10.89	16.13	21.87	14.37	0.19	0.21	532.45	
11βOHD	8	100*(THS)/ (THE+THF+5αTHF)	1.14	1.53	1.09	0.98	1.27	1.62	1.41	0.94	1.23	303.13	
3βHSD2D	9	DHEA/ (THE+THF+5αTHF)	0.01	0.00	0.04	0.01	0.01	0.00	0.00	0.07	0.03	0.62	
	10	5PT/ (THE+THF+5αTHF)	0.02	0.01	0.04	0.03	0.02	0.01	0.01	0.04	0.04	5.41	
	11	5PT/ PTONE	5.00	4.59	18.73	5.96	4.00	3.96	5.61	18.75	18.98	183.47	
5αRD	12	Et/ An	0.42	0.61	0.97	0.25	0.74	0.63	0.61	0.82	1.10	4.37	
	13	THB/ 5αTHB	0.40	0.19	0.46	0.66	0.36	0.19	0.21	0.40	0.50	5.64	
	14	THF/ 5αTHF	0.37	0.41	1.26	0.30	0.50	0.41	0.43	1.05	1.43	46.50	

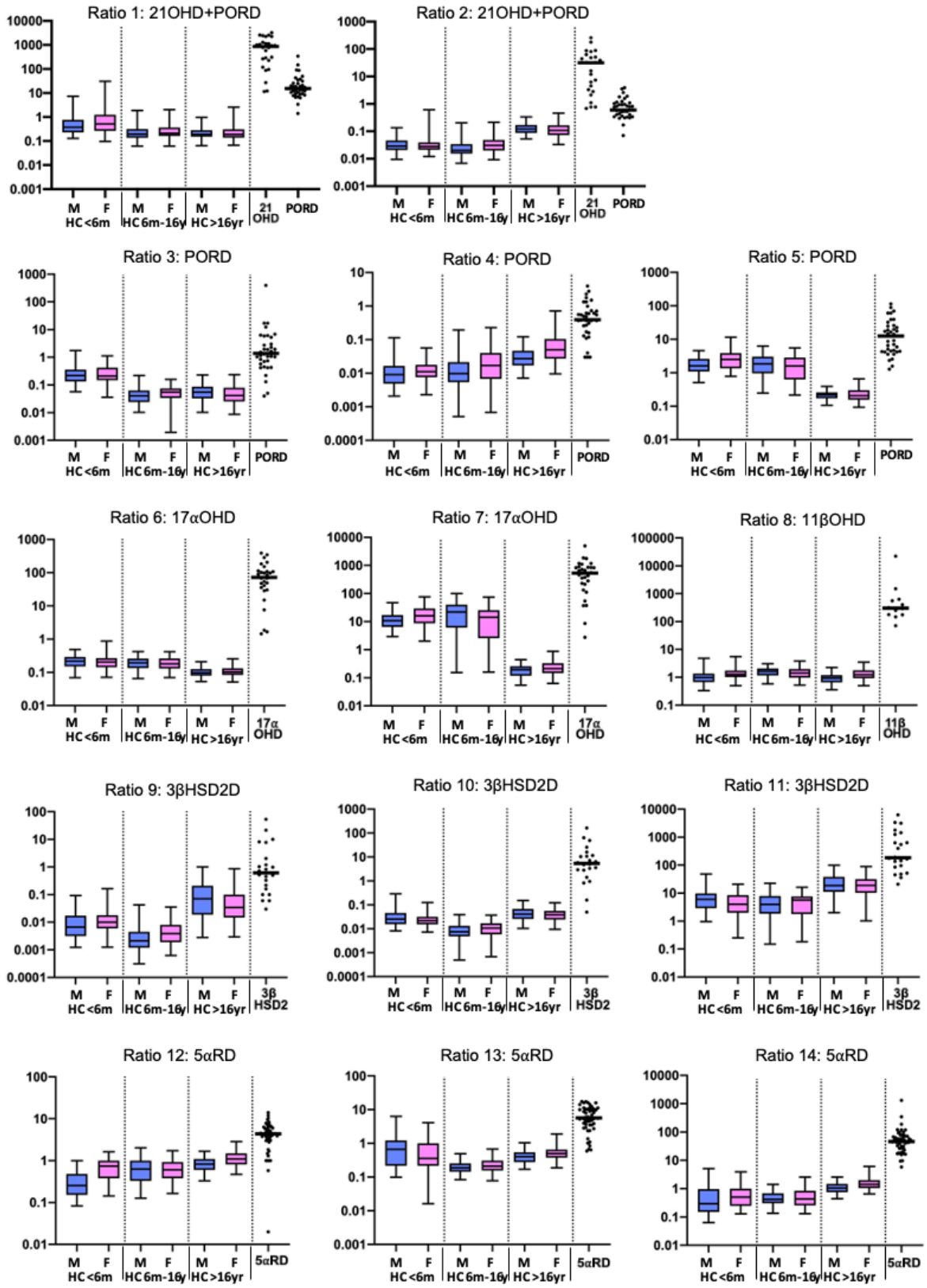


Figure 3.11: Ratio Performance by age, sex, and condition

Figure 3.11: *Healthy control cohort is split by sex (blue = male, pink = female) and into 3 defined age groups. Data is displayed by a box and whisker plot where the whiskers depict the 2.5th and 97.5th percentiles. Patient data is displayed as a scatter plot with median displayed by a horizontal line. Results are shown on a logarithmic scale. Each panel depicts performance for a single ratio. The patient data shown for each ratio is only from patients with the disease intended to be diagnosed by a given ratio.*

3.6. Discussion

This chapter shows an in-depth descriptive analysis of the urinary steroid metabolome from the cohort of patients we have collected. Through this analysis, some important characteristics have been identified relevant to our aim for designing a simplified tool to aid interpretation of urinary steroid metabolome data, for normative threshold setting and the clinical utility of these 14 metabolite ratios.

None of the ratios examined here showed a normal distribution. The distribution for most ratios tended to have a significant positive skew with a high degree of kurtosis and substantial outliers. Some ratios had a log-normal distribution, but this was not necessarily consistent for the same ratio across each of the 3 age groups. As the data did not permit normalisation, we proceeded with non-parametric statistical tests. This is relevant for normative value threshold setting, as mean and standard deviation would be inappropriate markers. The median and a percentile of normal would be better suited to the characteristics of the data. Previously published studies, for the majority, use either the 95^{th} or 97.5^{th} percentile as an upper limit of normal (ULN), but

a multiple of the median approach has been used for individual metabolite quantification reference values also (Zalas et al., 2018). Given the properties of the data with the high degree of skew and significant outliers, coupled with the exaggerated nature of changes observed in the patient populations, selecting a higher threshold of 97.5th percentile would be appropriate.

The early neonatal period is a challenging time to diagnose disorders of steroidogenesis (Honour & Rumsby, 1993; Imperato-McGinley, Gautier, Pichardo, et al., 1986; Yong et al., 1988) and the urinary steroid metabolome is known to be complex (Shackleton et al., 1979). The delayed appearance of THB, THF and their 5α reduced counterparts makes the diagnosis of 5α-reductase deficiency challenging from urinary steroid metabolome analysis, particularly as the 5α–reduced counterparts seem to appear later. This is more profound for THB and 5αTHB, where we can see the effect of this as the 0–1-month age group shows the highest values. In agreement with Shackleton et al (Shackleton et al., 1979), we show an abundance of 3β-hydroxy-5-ene steroids in the early neonatal urinary steroid metabolome.

When compared to published absolute 24-hour urine metabolite concentrations, we observed far less profound sex related variability with ratio values (Ackermann et al., 2019; Chan et al., 2008). Age made a statistically more significant difference than sex to the ratio values observed in the healthy population, an observation also made by Remer T et al (Remer et al., 2005). Lucas-Herald et al (Lucas-Herald et al., 2015) found no significant age or sex difference between urinary steroid metabolite ratios. In comparison to patient values, any observed healthy population sex related variability in our data was much less significant, with patient values being far in excess for the

majority of ratios. For some ratios, there was complete separation for patient values from healthy controls but for others it seems age is an important factor to keep in consideration. Looking at the changes in ratio values across the 10 age groupings, our 3 proposed age groupings seem to encompass the major developmental stages. Infancy appears to be distinct with a point of transition manifesting between 3 and 12 months of age. Adrenarche related changes are observable between the 1-4 years and the 4-10 years age groups, with the 4-10 years age group being a pivot point between childhood and adulthood observed values. Our results are in agreement with the observations by Kim et al (Kim et al., 2020) who comprehensively characterised changes in the steroid metabolome from 2 to 6 years of age. Lucas-Herald et al (Lucas-Herald et al., 2015) proposed three childhood age groupings of less than 6 months, 6 months to 10 years, and 10 to 18 years. As this is a dynamic physiological period, more detailed age groupings may be warranted. For this study we did not have sufficient participant numbers in this critical age range for further sub-analysis. For some ratios, the phenomenon of adrenopause appears to be demonstrable with a decline observed for certain ratios in the over 50 years age group – suggesting a decrease in production of androgen precursors and/or increased production of glucocorticoid metabolites. There is an argument for including a 4th over 50 years age group to comprehensively represent all life stages, however, it would be exceptionally rare for a patient to present in this age group, so this seems less relevant for a diagnostic test for this study. When developing a test for treatment monitoring for disorders of steroidogenesis, this would be an important consideration to include.

21-hydroxylase activity demonstrated through this study mirrors the findings of Remer et al (Remer et al., 2005) who looked in detail at urinary markers of adrenarche from

2 to 18 years, that there is a fall in activity with increasing age from 4 years to 18 years. HSD3B2 activity demonstrated here also mirrors previously published data (Gell et al., 1996; Remer et al., 2005; Rich et al., 1981), that activity declines with age, with a stepwise reduction around the time of adrenarche/puberty. Similar to our data, Remer et al found variable significance regards sex-specific differences in enzyme activity. Ratios that include androgen metabolites as the denominator are understandably affected by gonadal function and the post pubertal rise in androgens leads to low ratio values. Ratios containing DHEA, 17-OHPreg, progesterone and 17OHP metabolites as the numerator demonstrate substantially higher values post puberty compared to pre-puberty. Looking at the ratio of 17OHP over androgen metabolites, however, shows that the post pubertal rise in androgens is of greater magnitude than the post pubertal rise in 17OHP as this ratio demonstrates considerably lower values in adulthood. Interestingly, this ratio did not demonstrate any statistically significant sex related difference when comparing adult male and female groups. Mineralocorticoid and glucocorticoid concentrations appear stable throughout life as demonstrated by a ratio of their metabolites used for discriminating 17αOHD, and also by the ratio of 11-deoxycortisol over glucocorticoid metabolites used for the diagnosis of 11βOHD.

Rousson V et al (Rousson et al., 2021) published a comprehensive analysis of previously published steroid metabolite ratios felt to reflect steroidogenic activity. They looked at 65 ratios. Most ratios they examined showed a similar positive or right skew distribution. They achieved “near normal” distribution through transformation and demonstrated both age- and sex-specific differences for almost all ratios. Similar to our findings, they also observed differing patterns for different ratios used for approximating a single enzyme’s activity. For HSD3B2, they also observed an

increase in enzyme activity in the older adult age groups compared to younger adult groups. They also found a slight decrease in 5α-reductase activity with increasing age.

3.7. Conclusion

Steroidogenesis is a complex pathway and isn't unidirectional, therefore while these ratios may not all perfectly equate to enzymatic activity, they do accurately represent normal physiological processes. While statistical significance was achieved for age- and sex-specific groups, age-specific differences appear more significant than sex-specific differences, and neither was as profound as the difference observed between healthy and patient values. A defective enzyme leading to a block in this process rarely results in a subtle alteration in the steroid metabolome, as we have demonstrated here, therefore we should expect an acceptable sensitivity for any clinical diagnostic tool utilising these ratios. A careful algorithm for interpretation should be developed to assist the user and improve performance.

4. Use of Biochemical Ratios for the Differential Diagnosis of Inborn Steroidogenic Disorders

4.1. Introduction

Multiple studies describe the characteristic pattern of altered steroidogenesis and subsequent metabolites observed for each of the inborn steroidogenic disorders (Bongiovanni, 1962; Bongiovanni & Eberlein, 1956; Neres et al., 2010; Peterson et al., 1985b; Shackleton et al., 1986). In 1980, Shackleton (Shackleton et al., 1980) constructed an atlas of gas chromatographic profiles in health and disease and comprehensively outlined the urinary steroid metabolome changes in a variety of inborn disorders of steroidogenesis. In this manuscript, the authors proposed measurement of "appropriate relative steroid excretions" in the form of ratios of precursor metabolite/s to product metabolite/s. For a given inherited mutation in a gene encoding a particular enzyme in the steroidogenesis pathway, a non-functioning or reduced functioning enzyme is produced. This leads to an "enzymatic block" in the steroidogenic pathway where there is a reduction in downstream products and a build-up of precursor steroids (Baranowski et al., 2018). This is reflected in the urinary steroid metabolome, therefore, ratios of urinary metabolites of precursor to product steroids can be used as surrogate markers of enzyme activity and support interpretation of urinary steroid metabolome data (Caulfield et al., 2002). The additional benefit of using ratios of metabolites is that you can apply this approach to single spot urine samples as well as 24-hour collections, making this a much more practical approach for use with young children.

There have been limited numbers of studies which critically assess the diagnostic performance of these ratios. Previous studies are mostly limited by small numbers or looking at conditions in isolation compared to healthy controls. **Table 4.1** shows

examples of previous original work describing the utility of biochemical ratios to support urinary steroid metabolome interpretation.

Lucas-Herald (Lucas-Herald et al., 2015) noted that for patients presenting with clinical characteristics to warrant urinary steroid metabolome examination, but with ultimately no diagnosis confirmed, they may be found to have abnormally raised diagnostic ratios when using the 95th percentile of healthy as an upper limit of normal. They further noted that for the patients who were confirmed to have an inborn disorder of steroidogenesis, alongside the relevant diagnostic steroid metabolite ratios, ratios for other disorders were also found to be raised.

Caulfield (Caulfield et al., 2002) illustrated the diagnostic difficulty with discriminating between 21OHD and 3βHSD2D as both groups of patients display abnormal ratios intended for the diagnosis of the other. They proposed instruction for how to discriminate between the two conditions using the ratio of 5PT over PTONE. Aside from this comparison, and comparing to healthy controls, no further specificity analysis was undertaken including the other patient samples they presented. Koyama Y (Koyama et al., 2012) proposed two separate diagnostic thresholds to differentiate 21OHD from PORD and from healthy controls in Japanese infants based on urinary steroid metabolome analysis. They also proposed additional measurement of 11β-hydroxyandrosterone would aid further with correct disease diagnosis, which they presented as an absolute metabolite quantification rather than inclusion within a ratio.

This body of work highlights the need for larger studies examining the utility of these ratios to support interpretation of urinary steroid metabolome analysis. While these

original studies provide strong evidence of the utility of steroid metabolite ratios for diagnosis of the disease of interest, they don't necessarily provide a realistic estimate of performance based on how the ratios are likely to be used and in the target population the test is intended for. Careful consideration is needed for construction of reference intervals and clear instruction for how these ratios should be interpreted to arrive at a single diagnosis.

4.2. Objective

The objective we present in this chapter was to create a robust and straightforward algorithm for urinary steroid metabolome interpretation, built using the previously described 14 biochemical ratios of precursor to product steroid metabolites. We assessed the performance of this diagnostic tool using a database of comprehensive urine steroid metabolome data, quantified by GC-MS, from healthy controls and patients with definitively determined diagnoses.

Table 4.1: Published evidence for the utility of urinary steroid metabolite ratios or their components

Reference	Participants included
Caulfield MP et al. JCEM 2002.	21OHD n=31, 11βOHD n=2, 17αOHD n=1, 3βHSD2D n=1, Healthy n=59
Lucas- Herald AK, et al. Int J Pediatr Endocrinol. 2015	21OHD n=4, 11βOHD n=2, 5αRD n=2, Symptomatic cases with no diagnosis n=85, Healthy n=252
Bileck A, et al. JSBMB 2020	21OHD n=4, 11βOHD n=1, 17αOHD n=2, 5αRD n=3, Healthy n=40
Bongiovanni AM, et al. J Clinical Investigation 1962	21OHD n=5, 3βHSD2D n=6, Healthy n=5
Idkowiak J, et al. JCEM 2012.	PORD n=21, 17αOHD n=20, 17,20lyase n=6, Cytb5 n=3, Healthy n=98
Koyama Yuhai et al. Clinical Chemistry 2012	21OHD n=29, PORD n=9, Healthy babies with transiently raised 17OHP n=67, Heathy n=1341.
Malunowicz EM et al. Horm Res 1997.	161 with clinical suspicion of CAH: 21OHD n=61, 3βHSD2 n=1, Healthy n=99. Healthy preterm babies (no suspicion of CAH), n=20
Homma K, et al. JCEM 2004.	21OHD n=59, Healthy babies with transiently raised 17OHP n=83, Healthy n=62
Kamrath C et al. JSBMB 2016	21OHD n=95. Healthy n=261
Mussig K et al. Exp Clin Endocrinol Diabetes 2006	21OHD n=1
Vierhapper H et al. J Steroid Biochem. 1985	21OHD n=23
Shackleton C et al. Steroids. 1986	21OHD n=6, healthy n=17
Wudy S et al. Journal of Endocrinology. 2000	21OHD n = 11
Nguyen HH et al. JSBMB 2016	11βOHD n=6
Eberlein W, Bongionvanni AM. J Biol Chem. 1956	11βOHD n=1
Krone N et al, JCEM 2012.	PORD n=23, Healthy n=U/K
Homma K et al. JCEM 2006	PORD n=22, Healthy male n=854, Healthy female n=909
Idkowiak J, et al. JCEM 2011.	PORD n=7, Healthy n=12
Shackleton C et al. American Journal of Medical Genetics 2004	PORD n= 9, Healthy n=34
Dean HJ, et al. JCEM 1984	17αOHD n=1
D'Armiento M et al. JCEM 1983	17αOHD n=2
Sun M et al. Eur J Endocrinol. 2021	17αOHD n=8, Healthy n=28
Neres MS, et al. Arq Bras Endocrinol Metabol 2010	17αOHD n=20, Healthy n=11
Rosenfield RL et al. JCEM 1980	3βHSD2D n=1
Shackleton C. 2008	3βHSD2D n=8, Healthy n= 74
Imperato-McGinley et al JCEM 1986	5αRD n=3
Peterson RE et al. Clinical Endocrinology 1985.	5αRD n=48, Healthy n=26
Berra M et al. EJE 2011	5αRD n=14, 46,XY female (not 5αRD) n=13.

4.3. Methods

The study design and detailed cohort description are included in chapter 2, section 1. Please refer to **Table 2.1** for a detailed description of the cohort included in this study. All patients entered into the study were confirmed to have the stated disease, i.e., to harbour disease causing mutations. All healthy controls were screened by means of a questionnaire to ensure that they had no medical condition or regular medication that would interfere with steroidogenesis.

4.3.1. Sub-Group Definitions

To create informative reference ranges, we separated the healthy cohort by sex and into 3 age groups as described in chapter 3: Under 6 months of age, 6 months to 16 years and over 16 years of age. Due to the changing metabolome observed in early neonatal life and the presence of adult type metabolites being not consistently measurable until 6 months of age, we decided to separate infants under 6 months into a specific group. The 6 months to 16 years cohort will contain children moving through the spectrum of pre to post puberty / adrenarche. As pubertal staging information was not routinely collected for these healthy controls and given the age range considered normal for adrenarche and puberty, we kept this group together as one. The over 16 years of age group represents the post pubertal cohort.

4.4. Test Methods

4.4.1. The Index Test

4.4.1.1. Biochemical Ratios and Algorithm for Use

The ratios chosen to inform this method, along with the disease they are intended to diagnose are listed in **Table 2.3.**

Figure 4.1 shows the simple algorithm we created, designed for use even by non-experts in the field, for interpretation of urinary steroid metabolome quantification results. It states that calculated ratio values are presented as proportions of the upper limit of normal, then the proportionally highest magnitude ratio value defines the diagnosis. If the highest magnitude ratio is diagnostic of 21OHD or PORD, then we look to the second highest ratio for confirmation as to which of these is the diagnosis. If the second highest ratio is diagnostic of PORD or 17αOHD, then the diagnosis is PORD. If the second highest ratio is also diagnostic of 21OHD, or if the second highest ratio is diagnostic of an enzyme defect that is not 21OHD, PORD or 17αOHD, then the diagnosis is 21OHD. The order of magnitude for the different ratios within the healthy population differs greatly. To directly compare an individual's calculated ratio values, we scaled the values by dividing them by their respective upper limit of normal. This gives us a ratio value as a proportion of the healthy upper threshold which was then expressed as a percentage. Any value greater than 100% is above the upper limit of normal so is therefore considered abnormal.

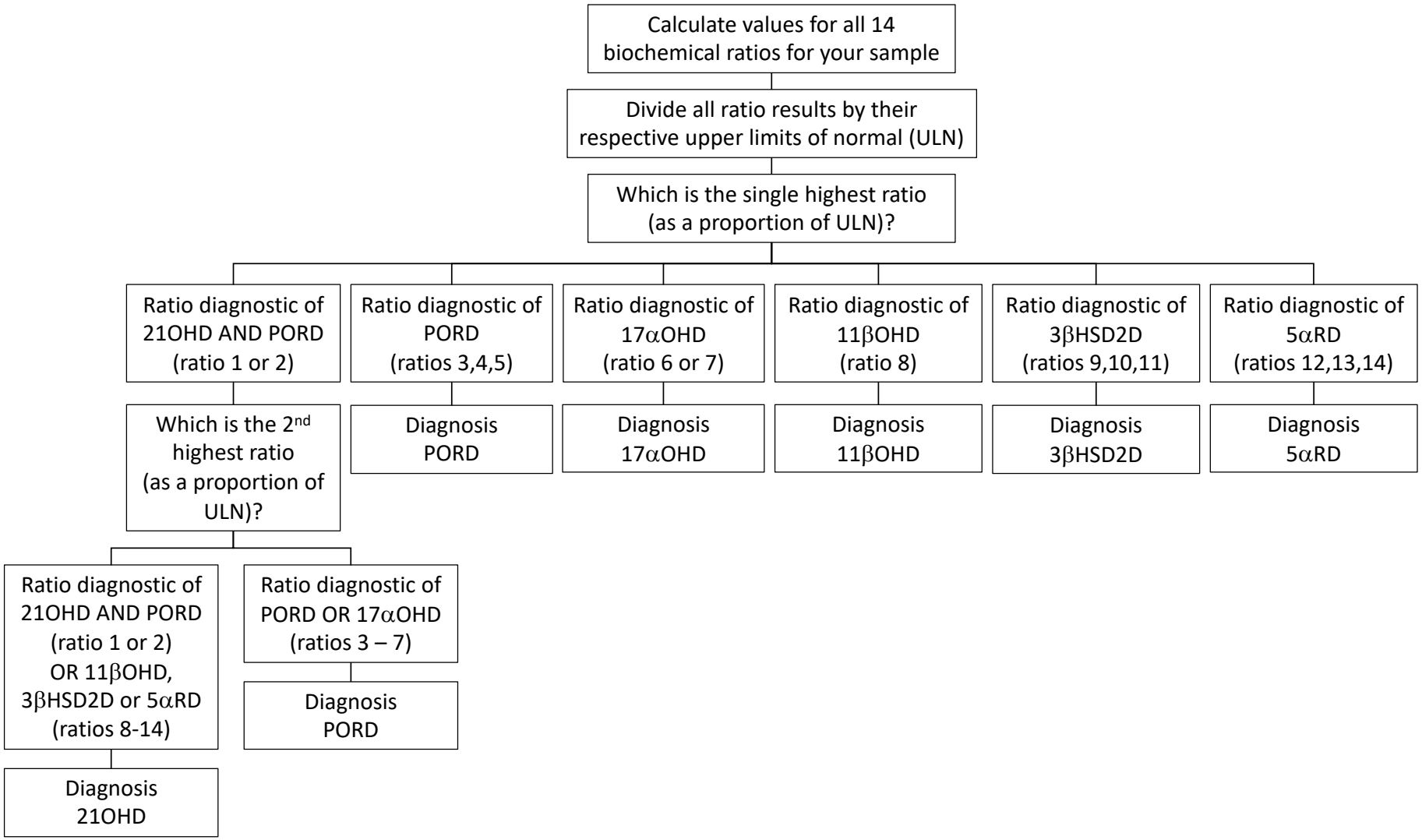


Figure 4.1: Decision tool for interpretation of urinary steroid metabolome data. *To be used in conjunction with Table 2.3.*

No.	Urinary Steroid Metabolite Ratio	Enzyme / Co-Factor Activity	Steroidogenesis Disorder Disruption is Indicative of
1	100*PTONE/(THE+THF+5αTHF)	CYP21A2 + POR	21-hydroxylase deficiency or P450 oxidoreductase deficiency
2	(17HP+PT)/(THE+THF+5αTHF)	CYP21A2 + POR	21-hydroxylase deficiency or P450 oxidoreductase deficiency
3	PD/(THE+THF+5αTHF)	CYP17A1 + POR	P450 oxidoreductase deficiency
4	5PD/(THE+THF+5αTHF)	CYP17A1 + POR	P450 oxidoreductase deficiency
5	(17HP+PT)/(An+Et)	CYP17A1 + POR	P450 oxidoreductase deficiency
6	(THA+5αTHA+THB+5αTHB)/ (THE+THF+5αTHF)	CYP17A1 + POR	17α-hydroxylase deficiency
7	(THA+5αTHA+THB+5αTHB)/ (An+Et)	CYP17A1 + POR	17α-hydroxylase deficiency
8	100*THS/(THE+THF+5αTHF)	CYP11B1	11β-hydroxylase deficiency
9	DHEA/(THE+THF+5αTHF)	HSD3B2	3β-hydroxysteroid dehydrogenase deficiency
10	5PT/(THE+THF+5αTHF)	HSD3B2	3β-hydroxysteroid dehydrogenase deficiency
11	5PT/PTONE	HSD3B2	3β-hydroxysteroid dehydrogenase deficiency
12	Et/An	SRD5A2	5α-reductase type 2 deficiency
13	THB/5αTHB	SRD5A2	5α-reductase type 2 deficiency
14	THF/5αTHF	SRD5A2	5α-reductase type 2 deficiency

Table 2.3 *(from chapter 2)*

4.4.1.2. Definition and Rationale for Test Positive Cut-Offs:

Two approaches for creating the reference intervals for our index test are presented here. We have used the whole cohort to create reference data for both methods but also show results obtained through a cross validation approach using 5 subdivisions of the cohort.

The first approach used the 97.5th percentile as the upper limit of normal, determined for each sub-group from the healthy control cohort.

The second approach used a multiple of age- and sex-specific medians for each biochemical ratio as the upper limit of normal. Sub-groups used the same definitions as already described above. In order to create thresholds using a multiple of the median as the upper limit of normal, which are generalisable to all examined age- and sex-specific subgroups even where we have no patients for comparison, we had to make one critical assumption: Although the median for the healthy group may vary, the difference between patient values when compared to the relevant healthy age- and sex-specific median remains static for all groups for a given ratio. By making this assumption we could determine a single “multiple” for each diagnostic ratio to apply to each age- and sex-specific median for optimum separation of patients from healthy controls. For a given ratio, calculated values for all patients and healthy controls were divided by their respective age- and sex-specific median. These ratio values as a proportion of the median were then used to inform a Receiver Operating Characteristic (ROC) analysis. This allowed examination of how the sensitivity and specificity were affected by varying a single multiplication factor applied to each age- and sex-specific median for a given ratio. Thresholds were initially selected using a binary disease vs healthy control comparison and refined using a heuristic approach according to where misclassifications occurred to maximise sensitivity. Threshold refinement was limited to varying between multiple values across the corner position of the ROC curve, between the point where the rate of sensitivity decline increases to the point of maximum sensitivity.

There was no blinding of participant diagnoses during the process of test parameter construction or during test performance analysis. As a whole cohort analysis approach

was adopted, the same cohort was used for test parameter construction as for performance analysis.

4.4.2. The Reference Standard

We have not compared to a reference standard for urinary steroid metabolome analysis for this study. The reference standard for biochemical diagnosis might be considered to be the result of steroid metabolome analysis and interpretation by an expert. However, urinary steroid metabolome analysis is more complex than looking at absolute metabolite concentrations as quantified in selected ion mode and not all samples were run in both selected ion and full scan modes. It was also felt that it would be an unrealistic undertaking to ask an expert in the field to analyse 1007 samples.

4.5. Analysis Methods

4.5.1. Diagnostic Accuracy Estimates and Analyses of Variability

Discriminative ability of each individual diagnostic ratio was assessed by comparison of values in the target class compared to all other classes. The index test provides discrete class labels so there are no indeterminate results to consider. Sensitivity and specificity are the primary measures of diagnostic accuracy used in this study. Sensitivity is shown in three ways. Firstly, sensitivity to assign a diagnosis of any disease (number of patients assigned a diagnosis of any patient class over total number of patients). Secondly, sensitivity to assign the diagnosis of the correct

disease (number of patients assigned to the correct disease class over total number of patients). Lastly, we also show class-wise sensitivity for each disease class (Number of correctly assigned patients over total number of patients for each individual class). For test precision, 95% confidence intervals are provided.

To examine for variability, we took a cross validation approach. The cohort was split into 80% of samples allocated for training and the remaining 20% of samples allocated as a test cohort. 5 folds of cross validation were used with the same allocation of samples into training and testing cohorts, as used for the machine learning approach (see chapter 5). The relative stability of the 97.5th percentile compared to the median within the healthy cohort was examined by calculating the mean and standard deviation of the median across the five training folds of cross-validation. To assess for potential variability in diagnostic accuracy with using the 97.5th percentile as the upper limit of normal, in addition to the whole cohort analysis, a cross validation approach was also taken. Thresholds were determined per fold from the healthy class in the training allocation and performance assessed by applying these to participants in the corresponding test allocation. Performance results from cross validation are shown as mean and standard deviation across the 5 different arrangements. For the multiple of median approach, a cross validation approach was also used to assess for potential variability in the ROC analysis only. For the 5 folds of training data, 5 ROC curves were determined and plotted on the same graph to illustrate potential variability in the analysis and Area under the curve (AUC) values are compared for all 5 curves.

4.5.2. Statistical Analysis

Statistical analyses have been conducted and figures created using GraphPad Prism (version 9.1.0), StataCorp Stata (version 16), Microsoft Excel (version 16.48) or MathWorks MATLAB (R2020a). For computing confidence intervals, the Wilson method was used. For ROC analysis, area under the curve (AUC), standard error (Std error), 95% confidence interval (CI) and p value are presented for each curve. Standard error and 95% confidence interval of the curve were computed using a nonparametric method that does not make any assumptions about the distributions of test results in the patient and control groups. The p value reported tests the null hypothesis that the AUC equals 0.5. For sensitivity and specificity derived from the ROC analysis, confidence intervals are computed using the Clopper method (Clopper & Pearson, 1934).

4.5.3. Missing Values

Figure 2.1 shows a detailed illustration of the missing metabolite values in the dataset. For most, the metabolite missing is not included in any of the biochemical ratios. Where a subject had missing metabolite values that are included in the biochemical ratios, if one of a combination of numerator or denominator values was missing then the ratio was calculated from the remaining metabolite values with no compensation or imputation. For ratios using a single numerator/denominator, if this was not measured, the ratio was discounted for that subject and decisions made from the remaining ratios.

4.6. Results

4.6.1. Discriminative Ability of Biochemical Ratios

For all 14 biochemical ratios, calculated ratio values were plotted for all healthy controls and all patient groups, shown in **Figure 4.2.** Data for the healthy cohort was split into the three defined age groups and compared to values for subjects in all patient groups, with each patient group displayed separately. Patient groups were not subdivided by age. Values for healthy controls are displayed as box and whisker plots, boxes showing median and interquartile range and whiskers depict the 2.5^{th} / 97.5^{th} percentiles. Patient results are displayed as scatter plots with the median highlighted.

For each of the ratios used, the target patient group (indicated by the grey panel) showed higher values than the three healthy control age groups, with good separation observed between controls and patients with the disease of intent. For some ratios, conditions other than the target condition also show values above that observed in the healthy control groups and, in some instances, having a similar or even higher magnitude than the disease of intent

In ratio 1 and 2, the target conditions were 21OHD and PORD. For ratio 1, these groups showed the greatest values, however, patients with 17αOHD and 3βHSD2D also showed values greater than healthy controls. For ratio 2, all patient groups except 5αRD showed values greater than healthy controls and patients with 3βHSD2D showed values greater than the PORD group. The two ratios showed excellent separation of patients with 21OHD from healthy controls.

For ratios 3, 4 and 5, the target patient group was PORD. For ratio 3, the PORD group displayed greater values than the healthy control groups, however, the 21OHD, 17αOHD and 3βHSD2D groups all showed values greater than healthy controls. Patients with 17αOHD showed values significantly greater than PORD. For ratio 4, all patient groups apart from 5αRD showed values greater than healthy controls. The 21OHD, 17αOHD and 3βHSD2D groups showed values greater than PORD. For ratio 5, the PORD group had values greater than healthy controls and all patient groups apart from 21OHD which showed equivalently high values. This ratio has been previously suggested to represent 17/20 lyase activity and felt to be diagnostic of 17αOHD and PORD. We have assigned it here solely as a ratio for the diagnosis of PORD, as patients with 17αOHD showed values equivalent to healthy controls.

For ratios 6 and 7, the target condition was 17αOHD. For ratio 6, this group showed the greatest magnitude values by far. Complete separation from healthy controls was observed and while 21OHD and PORD showed values greater than healthy controls, there was minimal overlap of the 17αOHD group with other patient groups. For ratio 7, the 17αOHD group demonstrated the highest magnitude values with some overlap seen with the youngest two healthy control groups. It must be noted here that age of the patient groups was not accounted for in this figure. When age was accounted for, there was complete separation of this patient group observed from all healthy control groups.

For ratio 8, 11βOHD was the target group and demonstrated the highest magnitude values above healthy controls and all other patient groups. There was complete separation from all healthy control groups observed.

For ratios 9, 10 and 11, 3βHSD2D was the target group. For ratio 9, 3βHSD2D and 11βOHD showed the highest magnitude values, tending to be greater than healthy controls but with some overlap observed. Other patient groups showed similar values to healthy controls. For ratio 10, the 3βHSD2D group showed the highest values of all groups examined. 21OHD and 11βOHD showed values greater than healthy controls but not to the same magnitude. For ratio 11, values observed for the 3βHSD2D group were greater than healthy controls, however, the 11βOHD patient group showed values slightly higher. Other patient groups showed values similar or lower than healthy controls. 21OHD was significantly lower than healthy controls with almost perfect separation.

For ratios 12, 13 and 14, 5αRD was the target condition. Values demonstrated by this group were highest for all three ratios above all other groups. Ratio 14 demonstrated complete separation of the 5αRD group from the healthy groups.

Figure 4.3, Figure 4.4, Figure 4.5, Figure 4.6, Figure 4.7 and **Figure 4.8** show the results when all ratios were applied to each patient group in isolation. The ratio/s that were intended to be diagnostic for each condition are highlighted by the grey panel. The top panels show a comparison of the absolute ratio quantification, and the bottom panels show the ratio values as a proportion of the age matched 97.5th percentile, using the same 3 age groupings as previously. These figures illustrate that despite multiple ratios being raised above healthy control values for each condition, the relevant ratios tended to be the highest observed within their given target group. This

further improved when results were standardised for the differing scale by using results as a proportion of a proposed upper limit of normal.

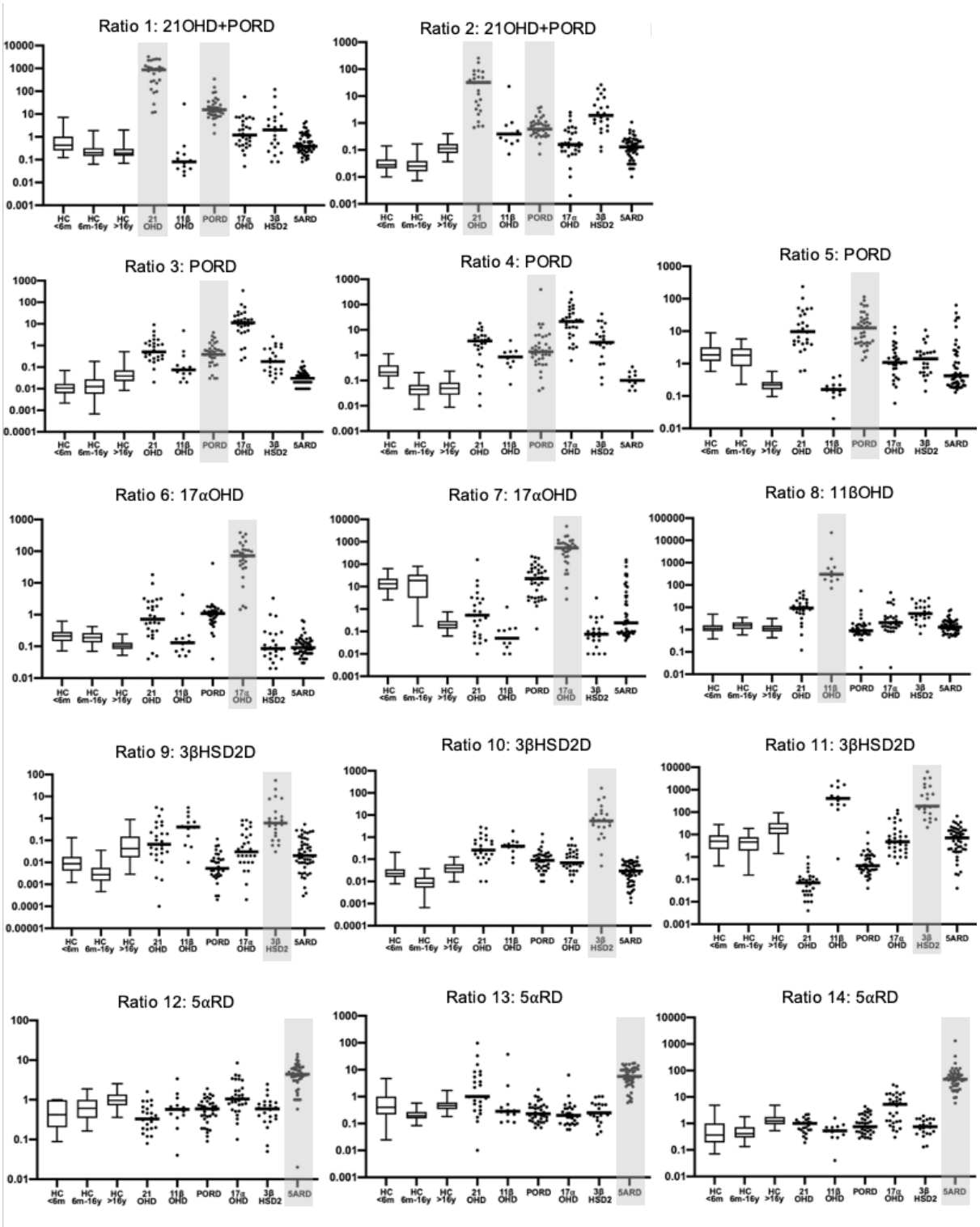


Figure 4.2: Ratio Performance by age and condition

Figure 4.2: *The healthy control cohort is split into 3 defined age groups (less than 6 months, 6 months to 16 years and over 16 years) and data is displayed by a box and whisker plot where the whiskers depict the 2.5th and 97.5th percentiles. Patient data is displayed as a scatter plot with median indicated by a horizontal line. Results are shown on a logarithmic scale. Each panel depicts a particular ratio, and for each ratio, the intended disease to be diagnosed is highlighted by a grey vertical panel.*

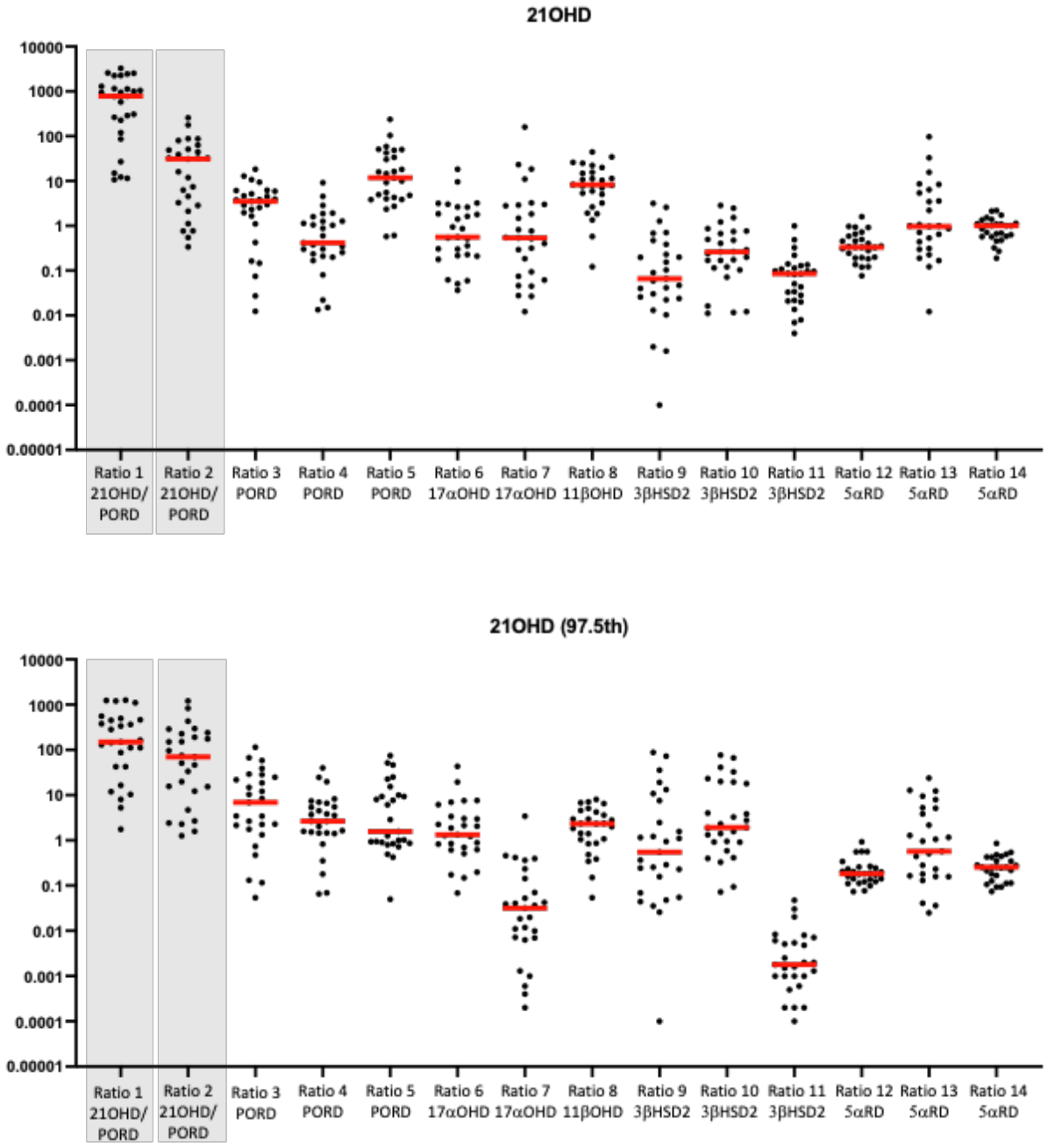


Figure 4.3: Ratio performance by condition, 21OHD

14 urinary steroid metabolite ratios calculated for patients with 21OHD, N=27. Results are shown as a scatter plot, a single dot per subject, median is displayed by a red horizontal line. Top panel: Ratio values as absolute quantification. Bottom panel: Results shown as a proportion of the 97.5th percentile of the corresponding age-matched healthy control cohort group.

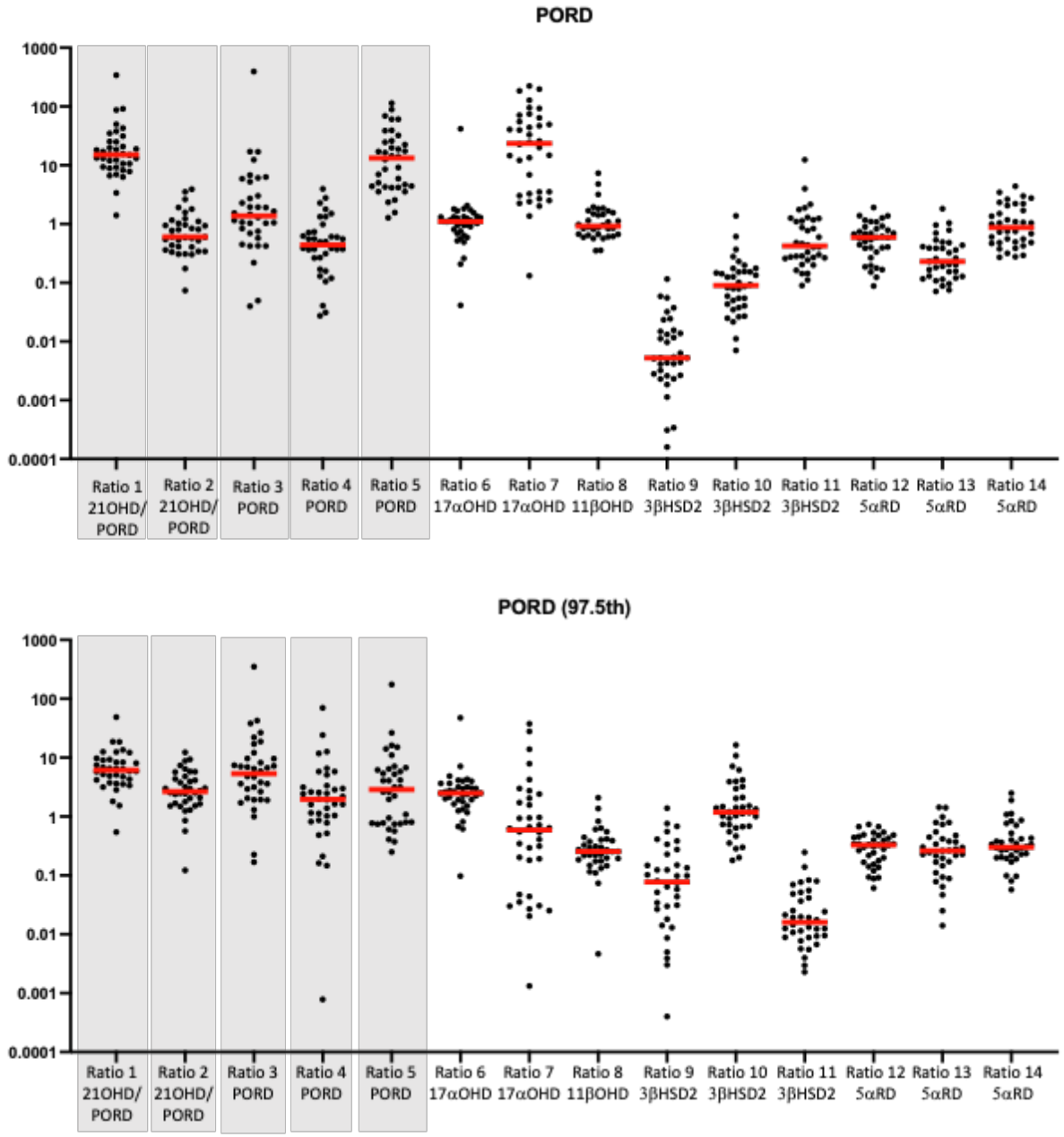


Figure 4.4: Ratio performance by condition, PORD

14 urinary steroid metabolite ratios calculated for patients with PORD, N=36. Results are shown as a scatter plot, a single dot per subject, median is displayed by a red horizontal line. Top panel: Ratio values as absolute quantification. Bottom panel: Results shown as a proportion of the 97.5th percentile of the corresponding age-matched healthy control cohort group.

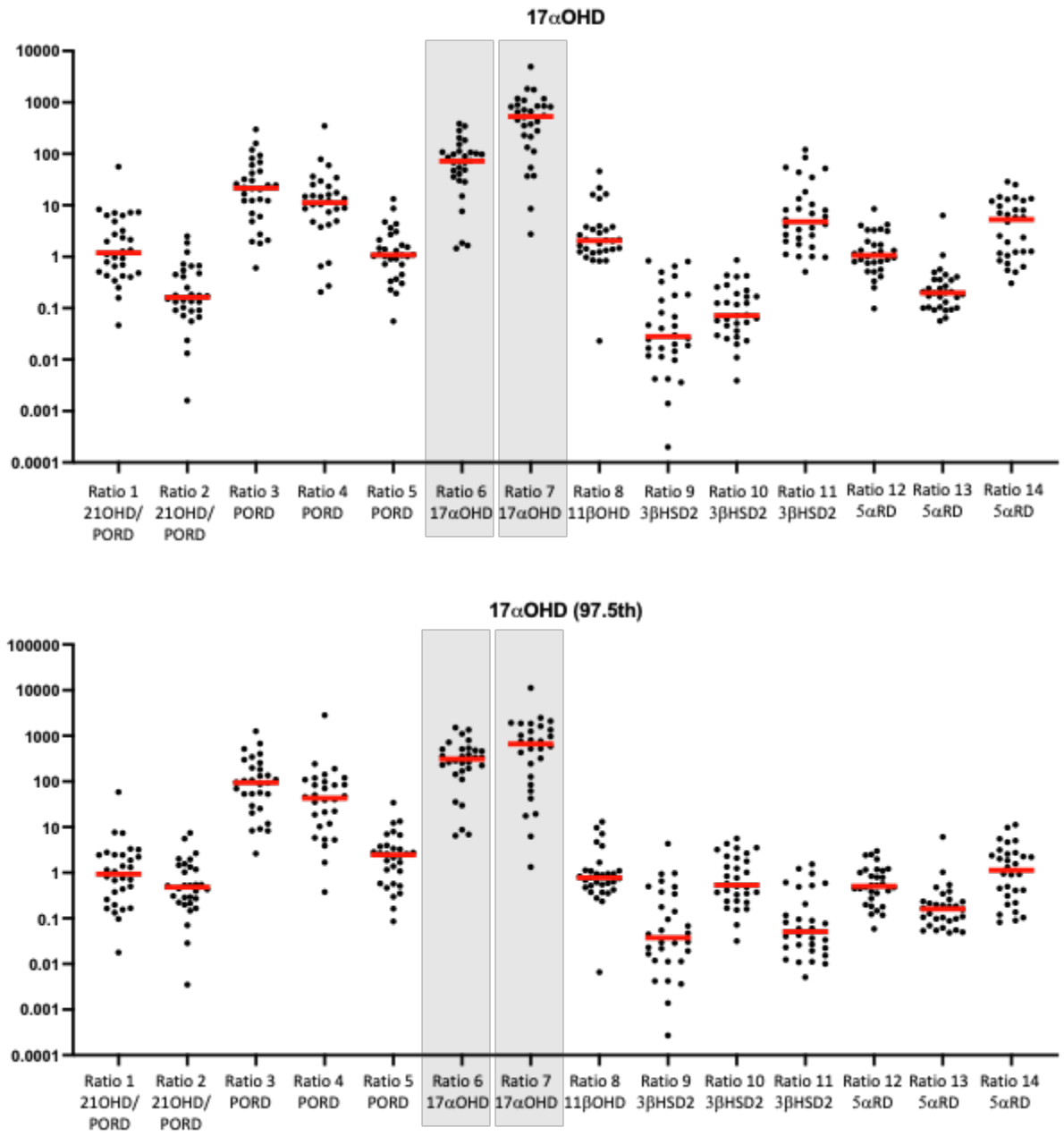


Figure 4.5: Ratio performance by condition, 17αOHD

14 urinary steroid metabolite ratios calculated for patients with 17αOHD, N=30. Results shown as a scatter plot, a single dot per patient, median is displayed by a red horizontal line. Top panel: Ratio values as absolute quantification. Bottom panel: Results shown as a proportion of the 97.5th percentile of the corresponding age-matched healthy control cohort group.

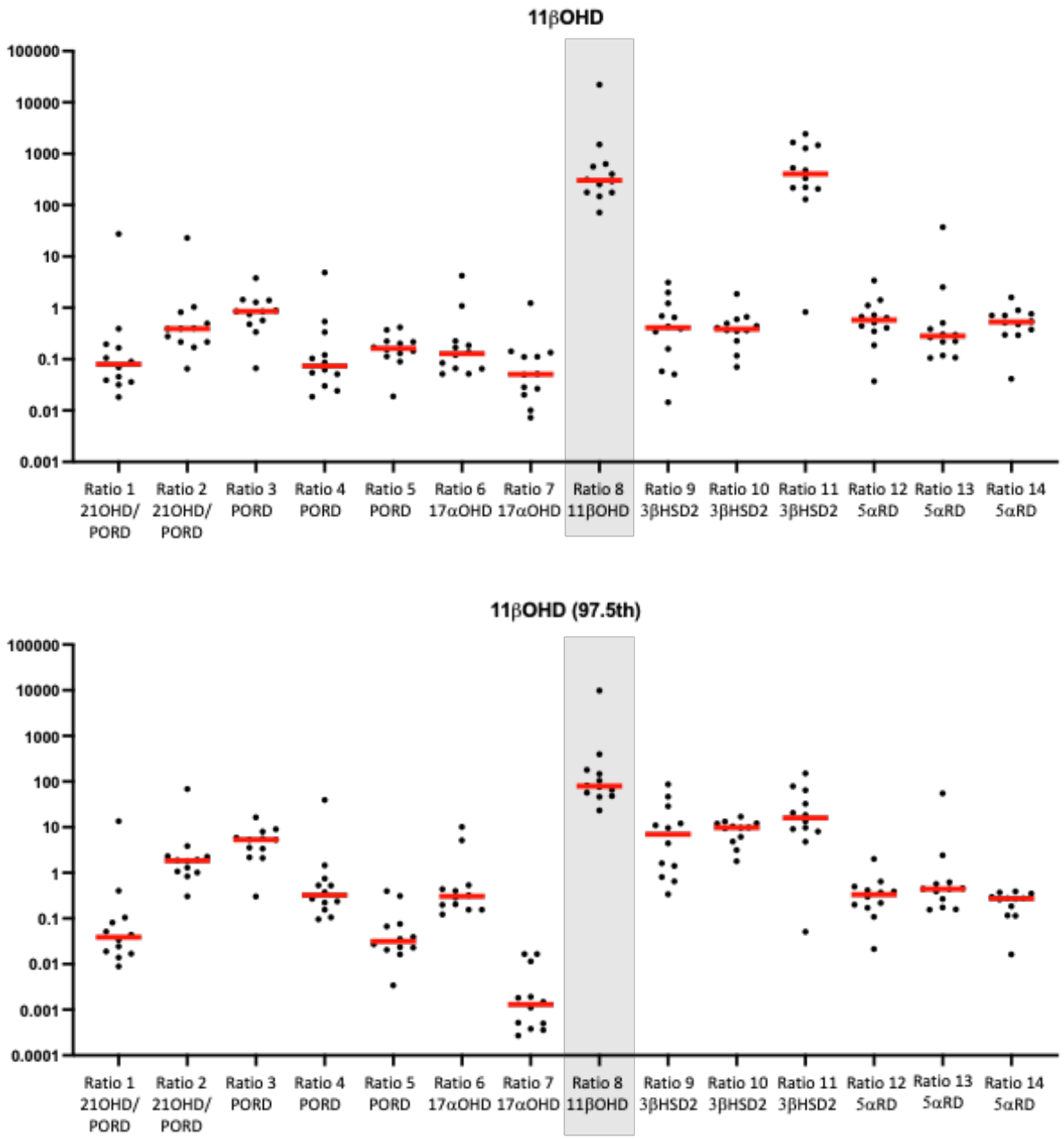


Figure 4.6: Ratio performance by condition, 11βOHD

14 urinary steroid metabolite ratios calculated for patients with 11βOHD, N=12. Results shown as a scatter plot, a single dot per subject, median is displayed by a red horizontal line. Top panel: Ratio values as absolute quantification. Bottom panel: Results shown as a proportion of the 97.5th percentile of the corresponding age-matched healthy control cohort group.

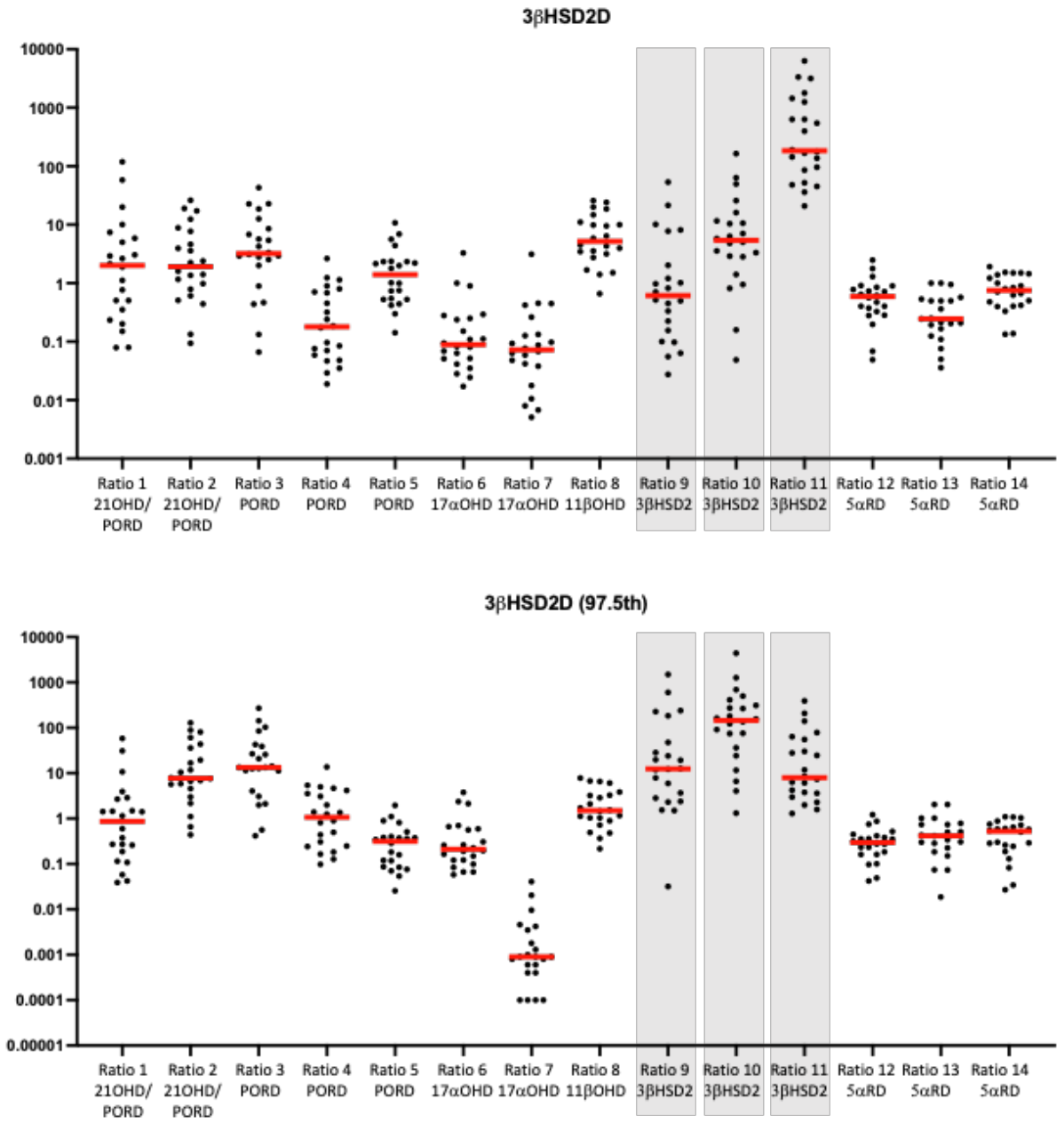


Figure 4.7: Ratio performance by condition, 3βHSD2D

14 urinary steroid metabolite ratios calculated for patients with 3βHSD2D, N=22. Results shown as a scatter plot, a single dot per patient, median displayed by a red horizontal line. Top panel: Ratio values as absolute quantification. Bottom panel: Results shown as a proportion of the 97.5th percentile of the corresponding age-matched healthy control cohort group.

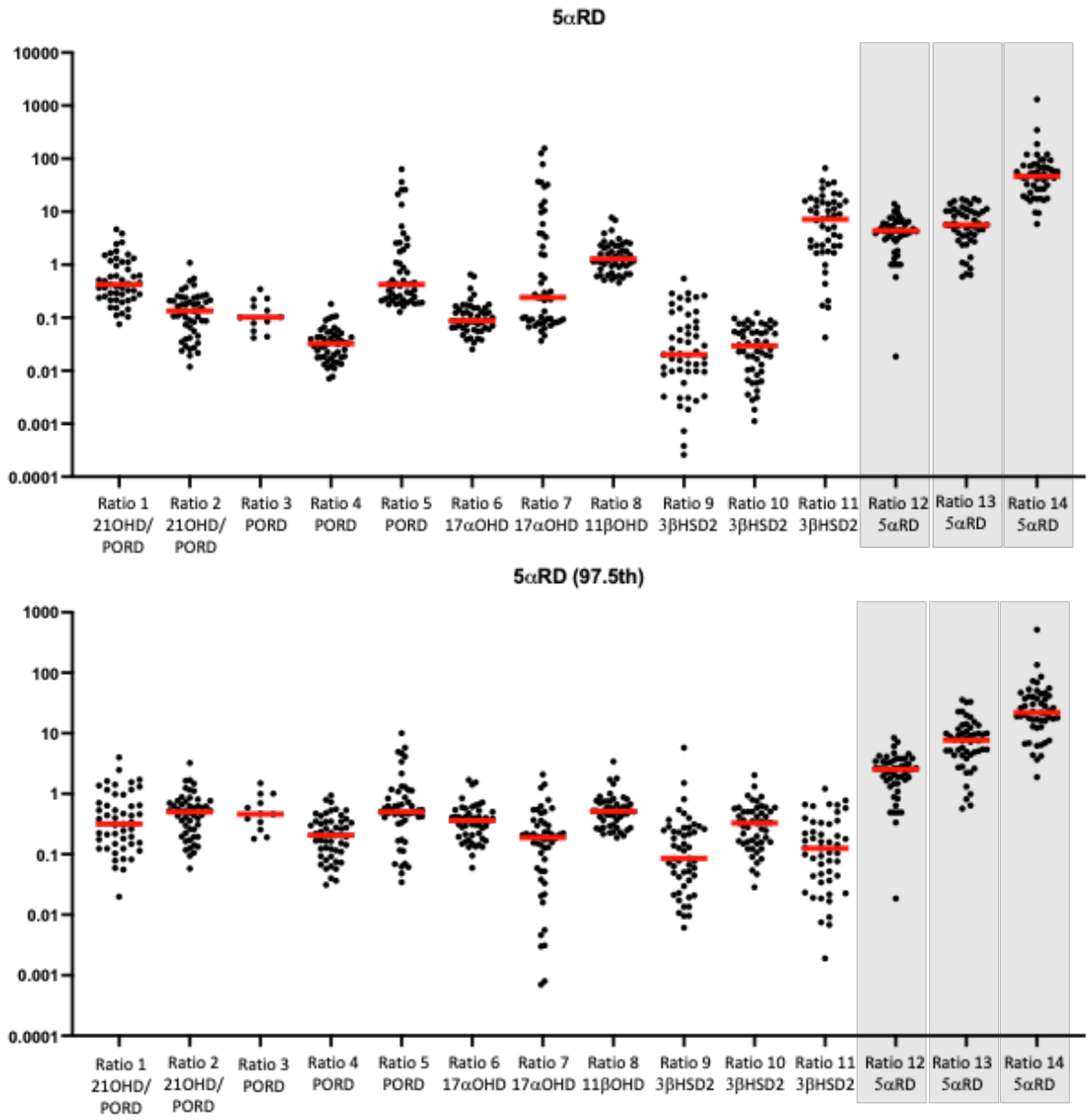


Figure 4.8: Ratio performance by condition, 5αRD

14 urinary steroid metabolite ratios calculated for patients with 5αRD, N=51. Results shown as a scatter plot, a single dot per subject, median displayed by a red horizontal line. Top panel: Ratio values as absolute quantification. Bottom panel: Results shown as a proportion of the 97.5th percentile of the corresponding age-matched healthy control cohort group.

4.6.2. Performance Using the 97.5th Percentile as the Upper Limit of Normal

The class-wise performance of the biochemical ratios when applied using our simple algorithm and using age- and sex-specific thresholds with a whole cohort analysis are shown in **Table 4.2**. Confidence intervals were constructed using the Wilson method and were not derived from cross validation where results were obtained using the whole cohort. Performance results are presented as a matrix of true vs. predicted diagnosis (also known as a "confusion matrix"). Sensitivity to diagnose a patient as a patient (regardless of specific diagnosis) was 100% (95% CI [97.9% – 100%]). Sensitivity to diagnose patients with the correct diagnosis was 89.9% (95% CI [84.5% – 93.5%]), meaning that 10.1% of patients were identified as patients but diagnosed with the wrong condition. Class-wise sensitivity ranged from 77.8% for PORD to 100% for 5αRD and 3βHSD2D. Specificity was 80% (95% CI [77.1% – 82.6%]).

Performance with only age-specific thresholds in a whole cohort analysis are shown in **Table 4.3**. Confidence intervals were constructed using the Wilson method and were not derived from cross validation where results were obtained using the whole cohort. Average (Lupien et al.) performance with standard deviation across five folds of cross validation, using age- and age- and sex-specific thresholds are shown in **Table 4.4** and corresponding matrix of average performance per class in **Table 4.5.** For results obtained using a five-fold cross validation approach, mean and standard deviation are shown over the five folds rather than confidence intervals.

Table 4.2: Matrix of true vs predicted results - Biochemical Ratios, 97.5th percentile as upper limit of normal

Age- and sex-specific thresholds used. Diagonal (shaded grey) are correctly identified per class. Sensitivity (Patients identified as having any disease) = 100% (95% CI 97.9% - 100%), Sensitivity (Patients identified as having the correct disease = 89.9% (95% CI 84.5% - 93.5%), Specificity (Healthy individuals correctly identified at healthy) = 80% (95% CI 77.1% - 82.6%).

	Predicted Healthy	Predicted 21OHD	Predicted 11βOHD	Predicted PORD	Predicted 17αOHD	Predicted 3βHSD2D	Predicted 5αRD	Total per class	Class-wise Sensitivity [95%CI]
true Healthy	663	19	12	34	30	33	38	829	80.0% [77.1, 82.6]
true 21OHD	0	22	0	5	0	0	0	27	81.5% [63.3, 91.8]
true 11βOHD	0	0	10	0	0	2	0	12	83.3% [55.2, 95.3]
true PORD	0	6	0	28	2	0	0	36	77.8% [61.9, 88.3]
true 17αOHD	0	0	0	3	27	0	0	30	90.0% [74.4, 96.5]
true 3βHSD2D	0	0	0	0	0	22	0	22	100.0% [85.1, 100]
true 5αRD	0	0	0	0	0	0	51	51	100.0% [93.0, 100]

Table 4.3: Matrix of true vs predicted results- Biochemical Ratios, 97.5th percentile as upper limit of normal

Age-specific thresholds used. Diagonal (shaded grey) are correctly identified per class. Sensitivity (Patients identified as having any disease) = 100% (95% CI 97.9% - 100%), Sensitivity (Patients identified as having the correct disease = 88.2% (95% CI 82.6% - 92.2%), Specificity (Healthy individuals correctly identified at healthy) = 80% (95% CI 77.1% - 82.6%).

	Predicted Healthy	Predicted 21OHD	Predicted 11βOHD	Predicted PORD	Predicted 17αOHD	Predicted 3βHSD2D	Predicted 5αRD	Total per class	Class-wise Sensitivity [95%CI]
true Healthy	663	21	11	38	28	31	37	829	80.0% [77.1,82.6]
true 21OHD	0	22	0	5	0	0	0	27	81.5% [63.3,91.8]
true 11βOHD	0	0	10	0	0	2	0	12	83.3% [55.2,95.3]
true PORD	0	7	0	25	4	0	0	36	69.4% [53.1,82.0]
true 17αOHD	0	0	0	3	27	0	0	30	90.0% [74.4,96.5]
true 3βHSD2D	0	0	0	0	0	22	0	22	100.0% [85.1,100]
true 5αRD	0	0	0	0	0	0	51	51	100.0% [93.0,100]

Table 4.4: Performance of biochemical ratios across 5 folds of cross validation

Age- and sex-specific 97.5% percentile of healthy cohort in each training fold used as upper limit of normal. Mean and standard deviation of test fold performance over 5 folds of cross validation is shown.

	Mean	std
specificity	72.7%	0.04
sensitivity (any disease)	100.0%	0.00
sensitivity (correct disease)	90.4%	0.05
21OHD	81.3%	0.16
11βOHD	83.3%	0.21
PORD	83.6%	0.04
17αOHD	85.3%	0.15
3βHSD2D	100.0%	0.00
5αRD	100.0%	0.00

Table 4.5: Matrix of true vs predicted results for biochemical ratios across 5 folds of cross validation

Average results across five test folds are shown.

	Predicted Healthy	Predicted 21OHD	Predicted 11βOHD	Predicted PORD	Predicted 17αOHD	Predicted 3βHSD2D	Predicted 5αRD	Total per class
true Healthy	120.6	9.8	3.4	7.8	5.2	8.8	10.2	165.8
true 21OHD	0	4.4	0	1	0	0	0	5.4
true 11βOHD	0		2	0	0	0.4	0	2.4
true PORD	0	1	0	6	0.2	0	0	7.2
true 17αOHD	0	0	0	0.8	5.2	0	0	6
true 3βHSD2D	0	0	0	0	0	4.4	0	4.4
true 5αRD	0	0	0	0	0	0	10.2	10.2

4.6.2.1. Analysis of the misclassified individuals, using the 97.5th percentile as upper limit of normal, age- and sex-specific thresholds

The matrices above show where misclassifications occur. Predictably, there was some confusion between the PORD, 21OHD and 17αOHD classes. There was some misclassification observed between 11βOHD and 3βHSD2D. Androgens are invariably high in 11βOHD and downstream glucocorticoids low, well known characteristic features of the condition, which explains the misclassification predominantly occurring due to the DHEA/THE+THF+5αTHF ratio.

Approximately equal numbers of healthy control misclassifications occurred with PORD, 17αOHD, 3βHSD2D and 5αRD. **Table 4.6** shows the healthy participant misclassifications in more detail. For male participants the most frequent age group where misclassifications occurred was the 6 months to 16 years age group (n=31), presumably an effect related to adrenarche / pubarche. For the female healthy participants, the most frequent age group where misclassifications occurred was the over 16 years age group (n=44), the most frequent misclassifications occurred with 17αOHD and 3βHSD2D (n=10 for both), closely followed by PORD and 5αRD (n=9 for both). We hypothesise this is due to normal ovulatory cycle variation which is not accounted for in this analysis. Across all age groups for female participants, the most frequent misclassifications occurred for 5αRD (n=22) and 3βHSD2D (n=20). For male participants, the misclassifications were spread more evenly across the disease classes for all age groups. For both sexes, the least number of misclassifications overall occurred with 11βOHD (female n=4, male n=8) and 21OHD (female n=9, male n= 10).

Table 4.6: Healthy participants misclassified using the 97.5th percentile ULN

Numbers of healthy participants with each predicted patient diagnosis are shown. Left panel shows male participants split by age group; right panel shows female participants split by age group.

Male	< 6 m	6 m – 16 yr	>16 yr	Total
21OHD	1	5	4	10
11βOHD	3	3	2	8
PORD	4	6	7	17
17αOHD	5	4	5	14
3βHSD2D	3	6	4	13
5αRD	4	7	5	16
Total	20	31	27	78

Female	< 6 m	6 m – 16 yr	>16 yr	Total
21OHD	1	4	4	9
11βOHD	1	1	2	4
PORD	4	4	9	17
17αOHD	2	4	10	16
3βHSD2D	5	5	10	20
5αRD	9	4	9	22
Total	22	22	44	88

4.6.3. Performance using a Multiple of the Median as the Upper limit of Normal

4.6.3.1. Relative Stability of Median Compared to 97.5th Percentile

For each of the 14 biochemical ratios, across 5 folds of cross validation training sets of data, the age-specific median and the 97.5th percentiles were calculated from healthy participants. The standard deviation across the 5 values was calculated to demonstrate the relative stability of each metric. The results shown in **Table 4.7** demonstrate less variability with the median across the five training folds than the 97.5th percentile.

Table 4.7: Relative stability of median compared to 97.5th percentile

Using three age groups, less than 6 months (<6m), 6 months to 16 years (child) and over 16 years (Arlt et al.). Standard deviation of the median and 97.5th percentile across 5 folds of training data used for cross validation are shown. Top table shows results for the median. Bottom table shows results for the 97.5th percentile.

STD of Median	Ratio 1	Ratio 2	Ratio 3	Ratio 4	Ratio 5	Ratio 6	Ratio 7	Ratio 8	Ratio 9	Ratio 10	Ratio 11	Ratio 12	Ratio 13	Ratio 14
<6m	8.8e-05	0.001	1.6e-04	0.005	0.127	0.002	0.215	0.029	2.9e-04	3.6e-04	0.108	0.018	0.021	0.007
Child	3.1e-05	3.0e-04	4.7e-04	2.3e-04	0.039	0.007	1.297	0.040	1.5e-04	4.7e-04	0.364	0.024	0.003	0.003
Adult	8.2e-05	0.003	0.001	7.2e-04	0.002	0.002	0.002	0.008	0.005	6.0e-04	0.980	0.014	0.004	0.029

STD of 97.5th C	Ratio 1	Ratio 2	Ratio 3	Ratio 4	Ratio 5	Ratio 6	Ratio 7	Ratio 8	Ratio 9	Ratio 10	Ratio 11	Ratio 12	Ratio 13	Ratio 14
<6m	0.014	0.016	0.004	0.242	1.070	0.068	7.637	0.374	0.005	0.032	3.340	0.001	0.624	0.575
Child	0.003	0.022	0.008	0.019	0.390	0.016	2.612	0.230	5.3e-04	0.002	0.376	0.033	0.064	0.180
Adult	0.002	0.020	0.060	0.010	0.006	0.004	0.064	0.127	0.047	0.012	1.023	0.148	0.204	0.280

4.6.3.2. Identification of the optimum multiples

ROC curves are shown for all ratios, with AUC values in **Figure 4.9**. ROC curves illustrating variability using the 5 different groups for 5 folds of cross-validation, with the 5 AUC values are shown in **Figure 4.10.**

As sensitivity and specificity were felt to not carry equivalent value for a diagnostic test of our purpose, we did not limit the multiple selection to the maximisation of Youden's *J* statistic (Youden, 1950), where *J* = sensitivity + specificity - 1, the maximum value of *J* indicating the position of maximum compromise of sensitivity and specificity. The position on each ROC curve where the multiple afforded the maximum sensitivity and the position where the rate of decline in sensitivity increases (indicating the two extreme points of the corner) were identified. This is shown in **Table 4.8.** We hypothesise that the "optimum" multiple is between these two thresholds. Initially, we applied the latter value as an upper threshold of normal and incrementally increased each of the multiples as necessary according to where misclassifications occurred to maximise the sensitivity (to diagnose a patient as the correct class). **Table 4.8** also shows the initial multiples selected for each ratio. Results using these thresholds are shown in **Table 4.9;** Sensitivity (Patients identified as having any disease) = 100% (95% CI 97.9-100), Sensitivity (Patients identified as having the correct disease = 93.8% (95% CI 89.3-96.5), Specificity (Healthy individuals correctly identified at healthy) = 76.7% (95% CI 73.7-79.5). Class wise sensitivity varied from 81.5% (21OHD, 95% CI 63.3-91.8) to 100% (11βOHD, 3βHSD2D, 5αRD).

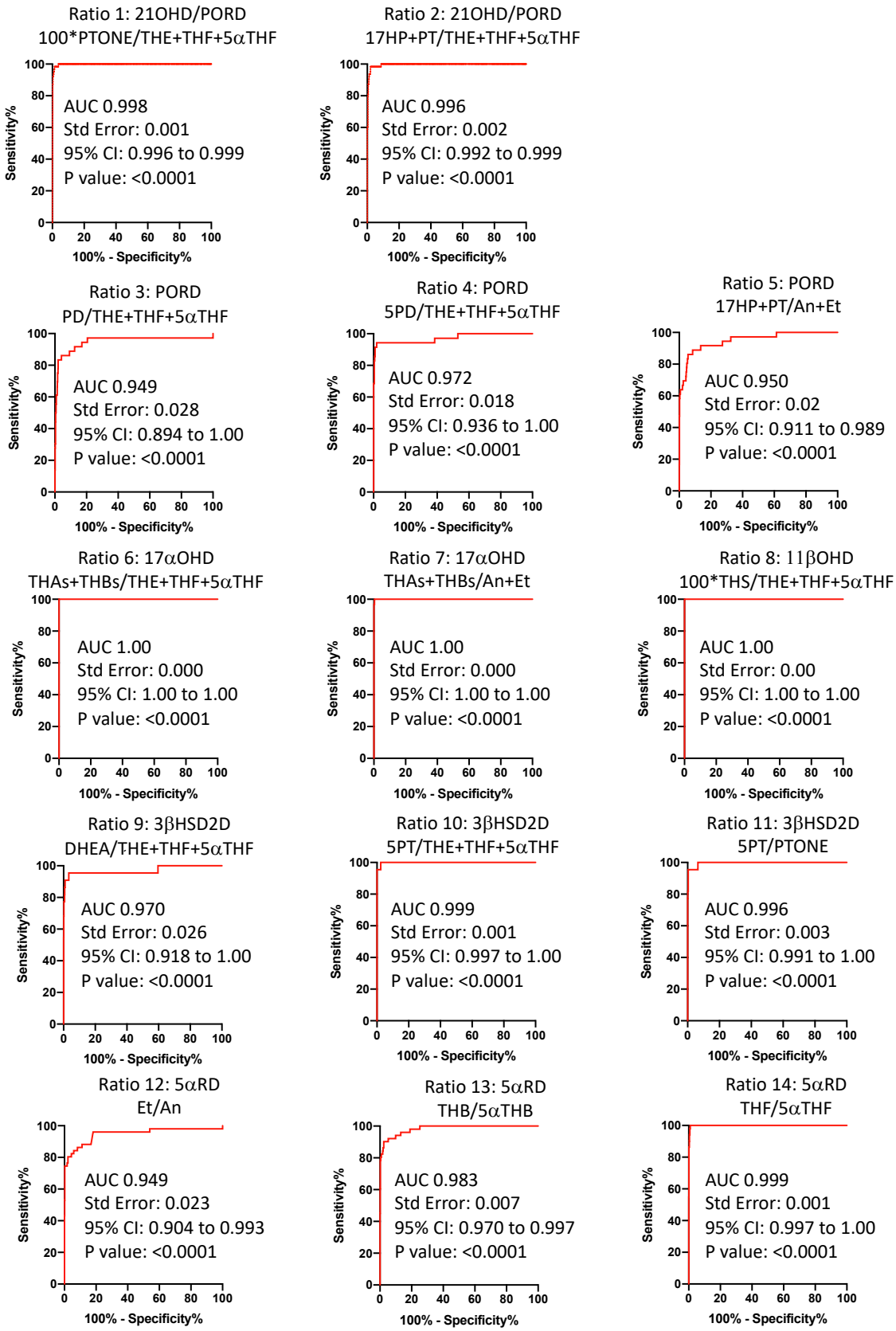


Figure 4.9: ROC curves for multiple of the median approach *Age- and sex-related medians used. Whole cohort informs analysis.*

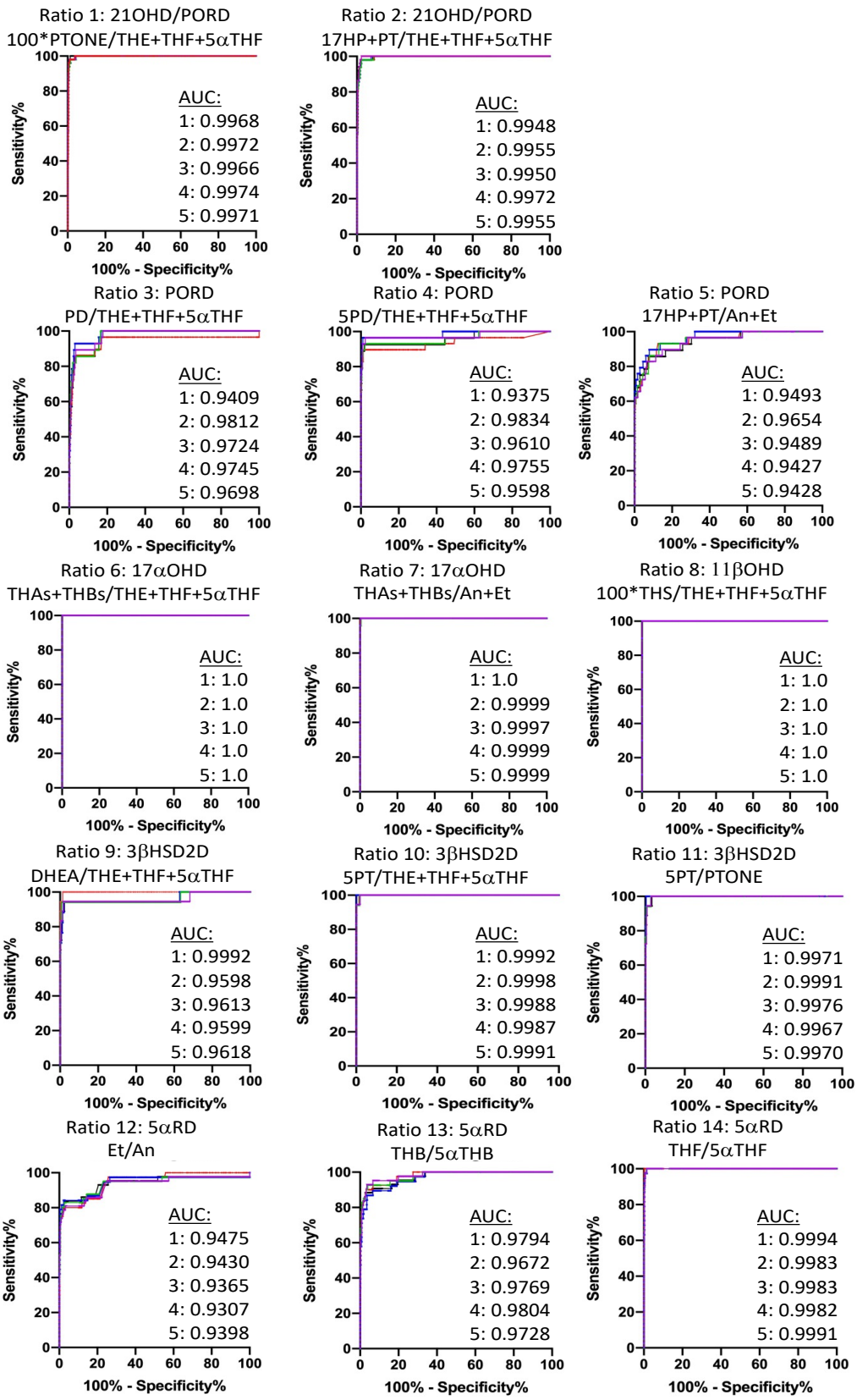


Figure 4.10: ROC curves for multiple of the median approach

Age- and sex-related medians used. 5 ROC curves from 5 sets of training data shown

Table 4.8: Multiple of the median, corner positions of ROC curve analysis

Thresholds which equated to the point of maximum sensitivity and the point where sensitivity exponentially declines. Sensitivity and specificity with confidence intervals for each threshold are shown.

Ratio	Threshold 1 (maximum sensitivity)	Sensitivity *(95% CI)*	Specificity *(95% CI)*	Threshold 2 (sensitivity declines)	Sensitivity *(95% CI)*	Specificity (95% CI)	**Multiple Applied to the Median**
100*PT'ONE/(THE+THF+5αTHF)	**> 7.41**	100% *(94.3 – 100)*	96.2% *(94.7 – 97.3)*	**> 15.77**	98.4% *(91.5 – 99.9)*	98.7% *(97.6 – 99.3)*	**15.77**
17HP+PT/(THE+THF+5αTHF)	**> 2.65**	100% *(94.3 – 100)*	91.3% *(89.2 – 93.1)*	**> 5.27**	98.4% *(91.5 – 99.9)*	98.0% *(96.7 - 98.7)*	**5.27**
PD/(THE+THF+5αTHF)	**> 2.10**	97.2% *(85.8 – 99.9)*	79.5% *(76.6 – 82.1)*	**> 7.39**	86.1% *(71.3 – 93.9)*	95.9% *(94.3 – 97.0)*	**5.37**
5PD/(THE+THF+5αTHF)	**> 0.96**	100% *(90.1 – 100)*	47.0% *(43.3 – 50.7)*	**> 5.44**	94.3% *(81.4 – 99.0)*	98.0% (96.7 – 98.8)	**5.44**
(17HP+PT)/(An+Et)	**> 0.85**	100% *(90.4 – 100)*	38.6% *(35.3 – 42.0)*	**> 2.54**	86.1% *(71.3 – 93.9)*	94.5% *(92.7. - 95.8)*	**2.37**
(THA+5αTHA+THB+5αTHB)/ (THE+THF+5αTHF)	**> 10.6**	100% *(88.6 – 100)*	100% *(95.5 – 100)*	**> 10.6**	100% *(88.6 – 100)*	100% *(95.5 – 100)*	**2.39**
(THA+5αTHA+THB+5αTHB)/ (An+Et)	**> 6.02**	100% *(88.6 – 100)*	99.6% *(98.9 – 99.9)*	**> 6.02**	100% *(88.6 – 100)*	99.6% *(98.9 – 99.9)*	**6.02**
100*THS/(THE+THF+5αTHF)	**> 35.1**	100% *(75.8 – 100)*	100% *(99.5 – 100)*	**> 35.1**	100% *(75.8 – 100)*	100% *(99.5 – 100)*	**2.41**
DHEA/(THE+THF+5αTHF)	**> 0.79**	100% *(85.1 – 100)*	40.4% *(37.1 – 43.8)*	**> 25.5**	90.9% *(72.2 – 98.4)*	99.0% *(98.1 – 99.5)*	**25.5**
5PT/(THE+THF+5αTHF)	**> 4.54**	100% *(85.1 – 100)*	97.6% *(96.3 – 98.4)*	**> 21.1**	95.5% *(78.2 – 99.8)*	100% *(99.5 – 100)*	**4.99**
5PT/PT'ONE	**> 3.69**	100% *(85.1 – 100)*	93.5% *(91.6 – 95.0)*	**> 8.48**	95.5% *(78.2 – 99.8)*	99.6% *(98.9 – 99.9)*	**8.48**
Et/An	**> 0.95**	98.0% *(89.7 – 99.9)*	46.1% *(42.7 – 49.5)*	**> 3.35**	80.4% *(67.5 – 89.0)*	97.8% *(96.6 – 98.6)*	**3.35**
THB/5αTHB	**> 1.47**	100% *(93 – 100)*	74.8% *(71.7 – 77.6)*	**> 5.67**	90.2% *(79.0 – 95.7)*	97.6% *(96.3 – 98.4)*	**5.67**
THF/5αTHF	**> 8.95**	100% *(93 – 100)*	99.0% *(98.1 – 99.5)*	**> 8.95**	100% *(93 – 100)*	99.0% *(98.1 – 99.5)*	**6.49**

Table 4.9: Matrix of true vs predicted results for biochemical ratios, multiple of the median approach

*Results calculated using the decision tool in **Figure 4.1**. Multiples as determined in **Table 4.8** applied to age- and sex-specific healthy median used as upper limit of normal (ULN). Diagonal (shaded grey) are correctly identified individuals in each class. Sensitivity (Patients identified as having any disease) = 100% (95% CI 97.9% - 100%), Sensitivity (Patients identified as having the correct disease = 93.8% (95% CI 89.3% - 96.5%), Specificity (Healthy individuals identified as healthy) = 76.7% (95% CI 73.7% - 79.5%).*

	Predicted Healthy	Predicted 21OHD	Predicted 11βOHD	Predicted PORD	Predicted 17αOHD	Predicted 3βHSD2D	Predicted 5αRD	Total per class	Class-wise Sensitivity [95%CI]
True Healthy	636	12	28	88	15	15	35	829	76.7% [73.7,79.5]
true 21OHD	0	22	0	5	0	0	0	27	81.5% [63.3,91.8]
true 11βOHD	0	0	12	0	0	0	0	12	100% [75.8, 100]
true PORD	0	3	0	33	0	0	0	36	91.7% [78.2,97.1]
true 17αOHD	0	0	0	3	27	0	0	30	90.0% [74.4, 96.5]
true 3βHSD2D	0	0	0	0	0	22	0	22	100.0% [85.1, 100]
true 5αRD	0	0	0	0	0	0	51	51	100.0% [93.0, 100]

4.6.3.3. Multiple of the Median Method Refinement using "Cost-Weighted Decision Making"

In the next chapter, we describe a cost weighting approach, where we define a priority order for classification to optimising the machine learning model. In descending order of importance: The 1st priority is to correctly identify patients as patients (of any class). The 2nd priority is to correctly identify 21OHD patients as this is the most common disorder. The 3rd priority is to correctly identify CAH patients as CAH patients (not 5αRD). The 4th priority is correct subclass diagnosis for patients. The 5th priority is correct identification of healthy class. In a heuristic approach, to show the optimum performance of a biochemical ratio-based classifier using the multiples of the median, we applied this same cost weighting to adapt our decision tool. We already satisfy the first priority, with no patients diagnosed as healthy. 4 of 5 misclassifications of 21OHD patients occur because of one ratio ((17HP+PT)/(An+Et)). We implemented an additional criterion to the algorithm whereby if this ratio is the highest, then we look at the second highest ratio to define diagnosis. If the second highest ratio is intended for 21OHD, then the diagnosis is 21OHD, otherwise the diagnosis is PORD. This addition improved 21OHD class performance without impacting PORD class performance. No patients with CAH are mistaken for having 5αRD. The 11βOHD class required a dramatic reduction in the multiple chosen for the median in order to maximise sensitivity, due to raised DHEA/Glucocorticoid metabolites ratio (ratio 9) causing confusion with 3βHSD2D. Patients with 3βHSD2D did not show elevated values for ratio 8 (intended for 11βOHD). Therefore, if the value for ratio 8 is elevated, then the diagnosis is more likely to be 11βOHD. Inclusion of this additional criterion allowed us to increase the threshold for ratio 8 to improve specificity, without compromising

sensitivity. The adjusted decision tool is shown in **Figure 4.11.** How the multiples chosen relate to whole number thresholds is shown in **Table 4.10.**

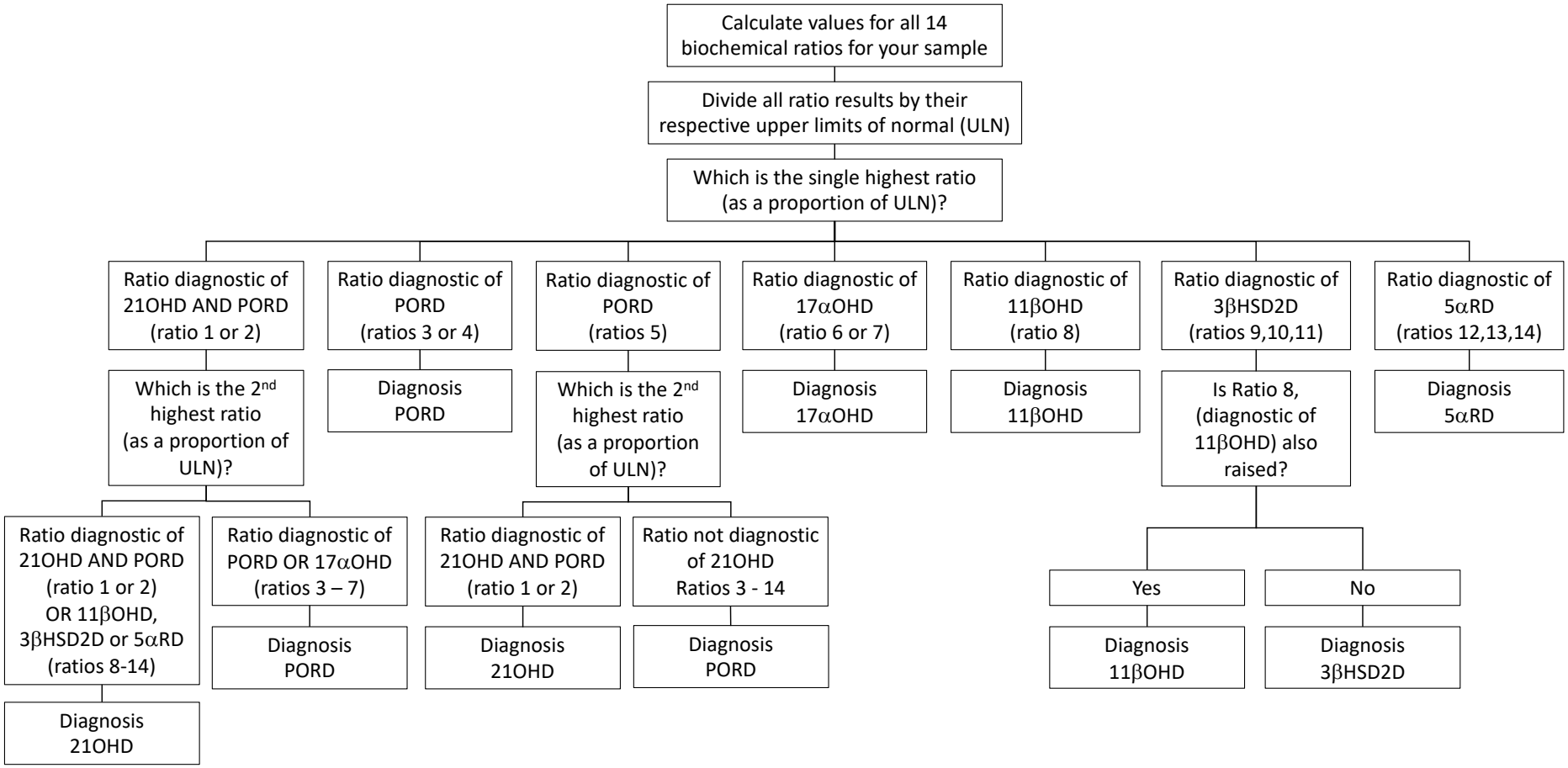


Figure 4.11: Refined decision tool for the interpretation of urinary steroid metabolome data *Use with table 2.3.*

Additional detail derived from heuristic decisions made in view of known characteristics of disease to minimise error in classification.

Table 4.10: Final multiples applied to each median and determined thresholds

Sensitivity and specificity plus confidence intervals shown for each multiple from ROC analysis. Absolute age- and sex-specific reference intervals calculated using these multiples are stated.

Ratio	Multiple Applied to the Median	Sensitivity (95%CI)	Specificity (95%CI)	Threshold Male <6m	Threshold Female <6m	Threshold Male 6m- 16yr	Threshold Female 6m – 16yr	Threshold Male >16yr	Threshold Female >16yr
1) 100*PT'ONE/ (THE+THF+5αTHF)	15.77	98.4 (91.5 – 99.9)	98.7 (97.6 – 99.3)	5.93	8.10	3.09	3.33	2.99	2.99
2) 17HP+PT/ (THE+THF+5αTHF)	5.27	98.4 (91.5 – 99.9)	98.0 (96.7 – 98.7)	0.15	0.15	0.11	0.16	0.64	0.56
3) PD/(THE+THF+5αTHF)	5.37	86.1 (71.3 – 93.9)	93.6 (91.7 – 95.1)	0.05	0.06	0.05	0.09	0.15	0.27
4) 5PD/(THE+THF+5αTHF)	5.44	94.3 (81.4 – 99.0)	98.0 (96.7 – 98.8)	1.21	1.15	0.22	0.30	0.30	0.23
5) (17HP+PT)/(An+Et)	2.37	86.1 (71.3 – 93.9)	93.1 (91.2 – 94.7)	3.77	5.91	4.36	3.76	0.52	0.49
6) (THA+5αTHA+THB+5αTHB)/ (THE+THF+5αTHF)	2.39	100 (88.6 – 100)	97.8 (96.6 – 98.6)	0.52	0.49	0.46	0.43	0.24	0.25
7) (THA+5αTHA+THB+5αTHB)/ (An+Et)	6.02	100 (88.6 – 100)	99.6 (98.9 – 99.9)	65.54	97.07	131.68	86.53	1.16	1.26
8) 100*THS/(THE+THF+5αTHF)	19.4	100 (75.8 – 100)	99.9 (99.3 – 100)	19.09	24.71	31.41	27.44	18.24	23.86
9) DHEA/(THE+THF+5αTHF)	25.5	90.9 (72.2 – 98.4)	99.0 (98.1 – 99.5)	0.17	0.26	0.05	0.10	1.79	0.88
10) 5PT/(THE+THF+5αTHF)	4.99	95.5 (78.2 – 99.8)	97.9 (96.7 – 98.7)	0.13	0.11	0.04	0.05	0.21	0.19
11) 5PT/PT'ONE	8.48	95.5 (78.2 – 99.8)	99.6 (98.9 – 99.9)	50.50	33.92	33.56	47.61	159.00	160.99
12) Et/An	3.35	80.4 (67.5 – 89.0)	97.8 (96.6 – 98.6)	0.85	2.46	2.12	2.04	2.75	3.68
13) THB/5αTHB	5.67	90.2 (79 – 95.7)	97.6 (96.3 – 98.4)	3.77	2.04	1.07	1.19	2.28	2.82
14) THF/5αTHF	6.49	100 (93 - 100)	98.6 (97.5 – 99.2)	1.92	3.25	2.67	2.79	6.79	9.25

4.6.3.4. Performance results

Results using the decision tool in **Figure 4.11** and age- and sex-specific thresholds as shown in **Table 4.10** are shown in **Table 4.11**. Sensitivity to correctly identify patients as patients was 100% (95% CI 97.9 – 100), sensitivity to identify patients as having the correct diagnosis was 94.4% (95% CI 92.1 – 98.1), specificity was 79.1% (95% CI 76.2 – 81.8%). Class wise sensitivity ranged from 90% (17αOHD 95% CI 74.4 – 97.1) to 100% for 11βOHD (95% CI 75.8 – 100%), 3βHSD2D (95% CI 85.1 – 100%) and 5αRD (95% CI 93 – 100%).

All sensitivity metrics were superior for this approach than using the 97.5th percentile as ULN (see **Table 4.12**). However, it must be considered that the results provided for the multiple of the median approach are likely optimistic, similar to what would be considered “training performance”, as the defined thresholds and results were derived from the same cohort. There remained some misclassification between 21OHD, PORD and 17αOHD, but less than observed when using the 97.5th percentile. Healthy control misclassification (see **Table 4.13**) is a construct of the selected threshold values so reflects the method rather than the internal characteristics of the healthy cohort. 98 misclassifications were for female healthy subjects and 75 male subjects. The majority of misclassified healthy subjects were incorrectly assigned as PORD (n=91). Of these, 57 were female controls and 32 were female over 16 years. 22 were from the male 6month-16year age group. 23 males and 11 females under 6 months were misclassified as having 5αRD.

Table 4.11: Matrix of true vs predicted results for biochemical ratios, optimised multiple of the median approach

*Results calculated using the decision tool in **Figure 4.11**. Multiples as determined in **Table 4.10** applied to age- and sex-specific healthy median used as upper limit of normal (ULN). Diagonal (shaded grey) are correctly identified individuals in each class. Sensitivity to correctly identify patients as patients= 100% (95% CI 97.9 – 100), sensitivity to identify patients as having the correct diagnosis= 94.4% (95% CI 92.1 – 98.1), specificity= 79.1% (95% CI 76.2 – 81.8%).*

	Predicted Healthy	Predicted 21OHD	Predicted 11βOHD	Predicted PORD	Predicted 17αOHD	Predicted 3βHSD2D	Predicted 5αRD	Total per class	Class-wise Sensitivity [95%CI]
true Healthy	656	14	0	91	15	16	37	829	79.1% [76.2,81.8]
true 21OHD	0	26	0	1	0	0	0	27	96.3% [81.7, 99.3]
true 11βOHD	0	0	12	0	0	0	0	12	100% [75.8, 100]
true PORD	0	3	0	33	0	0	0	36	91.7% [78.2,97.1]
true 17αOHD	0	0	0	3	27	0	0	30	90.0% [74.4, 96.5]
true 3βHSD2D	0	0	0	0	0	22	0	22	100.0% [85.1, 100]
true 5αRD	0	0	0	0	0	0	51	51	100.0% [93.0, 100]

Table 4.12: Comparative performance of 97.5th percentile vs. Multiple of the Median approach.

	Biochemical Ratios, 97.5th percentile ULN [95%CI]	**Biochemical Ratios, Multiple of Median ULN** [95%CI]
Sensitivity (Any disease)	100% [97.9 - 100]	100% [97.9 - 100]
Sensitivity (Correct disease)	89.9% [84.5 - 93.5]	94.4% [92.1 - 98.1]
Healthy (Specificity)	80% [77.1 - 82.6]	79.1% [76.2 - 81.8]
21-hydroxylase deficiency	81.5% [63.3 - 91.8]	96.3% [81.7 - 99.3]
11β-hydroxylase deficiency	83.3% [55.2 - 95.3]	100% [75.8 - 100]
P450 oxidoreductase deficiency	77.8% [61.9 - 88.3]	91.7% [78.2 - 97.1]
17α-hydroxylase deficiency	90.0% [74.4 - 96.5]	90.0% [74.4 - 96.5]
3β-hydroxysteroid dehydrogenase type 2 deficiency	100.0% [85.1 - 100]	100.0% [85.1 - 100]
5α-reductase type 2 deficiency	100.0% [93.0 - 100]	100.0% [93.0 - 100]

Table 4.13: Healthy control misclassifications using the optimised multiple of the median approach. Age- and sex-specific thresholds used

	Female n = 98			Male n = 75		
	< 6m	6m-16yr	> 16yr	< 6m	6m-16yr	> 16yr
21OHD	2	1	3	1	7	0
PORD	10	15	32	9	22	3
17αOHD	7	1	5	0	1	1
11βOHD	0	0	0	0	0	0
3βHSD2D	2	1	6	5	2	0
5αRD	11	0	2	23	1	0
Total	32	18	48	38	33	4

4.7. Discussion

We present results for two approaches for using 14 previously published biochemical ratios to differentiate patients with inborn disorders of steroidogenesis from each other and from healthy controls. Use of the 97.5th percentile is a widely acceptable method for reference interval construction. We showed that the median, however, was a more stable marker than the 97.5th percentile. Therefore, reference intervals created using a multiple of the median may be more generalisable.

Our preliminary investigations highlight the non-exclusivity of these biochemical ratios, with raised values being observed outside their target population. This led to the development of a simple decision tool to assist with interpretation of results. We highlight the importance to show ratios as a proportion of a healthy threshold in order to scale results to enable direct comparison. Our preliminary work further highlighted the predictable difficulty in discriminating 21OHD, 17αOHD and PORD, confirmed by our classifier results. Given POR is an electron donor to CYP21A2 and CYP17A1 and shows features of both disorders, it is understandable that it would be difficult to create a diagnostic steroid metabolite ratio very specific to this disorder. It has previously been suggested that as 21OHD and PORD cohorts both show relevant ratio values above healthy controls, but 21OHD to a greater order of magnitude than PORD, an additional threshold could be used to differentiate the groups (Koyama et al., 2012). It is unclear whether this method could differentiate late onset forms of CAH. We encountered some confusion between 11βOHD and 3βHSD2D which was not anticipated, as patients with 11βOHD had significantly raised values for ratios intended for diagnosis of 3βHSD2D. Conversely, patients with 3βHSD2D did not show

significantly raised values for the ratio intended for diagnosis of 11βOHD, which was predictable as a raised 11-deoxycortisol is not a characteristic feature. We show for our multiple of the median approach how we can adjust for this confusion with a simple addition to the decision tool based on existing knowledge of these diseases. We observed no diagnostic confusion between 21OHD and 3βHSD2D using the 97.5th percentile ULN. 21OHD ratios were raised for our 3βHSD2D patients but were not proportionally highest in this group. However, using the multiple of the median approach, we were unable to reduce the thresholds to improve the 21OHD class performance without adjustment of the diagnostic tool because it would have resulted in a greater number of misclassifications within the 3βHSD2D and PORD classes.

Our upper thresholds determined by using the 97.5th percentile of healthy are of a similar order of magnitude to previously published values (de Jong et al., 2018; Lucas-Herald et al., 2015; Rousson et al., 2021). Using these thresholds, within the female healthy cohort, the over 16 years group showed the most misclassifications. We hypothesise this is explained by variability observed with the menstrual cycle and changes observed post menopause/adrenopause. Male healthy participants showed a comparatively more even distribution of misclassification with age, with the greatest number observed in the 6 months to 16 years age group, likely reflecting changes associated with adrenarche/puberty.

There were several challenges we encountered for this research objective. Both approaches shown use already published ratios of steroid metabolites, which is an inherent limitation of the study. New biochemical ratios could be determined which build on the increasing knowledge in this domain, such as the knowledge of alternative

pathways to active androgen synthesis and how they are relevant in these disorders. Our class sizes are relatively small, and we did not have age- and sex-matched patients for every subgroup threshold we wished to determine. We did not have a holdout, generalisation cohort. These are rare diseases and collecting a prospective test validation cohort would be unfeasible. Our healthy cohort was a healthy asymptomatic population, therefore does not fully represent the disease negative symptomatic population which would be the true negative class for our intended test. The pre-test probability of a positive test is not quantified and likely highly variable across centres. The prevalence of many of these diseases in the target symptomatic population is highly dependent on the geographical location. The cost of misdiagnosis if a patient is mistaken for a healthy control far outweighs the cost if a healthy control is mistaken for a patient. Therefore, our optimum threshold is likely to not be a balance of sensitivity and specificity as sensitivity is considered more important.

With these challenges in mind, choosing a percentile as an upper limit of normal, such as the 97.5th percentile, is the least biased and likely to be the most generalisable approach. Thresholds are determined by the numerically largest group; they are not affected by the patient cohort, and we can easily create age- and sex-specific thresholds even in the absence of matched patients. This approach is the more widely published for urinary steroid metabolome data (Ackermann et al., 2019; de Jong et al., 2018; Rousson et al., 2021). A multiple of the median approach is used for antenatal screening tests and has been proposed as a more stable strategy for determining reference intervals, including for urinary steroid metabolome data (Zalas et al., 2018). We took a heuristic approach to implement this strategy. We had to assume that patient data distribution for each ratio was static for each age- and sex-specific group

with respect to matched healthy controls in order to determine thresholds in the absence of matched patients. If we had a greater number of matched patients, it would have been feasible to pursue statistical modelling or a machine learning approach to determine optimum thresholds more specific to age and sex. In the absence of sufficient patient numbers, heuristic decisions were made for threshold optimisation based on occurrence of misclassifications. This method approximates training within a machine learning method- the most ideal fit for our data. Our optimisation approach runs the risk of "overfitting"- by optimising the method to be able to correctly diagnose the most patients out of those we have included in our cohort, we run the risk of creating a method that is not generalisable to the wider population, particularly when we are making decisions based on small class numbers. Therefore, the results obtained from our multiple of the median approach should be interpreted with caution as performance in a wider population has not been assessed. Performance estimates from whole cohort approaches are shown with confidence intervals and these are appropriately wide which further illustrates the consequence of small class sizes. By keeping the diagnostic tool as simple as possible, by splitting our subgroups to as few as possible and by using biological knowledge to drive the decisions we make to alter the diagnostic method created, we maintain some degree of potential generalisability. It would only be through testing this method against a prospective validation cohort that a more accurate population performance estimate can be made. While we cannot make conclusions about wider population performance, we can surmise that the approach presented here is a feasible method to construct thresholds of normality, with the ability to generalise thresholds for groups where healthy data may be present but patient data may be absent.

The initial decision tool was intentionally designed to be uncomplicated. The more complex version we present includes additional criteria based on known characteristics of these diseases and could be implemented to improve the 97.5th percentile approach. It is possible we could improve both presented classifiers further through multiple strategies. We could alter the number of included ratios, increase the complexity of the decision tool to include further prior known characteristic details for each disease, increase the number of age groups examined, include clinical information such as reported symptoms/signs, pubertal staging, or menstrual cycle details for post pubertal female participants.

4.8. Conclusion

We outline a simplified method using previously published ratios of steroid metabolites for standardising manual analysis of urinary steroid metabolome data as quantified by GC-MS. The calculation of 14 biochemical steroid ratios with our decision tool to aid interpretation performs well with high sensitivity and acceptable specificity. Further optimisation of this biochemical approach to the differential diagnosis of inborn steroidogenic disorders is likely possible, and we present a promising method using a multiple of the median approach.

5. Use of Machine Learning for Automated Differential Diagnosis of Inborn Steroidogenic Disorders

5.1. Introduction

The biochemical diagnosis of inborn disorders of steroidogenesis by examining urinary steroid metabolite profiles is long established. However, urinary steroid metabolome analysis is a highly specialised service, limited to only a select number of centres which house the required expertise for sample processing, analysis and interpretation. Classical descriptions of the observed urinary steroid metabolome for each enzymatic cause of these inborn disorders of steroidogenesis focus on a select number of metabolites per condition, characteristically observed to be grossly abnormal (Shackleton et al., 1980; Storbeck et al., 2019). Since the original descriptions of these conditions, knowledge of the biochemical process of steroidogenesis has increased and the steroidogenesis pathway as we know it has gained in complexity (Miller & Auchus, 2011). Aside from the discovery of P450 oxidoreductase as the causative defect in what was previously described as "Combined 17α–/21-hydroxylase deficiency" (Shackleton et al., 1980), other important advancements include the characterisation of the alternative pathway to DHT synthesis and the 11-oxygenated androgen pathway. Recent work has deduced that in 21-hydroxylase deficiency, not only are classical androgens upregulated, but also 11-oxygenated androgens and metabolites of the alternative pathway the DHT synthesis (Jones et al., 2017; Kamrath et al., 2012; Pretorius et al., 2017; Turcu et al., 2016). In P450 oxidoreductase deficiency, the mechanism behind the perplexing disparity at presentation of both undervirilisation in genetic males and over virilisation in genetic females observed in newborns has now been elucidated and is attributable to in utero activation of the alternative pathway to androgen synthesis (Reisch et al., 2019). This additional knowledge of the mechanism of steroidogenesis in healthy and patient populations

has potential implications for both diagnosis and treatment monitoring in this complex group of conditions (Jha et al., 2021; Jones et al., 2017; Kamrath et al., 2017). Due to the rarity of inborn disorders of steroidogenesis, it has not been comprehensively examined if there are significant observable alterations in the wider urinary steroid metabolome which could augment existing disease descriptions, assist with the accuracy of biochemical diagnosis, and improve our knowledge of the pathogenesis of these diseases.

Machine learning techniques are being increasingly used within biomedical domains and provide a powerful tool which can be used to identify intrinsic patterns within high complexity data. Once intrinsic patterns have been identified, classification schemes can be inferred from labelled example data and applied prospectively to a wider population of unlabelled, unclassified data. Urine steroid metabolomics, the combination of urinary steroid metabolite quantification through mass spectrometry techniques with machine learning based analysis, has been proposed as a highly sensitive and specific biomarker tool for adrenal disease (Arlt et al., 2011; Bancos et al., 2020; Chortis et al., 2020). Specifically, Arlt et al use urinary steroid metabolite profiling by GC-MS with a computational data analysis method called GMLVQ (Generalized Matrix Learning Vector Quantization) for discriminating benign from malignant adrenal tumours (Arlt et al., 2011).

The Learning Vector Quantization (LVQ) methods (Biehl et al., 2016; Kaden et al., 2014; Nova & Estévez, 2014) are nearest prototype classification schemes, where prototypes are learned typical representative examples of each class, as determined through a training process using labelled example data. Data for each sample is

mathematically expressed as a single vector in the input space. Prototype vectors are defined in the input space and during training the prototype vector moves according to the labelled training data so that sample vectors of the assigned class label are closest and other classes are further away. New samples are compared to all computed prototypes to ascertain which one it most resembles. A dissimilarity measure is employed to quantify resemblance to each class prototype, usually in terms of distance. For an unknown sample, dissimilarity to each prototype is calculated and the sample is labelled according to which one it is closest to (winner takes all approach). Decision boundaries in the input space (also known as Voronoi partitions) are determined by computing the equidistant positions between class prototypes, or the points where dissimilarity to two separate prototypes would be equal (Nova & Estévez, 2014). The prototypes capture the essential characteristics of the data specific to each class and it is possible to fully inspect these prototypes to examine their properties. This means these methods are particularly well suited to the healthcare domain due to their inherent transparency and interpretability (Hammer et al., 2011). LVQs can be easily adapted for multiclass problems, they are well suited to high dimensional data, are computationally efficient (Biehl et al., 2007) and can be forgiving regards missing values in a dataset without the need for imputation (Schneider, 2010).

Sato and Yamada (Sato & Yamada, 1995) proposed generalized learning vector quantization, a supervised training procedure, learning occurring through minimisation of the below cost function (see **Equation 5.1**). d represents the distance from sample vector ($\vec{x}_i$) to prototype vector ($\vec{w}_i$), J indicates the closest prototype to the training sample with the correct class label and K indicates the closest prototype with a

different class label. Therefore, λ_i represents the costs or the loss induced for a given training sample vector. E is the sum of all costs induced for all training sample vectors. Through an appropriate optimisation procedure, the aim of training is to find the optimum model parameters which minimise the value of E, meaning that labelled training samples of a given class are as close to their own labelled class prototype as possible and as far as possible away from the nearest incorrect class.

Equation 5.1

$$E = \sum_{i=1}^{N} f(\lambda_i), where\ \lambda_i = \frac{d_i^J - d_i^K}{d_i^J + d_i^K}$$

The original LVQ method (LVQ[1]) was proposed by Kohonen in 1986 (Kohonen, 1990) and since then several extensions to LVQ[1] have been proposed to enable users to improve performance according to the classification task. In Generalised Relevance Learning Vector Quantization (GRLVQ), an adaptive parameterised distance measure is implemented which weights features according to their importance for classification (Biehl et al., 2013; Hammer & Villmann, 2002). Generalized Matrix Learning Vector Quantization (GMLVQ) further builds on GRLVQ. A full adaptive relevance matrix is utilised in the distance measure and optimised during training alongside the prototypes. This strategy takes into account not only the relative weight of features according to their importance for classification but also features that may only be relevant in combination with other features (Biehl et al., 2013; Biehl et al., 2012; Schneider et al., 2009).

5.2. Aim and Objective

The objective which we present in this chapter was to create a data-driven but biologically relevant diagnostic tool for automated differential diagnosis of inborn disorders of adrenal steroidogenesis. To achieve this objective, we pursued a machine learning strategy. The model would be built using comprehensive urinary steroid metabolite measurements as quantified by GC-MS, collected from our cohort of 829 healthy controls and 178 patients with definitively determined diagnoses.

5.3. Biological Considerations

It is important for method accountability to retain maximum transparency and interpretability of any diagnostic method created. There is considerable overlap between steroidogenic enzyme defects, as shown in chapter 3, and the potential for misclassification will remain. It is important that we can explain why decisions are made, and importantly why misclassifications may occur. By using a data-driven approach, transparency affords the potential to gain domain knowledge in being able to understand the relative importance of different features for each disease class. The class sizes for each disease class in our cohort are highly imbalanced and do not accurately reflect the population prevalence of each disease nor the relative importance of each diagnosis to be correctly made. Furthermore, not all decisions are equally important. In routine clinical practice, what we create will be a tool in an armoury of potential investigations that a clinician can request for patients with a suspected Difference of Sex Development (Engels et al.) or Congenital Adrenal Hyperplasia (Hargitai et al.) diagnosis. The consequence of a healthy person being

misdiagnosed as a patient would be further investigations, such as hormonal and genetic tests, to confirm or refute any suggested diagnosis. However, if a patient is misdiagnosed as a healthy control, then the clinician may decide not to pursue any further investigations which would mean that the individual could be denied timely lifesaving glucocorticoid replacement therapy. Therefore, the sensitivity is much more important than the specificity; and identifying patients as having a problem is more important than correctly identifying exactly what the problem is.

5.4. Cohort description

The cohort and details of missing values are comprehensively described in chapter 2.

A detailed cohort description is included in **Table 2.1.** Subjects with less than 80% of the 32 steroid metabolites quantified were not included in the study database. **Figure 2.1** gives an overview of the missing metabolite values in the database, arising both systematically (missing not at random, MNAR) and non-systematically (missing at random, MAR).

5.5. Cross Validation Approach

Due to the small sample sizes and the lack of a prospective validation cohort, we used a cross validation approach to approximate generalizable performance values. We used 5 rounds of cross validation and we used stratified sampling to allocate an approximately equal distribution of samples for each class into each fold. Samples were ordered 1-1007. First 829 were healthy controls in ascending order of age. Then

followed each of the patient groups in sequence. All samples were sequentially allocated a number from 1 to 5, with labels systematically repeating down from sample 1 to sample 1007. This ensured an equal distribution of healthy controls respect to age in each fold. For the first round of cross validation, all samples allocated groups 1-4 were used for training and samples allocated group 5 were used for testing. For the 2nd round of cross validation, samples allocated groups 2-5 were used for training, with samples allocated 1 used for testing. Fold allocation continued in this way. This meant 80% of data was used for training and 20% of data used for testing performance in each of the five folds. The individuals allocated to each of the 5 test groups were unique to that group and unseen by the respective round of training.

5.6. Initial Exploratory Work

5.6.1. Methods

Features were constructed from all unique pairwise ratios of the 32 individual measured metabolites equating to 496 features. We opted to use ratios of metabolites to account for the fact that within our cohort we have a range of age groups and different collection methods (24-hour, timed collections, and random single spot urine collections), so the absolute values of single metabolites were not directly comparable between all individuals. Taking an unsupervised approach, using all features (all 496 pairwise ratios of measured metabolites), unlabelled samples were visualized using t-distributed stochastic neighbour embedding (t-SNE) that had been altered to operate

with angle distances to compare dissimilarity. Where there were missing values, dissimilarity was computed using the available dimensions only.

Once the visualisations had been created, samples were labelled post-hoc according to condition (patient group or healthy). To examine any age-related differences, individuals were highlighted by colour coding according to age. The youngest group is labelled in dark blue, then as age increases the colours move through the colour spectrum to the oldest group in dark red. The black samples were from adult participants (>18 years), but the exact age was unknown.

5.6.2. Results

Figure 5.1 and **Figure 5.2** show the results of the unsupervised embedding experiment. **Figure 5.1** shows all conditions on the same plot with healthy controls, with healthy controls coloured according to age as per the key to the right of the figure. **Figure 5.2** shows the healthy controls with individual conditions in isolation, extracted from the same visualization, with both healthy and patient groups coloured according to age. From these figures we can see that each of the patient conditions examined naturally cluster together, implying a distinct steroid metabolome pattern for each condition. There appeared to be some similarity between the 17αOHD group and the PORD group as these clusters overlap the most of all the groups. Interestingly, the 5αRD cohort appeared to project separately from the enzyme defects which cause CAH. 11βOHD, the smallest sized class, showed the least secure clustering. If we examine the healthy controls in the background of these two figures, they are shown as small dots which are coloured according to age as per the key to the right of the

figure. The healthy controls showed a striking clustering according to age and a clear functional transition with the youngest patients on the right of the figure, increasing in age as you move from right to left. In **Figure 5.2**, where the individual patient groups were also coloured according to age using the same colour scheme as healthy controls, patient groups did not display similar age-related differences.

We again used the angular t-SNE to visualise the healthy controls in isolation, without the patient groups, to confirm this age-related metabolome pattern was persistent. This is shown in **Figure 5.3**. We can observe from this figure that this relationship was a consistent finding and not simply a product of the embedding method. This implies that an individual's "adrenal age" can be visualized computationally through examining their steroid metabolome.

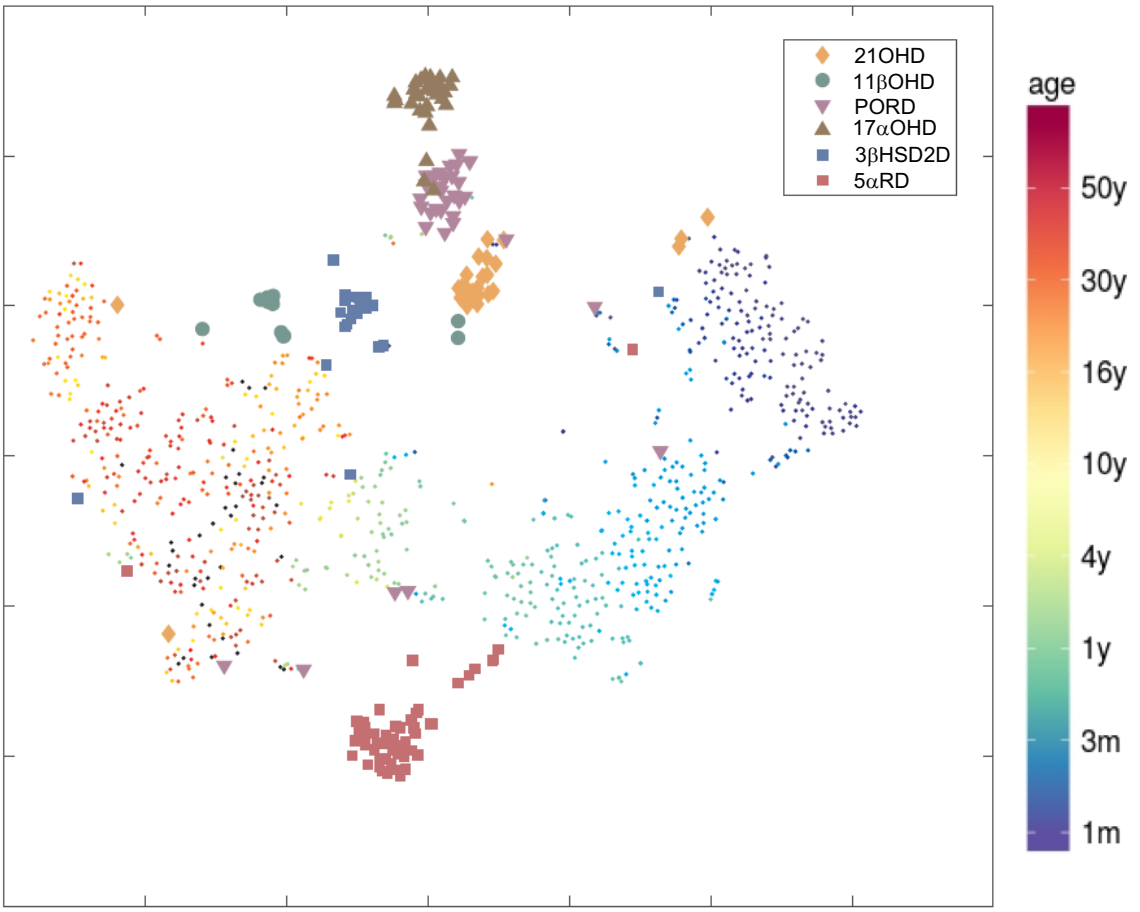


Figure 5.1: Unsupervised clustering using t-SNE

All conditions compared to healthy controls. Each individual patient participant is represented by a single symbol, the colour denotes which condition (see key at the top right of the figure). Healthy control participants are shows by dots with the colours relating to the age of the participant according to the scale to the right of the figure. The position of the symbol/dot determined by the pairwise interaction at all measure urinary steroid metabolites.

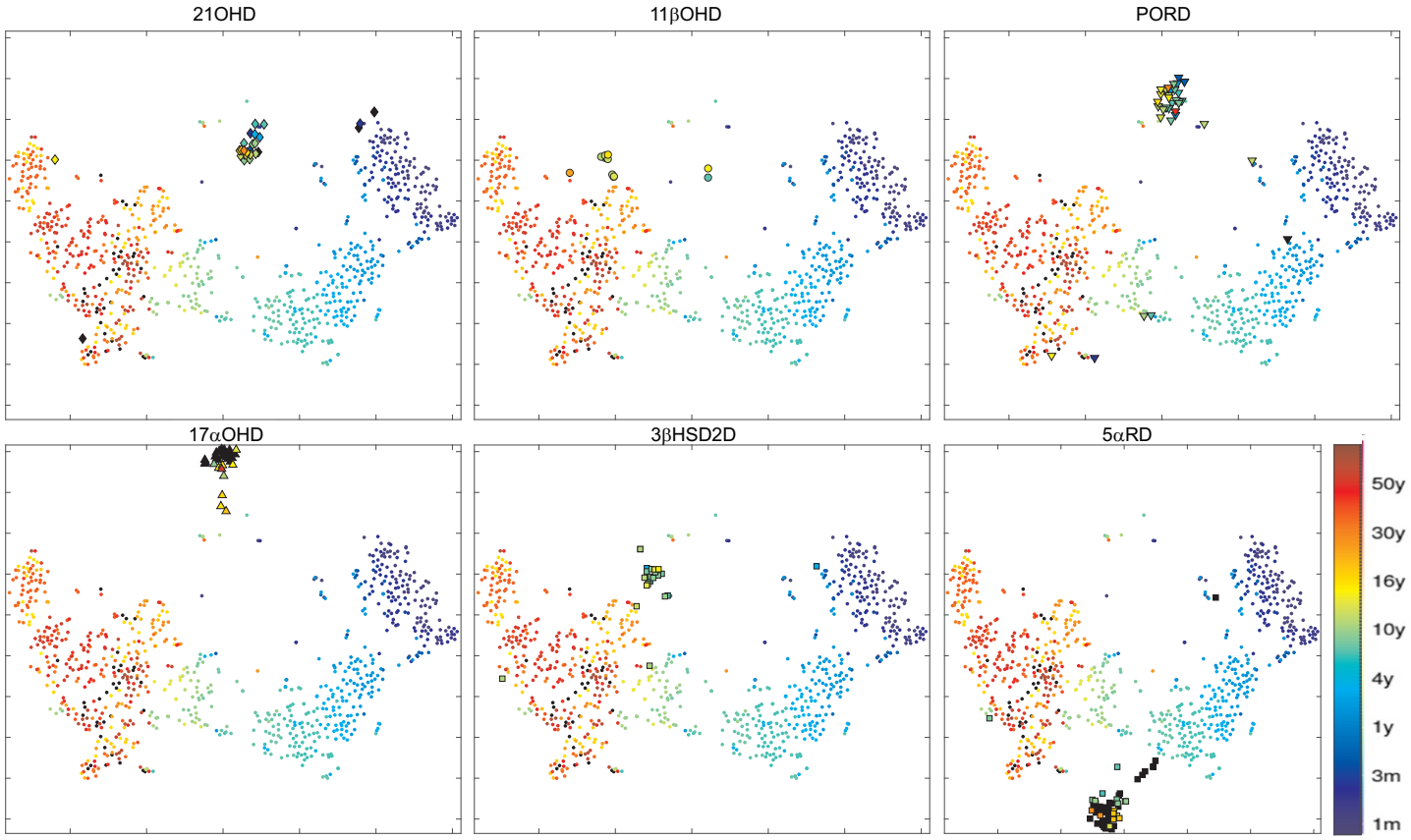


Figure 5.2: Unsupervised clustering using t-SNE

Each condition in isolation compared to healthy controls. Each individual patient participant is represented by a single symbol, healthy control participants are shows by dots. Colours relate to the age of the participant according to the scale to the bottom right of the figure. The position of the symbol/dot determined by the pairwise interaction at all measure urinary steroid metabolites

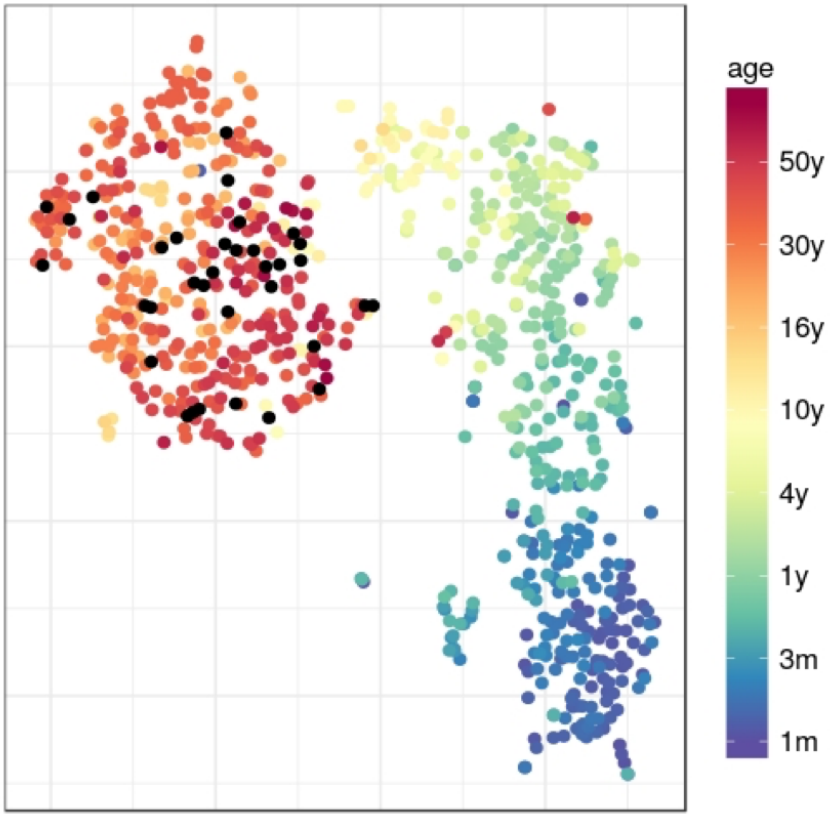


Figure 5.3: Unsupervised clustering using t-SNE: healthy controls in isolation

Each healthy control participant is shows by a single dot. Colours relate to the age of the participant according to the scale to the right of the figure. The position of the symbol/dot determined by the pairwise interaction at all measure urinary steroid metabolites.

5.6.3. Conclusions

We concluded from this unsupervised work that there was sufficient evidence to support proceeding with constructing a supervised learning classifier for the examined groups with no additional pre-processing of the features. Given the homogenous nature of metabolome findings in 11βOHD, as one raised metabolite is highly specific for this disease and other raised metabolites are common to other classes, we

speculated that we could improve performance of this class through feature reduction techniques and therefore continued inclusion of this class was justifiable. Using ratios of individual metabolites as features and using an angle-based dissimilarity measure appear effective strategies to ensure comparability between samples, reduce some inherent heterogeneity and limit the detrimental effect of missing metabolite values, while retaining the most important characteristics of the urine profiles. Based on these findings, we decided not to explicitly include age or sex information into the supervised classifier as clusters appear naturally in patient groups without this information and the normal observed age-related transition is not as apparent in the patient groups.

5.7. Supervised Learning Classifier Design

5.7.1. Angle Learning Vector Quantization

Our database included 24-hour urine collections as well as single spot urine sample collections, a wide age range for our participants and a significant amount of missing metabolite values. To overcome this heterogeneity and to make samples comparable, the constructed features used were ratios of metabolites rather than absolute metabolite values, as used in the unsupervised experiments. For 32 measured metabolites, that equates to 496 unique features. Ratios are z-score transformed for every training set in cross validation for the classification and the same transformation used in the respective test set.

In the context of both systematic and random missingness, this makes for a complex, high dimensional problem, with very low numbers of samples in some classes. Physiologically speaking, the relationship between metabolites is more important than

the absolute quantification of metabolites. With this in mind, and to handle the complexity of the classification task, we proposed a new variant of LVQ as an extension to GRLVQ and GMLVQ, termed Angle Learning Vector Quantization (ALVQ) (Bunte et al., 2016; Ghosh et al., 2022; Ghosh et al., 2020; Ghosh, 2021).

Angle Learning Vector Quantization (ALVQ) is a supervised, nearest prototype classification method which we applied to discriminate the different disease classes in our patient cohort from each other and from healthy controls. 496 pairwise features are used in combination, weighted through the use of adaptive relevances towards the more discriminatory features (automatically reducing the influence of less discriminatory features). Adaptive relevances and prototypes are determined through a supervised training procedure aiming to minimise a cost function which penalises incorrect classification. A prototype is created for each condition based on the training subset of data. These prototypes are artificially created sample vectors, derived through training. They represent the prototypical pattern of the most discriminatory features for each class, typical for differentiating a given class from the others. Features with missing metabolite values are ignored and sample vectors are constructed from the remaining available dimensions. This means that new data samples are compared to prototypes using only the observed dimensions. The model classifies test samples by comparing similarity of its steroid metabolome to each representative prototypic steroid metabolome, which it does by way of a parameterized adaptive angle-based distance measure. Sample vectors are classified on the unit hypersphere, meaning all vectors are of equivalent length, and a sample's class label is assigned according to which prototype they have the smallest distance to, in terms of geodesic distance on the hypersphere. The smaller the parameterized

angle, the closer the sample vector is to a given prototype vector. Classification is assigned on a "winner-takes-all" basis according to the closest prototype. See **Equation 5.2**- b identifies the relative similarity between sample vectors ($\vec{x}_{i,}$) and prototype vectors $\vec{w}_i^L$ parameterized by the matrix (Λ). Λ is a symmetric positive semi-definite matrix $\Omega^{\intercal}\Omega$ of rank M = 6, where Ω is a 496 x 6 matrix. J is the nearest correct prototype and K is the nearest incorrect prototype. For our example, we use the cosine b of the angle between vectors under the metric tensor Λ as our measure of similarity. This takes values from -1 (least similar) to +1 (most similar) which we transform using g_β to give us dissimilarity values from 0 to 1 such that the least similar values (-1) equate to the maximum dissimilarity (1) and the most similar values (+1) equate to minimum dissimilarity (0). The parameter β controls the translation of similarity to dissimilarity. Higher values of β imply sharper transition- increasing β results in a transformation which states that the closer the similarity value b is to +1 (most similar), the less influence it should have on refining the prototype vector during training (see **Figure 5.4**). This is because the prototype updating is driven by dissimilarity g_β. This approach means that the class samples which are further away from their respective prototype highly influence prototype position as this would be necessary to avoid a misclassification, whereas samples that are the most similar to the prototype don't need to influence the prototype position as unknown samples in that position are likely to fall into the correct Voronoi field and therefore be correctly identified regardless. **Figure 5.5** shows the effect of varying β on ALVQ performance. For this dataset it appeared that the classifier performance was robust to settings of β apart from extremely high values. This implies that the regime of prototype modification where the prototypes are updated only when the mismatch between prototypes and training examples is large (low similarity) is sufficient for constructing an effective classifier.

Note that the metric tensor Λ is normalised to trace 1. This is because multiplication of metric tensor by any positive constant does not have an effect on the similarity measure b.

Equation 5.2

$$d_i^L = g_\beta(b) \text{ with } \qquad b = b_\Lambda(\boldsymbol{x_i}, \boldsymbol{w}^L) = \frac{\boldsymbol{x}_i^\top \Lambda \boldsymbol{w}^L}{\|\boldsymbol{x}_i\|_\Lambda \|\boldsymbol{w}^L\|_\Lambda}$$

where

$$\|\boldsymbol{v}\|_\Lambda = \sqrt{\boldsymbol{v}^\top \Omega^\top \Omega \boldsymbol{v}} \ , \qquad \Lambda = \Omega^\top \Omega \ ,$$

$$g_\beta(b) = \frac{e^{-\beta(b-1)} - 1}{e^{(2\beta)} - 1}, \qquad \sum_i \Lambda_{ii} = 1 \text{ and } L \in \{J, K\} \ .$$

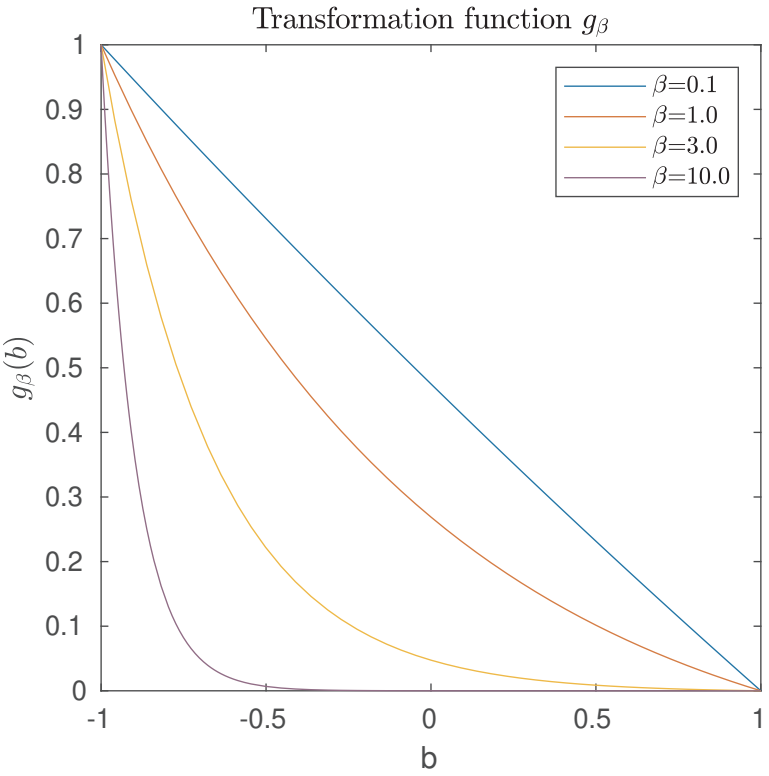


Figure 5.4: Transformation function β

X-axis depicts similarity [-1,1], y-axis depicts dissimilarity [0,1], coloured lines display the effect of varying β on the transformation of x to y.

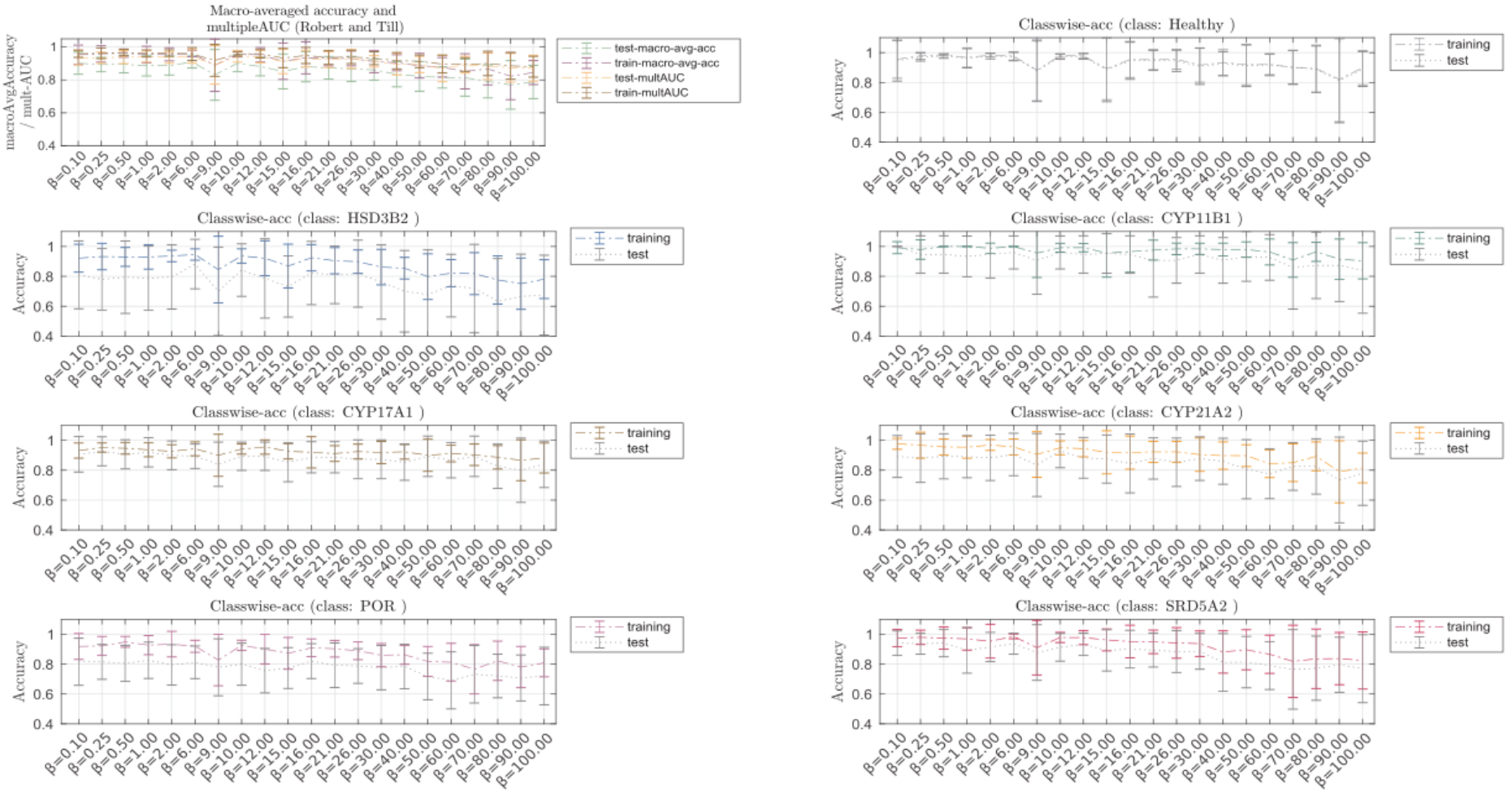


Figure 5.5: Effect of varying transformation factor β on performance of cost weighted deterministic ALVQ

The cost function shown in **Equation 5.1** is non-convex and can be optimised for example using stochastic gradient descent or conjugate gradient methods with derivatives for the parameters $\Phi \in \{\{w^j\}_{j=1}^k, \Omega\}$ shown in **Equation 5.3 to Equation 5.7:**

Equation 5.3

$$\frac{\partial E}{\partial \Phi} = \sum_{i=1}^{N} \frac{\partial f}{\partial \Phi} \frac{\partial \mu_i}{\partial \Phi} \qquad \text{with}$$

Equation 5.4

$$\frac{\partial \mu_i}{\partial \Omega} = \frac{2d^K}{(d^J + d^K)^2} \cdot \frac{\partial d^J}{\partial \Omega} - \frac{2d^J}{(d^J + d^K)^2} \cdot \frac{\partial d^K}{\partial \Omega}$$

Equation 5.5

$$\frac{\partial \mu_i}{\partial \boldsymbol{w}^J} = \frac{2d^K}{(d^J + d^K)^2} \cdot \frac{\partial d^J}{\partial \boldsymbol{w}^J} \qquad \text{and} \qquad \frac{\partial \mu_i}{\partial \boldsymbol{w}^K} = \frac{-2d^J}{(d^J + d^K)^2} \cdot \frac{\partial d^K}{\partial \boldsymbol{w}^K}$$

Equation 5.6

$$\frac{\partial d_i^L}{\partial \Phi} = \frac{\partial g_\beta(b_\Omega)}{\partial b_\Omega} \cdot \frac{\partial b_\Omega}{\partial \Phi}$$

Equation 5.7

$$\frac{\partial b_\Omega(\boldsymbol{x}_i, \boldsymbol{w}^L)}{\partial \boldsymbol{w}^L} = \frac{\boldsymbol{x}_i \Omega^\top \Omega \|\boldsymbol{w}^L\|_\Omega^2 - \boldsymbol{x}_i \Omega^\top \Omega \boldsymbol{w}^L \cdot \boldsymbol{w}^L \Omega^\top \Omega}{\|\boldsymbol{x}_i\|_\Omega \|\boldsymbol{w}^L\|_\Omega^3}$$

$$\frac{\partial b_\Omega(\boldsymbol{x_i}, \boldsymbol{w}^L)}{\partial \Omega_{mn}} = \frac{x_{i,n} \sum_j^D \Omega_{mj} w_j^L + w_n^L \sum_j^D \Omega_{mj} x_{i,j}}{\|\boldsymbol{x}_i\|_\Omega \|\boldsymbol{w}^L\|_\Omega}$$

$$- \boldsymbol{x}_i \Omega^\top \Omega \boldsymbol{w}^L \left[\frac{x_{i,n} \sum_j \Omega_{mj} x_{i,j}}{\|\boldsymbol{x}_i\|_\Omega^3 \|\boldsymbol{w}^L\|_\Omega} + \frac{w_n^L \sum_j \Omega_{mj} w_j^L}{\|\boldsymbol{x}_i\|_\Omega \|\boldsymbol{w}^L\|_\Omega^3} \right] ,$$

where $x_{i,n}$ denotes dimension *n* of vector x_i and *m* = 1, . . . ,*M* (recall that M denotes rank of the metric tensor).

It is expected that for a given condition, there will be a similar pattern of metabolites which are increased and decreased, therefore, there will be similar proportional characteristics between metabolites. This means ratio vectors of subjects of the same disease, constructed from the corresponding metabolite profiles, are expected to appear similar.

Using the cosine of the angle between vectors as a distance measure is ideally suited for our classification problem as it is the relationship between the metabolites that is important rather than the absolute quantification. An angle-based similarity measure has the benefit of being able to learn with vectors of variable dimension and naturally handles vectors of variable length because vector lengths are ignored when calculating angles. Angle based distance measures are more robust with respect to missing values compared to distance measures; decisions can be made using available dimensions, therefore ALVQ is less effected by random (MAR) or systematic (MNAR) missing values than LVQ versions operating with distances. This makes our method computationally very efficient because it is not required to implement any explicit techniques to handle missing values, such as imputation.

For more detail on the construction of ALVQ, please refer to the following references (Bunte et al., 2016; Ghosh et al., 2022; Ghosh et al., 2020; Ghosh, 2021).

5.7.2. Model Refinement Methods

5.7.2.1. Deterministic to Probabilistic Approaches

The first ALVQ model created was a deterministic model. This means that the output of the model would be a discrete label. Given the degree of overlap between the conditions, if patient misclassifications occurred, we hypothesised that the misclassified patients were likely to project near decision boundaries. This led to the proposal that from a medical standpoint, it would be valuable if we could numerically present the likelihood of a participant to belong to each of the classes, this could highlight samples near decision boundaries where there is less certainty of diagnosis. This would indicate to clinicians any cases where there may be some diagnostic uncertainty based solely on the urinary steroid metabolome analysis. This led to the creation of the Probabilistic variant of ALVQ (PALVQ) which provides a probability or percentage certainty of belonging to each of the investigated classes (Ghosh et al., 2022; Ghosh, 2021). Components of **Equation 5.2** are found as the numerator in **Equation 5.8**; classification is on the hypersphere and similarity of sample vector to prototype vectors is identified from the cosine of the angle between vectors giving values from -1 to +1. **Equation 5.8** defines the predicted class probability for a given sample ($\hat{p}(c|\vec{x}_i)$) as the transformed angle dissimilarity, parametrized by matrix Λ for a given class normalised by the sum of the identically handled angle dissimilarity calculated for all other classes. The transformation of this into a dissimilarity measure is different than for ALVQ as we want to transform this into a probability rather than a numerical distance, where the more similar the sample profiles are, the higher the probability (i.e., maximum similarity of +1 = maximum probability of +1, and vice versa

for minimum probability). This is shown in the second half of **Equation 5.8.** We use transformation factor θ, which provides the inverse of what we observe for the β transformation. The effect of altering the value of θ is shown in **Figure 5.6**. In preliminary studies (not shown) we found a robust performance with a θ value of 2 and so proceeded with this for all model iterations. **Equation 5.9** shows the divergence-based error or cost (H) of the predicted class ($\hat{p}(c|\vec{x}_i)$) compared to the known class label ($p(c|\vec{x}_i)$). Through optimisation procedures, learning aims to find the optimum model parameters to minimise this cost function. Minimisation takes place via numerical optimisation, such as stochastic gradient descent or in our case the Quasi Newton Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (Kroon, 2022). It is an iterative method for solving unconstrained nonlinear optimisation problems, based on descent directions where an approximation of the Hessian is gradually improved from gradient evaluations.

Training and Prediction processes of the proposed probabilistic ALVQ are shown in the below pseudocode:

Algorithm 1 Training of the probabilistic ALVQ

procedure PALVQ_TRAIN($\{(\vec{x}_i, y_i)\}_{i=1}^{N}, \Theta, M, \{\gamma_{cp}\}_{c,p=1,\ldots,C}$)
  initialize prototypes $\vec{w}_c$ with labels $c(\vec{w}_c) = c$ close to class means
  initialize entries $\Omega_{ml} \in [-1, 1]$ and divide by $\sqrt{\text{trace}(\Omega^\top \Omega)}$
  minimize costs $\sum_{c=1}^{C} \frac{1}{n_c} \left[\sum_{\vec{x}_i, s.t. y_i = c} \gamma_{c,\hat{y}_i} \hat{p}(c|\vec{x}_i) \ln \frac{\hat{p}(c|\vec{x}_i)}{p(c|\vec{x}_i)} \right]$ w.r.t. $\{\vec{w}_c, \Omega\}$
  normalize $\text{trace}(\Omega^\top \Omega) = 1$
  return LVQmodel $\{\vec{w}_c, c(\vec{w}_c), \Omega, \Theta\}$

Here $\{(\vec{x}_i, y_i)\}_{i=1}^{N}$ is the labeled training set, $\{\gamma_{cp}\}_{c,p=1,\ldots,C}$ the cost matrix to weight (mis-)classifications, M the rank of the metric, predicted labels $\hat{y}_i$ and Θ the parameter of the paramerized softmax $\hat{p}(c|\vec{x}_i)$ computed as:

$$\hat{p}(c|\vec{x}_i) = \frac{g_\Theta\left(\frac{\vec{x}_i \Lambda \vec{w}_c{}^\top}{\|\vec{x}_i\|_\Lambda \|\vec{w}_c\|_\Lambda}\right)}{\sum_j^C g_\Theta\left(\frac{\vec{x}_i \Lambda \vec{w}_j{}^\top}{\|\vec{x}_i\|_\Lambda \|\vec{w}_j\|_\Lambda}\right)} \quad \text{with} \quad g_\Theta(b) = \frac{e^{\Theta(b+1)} - 1}{e^{(2\Theta)} - 1} \text{ and } \Lambda = \Omega^\top \Omega \ .$$

Algorithm 2 Classification with the probabilistic ALVQ

procedure PALVQ_CLASSIFY($\{\vec{x}_i\}_{i=1}^{N}$, LVQmodel$\{\vec{w}_c, c(\vec{w}_c), \Omega, \Theta\}$)
  for $i = 1 \ldots N$ **do**
    for $c = 1 \ldots C$ **do**
      compute parameterized softmax $\hat{p}(c|\vec{x}_i)$
  return predicted labels $\hat{y}_i = \arg\max_c \hat{p}(c|\vec{x}_i)$ and probabilities $\hat{p}(c|\vec{x}_i)$

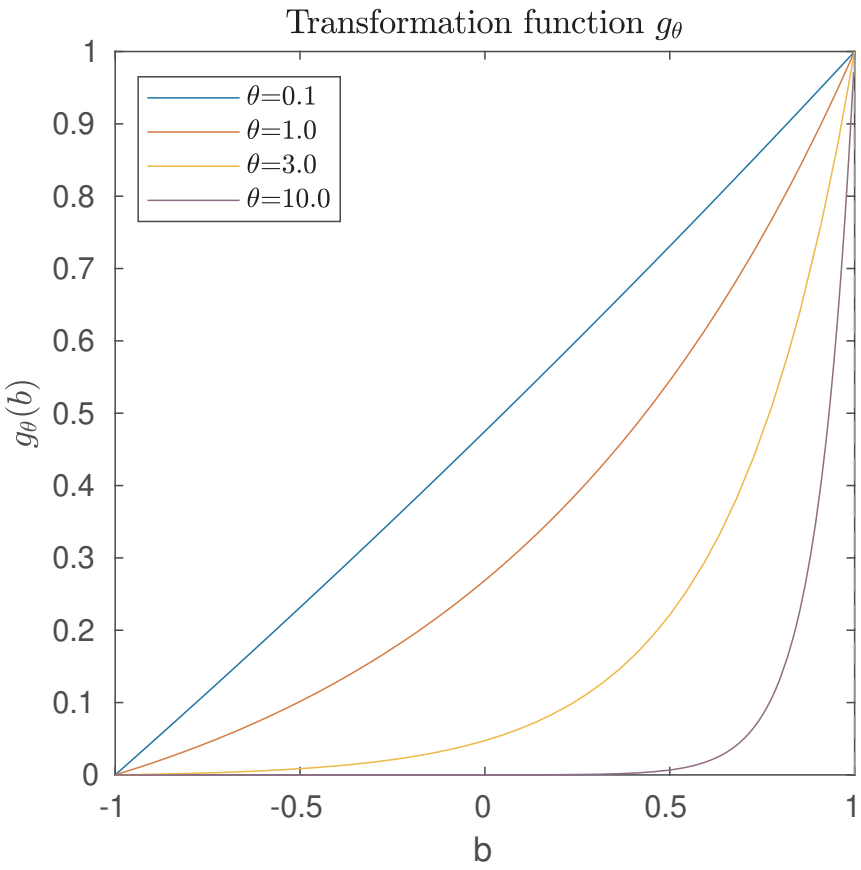


Figure 5.6: Transformation function θ

X-axis depicts similarity [-1,1], y-axis depicts dissimilarity [0,1], coloured lines display the effect of varying θ on the transformation of x to y.

Equation 5.8

$$\hat{p}(c|\vec{x}_i) = \frac{g_\theta\left(\frac{\vec{x}_i \Lambda \vec{w}^{c\intercal}}{\|\vec{x}_i\|_\Lambda \|\vec{w}^c\|_\Lambda}\right)}{\sum_j^C g_\theta\left(\frac{\vec{x}_i \Lambda \vec{w}^{j\intercal}}{\|\vec{x}_i\|_\Lambda \|\vec{w}^j\|_\Lambda}\right)} \; with \; g_\theta(b) = \frac{e^{\theta(b+1)} - 1}{e^{(2\theta)} - 1}$$

Equation 5.9

$$H = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} \hat{p}(c|\vec{x}_i) In \frac{\hat{p}(c|\vec{x}_i)}{p(c|\vec{x}_i)}$$

5.7.2.2. Combining the power of multiple learners

Ensembling is a well-known strategy to avoid overfitting and improve generalisation performance (Ying, 2019). Ensembling improves performance by combining variations in independently trained models (Breiman, 1996; Dietterich, 2002). The cost of this method is with computational power and memory, and with loss of interpretability. Random Forest, for example, is a robust method often used for classification which uses an ensemble of decision trees. To further develop from our "individual model" approach, we used two methods to construct models which benefit from combining multiple learners. First, we created an "Ensembled Model". Second, we created a "Geodesic Averaged Model". The ensembled model approach we pursued created 100 models per fold of cross validation with variability in classifiers observed due to random initialisation of the relevance matrix and of prototypes around class mean. For each fold of cross validation there are 100 results for each performance metric. The model determines the winning class by looking for the majority vote: what is the mode for each subject, or what is the most frequent class a given subject is classified as in the 100 models. In the geodesic averaged approach, 100 models are created but because we create only one prototype per class, we can calculate the geodesic average position of all 100 prototypes on the hypersphere for each class. We also calculated the geodesic average of 100 metric tensors. Construction of geodesic averages was necessary because both the prototypes and the metric tensors live in curved spaces (hypersphere and manifold of positive symmetric semidefinite matrices of a given rank *M,* respectively) and hence classical averaging does not guarantee that such averages will stay in the desired spaces. Full details can be found in (Ghosh et al., 2022). In this way, the 100 independently trained models are combined or

"averaged" to a single model per fold of cross validation. The geodesic averaged model is fed each fold of cross validation and so, in our case, we get 5 results for each performance metric, 1 result for each fold of cross validation. The mean and standard deviation are calculated from the results over the 5 folds of cross validation for all the performance metrics presented for all model variants. Model averaging is done intra-fold, meaning testing data remains unseen by the model so is not included in the determination of each prototype. Both are powerful techniques, but the averaged model has the additional benefit that it is a computationally simpler algorithm, and it retains full interpretability whereas the internal working logic of the ensembled model is more difficult to extract. Given the importance of transparency for wider acceptance of machine learning models, we decided to proceed with the geodesic averaged model rather than accept compromising interpretability. For more details of the ensembled and averaged PALVQ models, please see (Ghosh et al., 2022; Ghosh, 2021).

5.7.2.3. Introducing domain knowledge for cost weighting

In order to refine the model further we added in biological information in the form of a cost weighting matrix. Not all misclassifications are considered equal and to prioritise the correct classification of certain classes over others we introduced a domain knowledge driven cost weighting scheme. We provided the model with a matrix containing a hierarchical ranking of mistakes it could make, and during model training this informed the penalty given to settings which incur these misclassifications. The first medically driven priority is not to misclassify any patients as healthy controls, so this misclassification incurred the greatest penalty. The next priority would be not to misclassify patients with CAH as having 5αRD. The next priority would be not to

misclassify healthy controls as patients. The lowest priority, and therefore lowest penalty, was assigned to correct discrimination of the different causes of CAH. The greatest positive weighted priority was the diagonal or correct class wise determination for patient groups. The 21OHD class was given the most positive weight as this is the most common disorder within the patient classes and therefore the most important to correctly diagnose. We have compared results from oversampling with geodesic SMOTE to this domain knowledge led method of cost weighting.

5.7.2.4. Feature Reduction

We hypothesised that there were likely to be many common features between the different disease classes. Relatively proximal enzymatic blocks in the steroidogenesis pathway could be more difficult to differentiate from more distal enzymatic blocks as there is likely to be a high degree of overlap for many precursor-to-product metabolite ratios and fewer unique discriminatory ratios. Therefore, the next approach for model refinement that we used was to try to simplify the model through feature reduction. We hypothesized that this approach would improve class wise accuracy by removing the less relevant metabolite ratios which may be commonly high or low in more than one condition.

In order to reduce the number of features for the averaged cost weighted probabilistic ALVQ model, the 100 models informing each fold of cross validation were averaged resulting in 1 geodesic average model which was applied to the 5 rounds of training data. Any matrix LVQ model allows the analysis of its decision making based on the classification terms. For our probabilistic ALVQ model, a classification term is the

product of a sample vector dimension $x_{i,d1}$, the relevance matrix element $\Lambda_{d1,d2}$, and a prototype dimension w_{d2}^{c}, together with additional factors or transformations dependent on the dissimilarity measure. A sample x_i is classified as class *c* if the sum of the classification terms (T^{ic}) including prototype w^c is larger than for any other prototype (see **Equation 5.10**).

Equation 5.10

$$\hat{p}(c|\boldsymbol{x}_i) = \frac{g_\Theta\left(\sum_{F1=1,F2=1}^{D} T_{F1,F2}^{ic}\right)}{\sum_{j}^{C} g_\Theta\left(\sum_{F1=1,F2=1}^{D} T_{F1,F2}^{ic}\right)} \quad \text{with } T_{F1,F2}^{ic} = \frac{x_{i,F1}\Lambda_{F1,F2}w_{F2}^{c}}{\|\boldsymbol{x}_i\|_\Lambda \|\boldsymbol{w}^c\|_\Lambda} \ .$$

Relevances are weights for the decision collected in the positive semi-definite matrix Lambda (Λ). Because Λ is a metric tensor, feature-wise relevances can be directly extracted from the diagonal of Λ and sorted in descending order. The matrix of classification terms T^{ic} contains positive or negative entries indicating the correlation of x_i with the prototype w_c induced by the metric tensor Λ. Since the classification decision is based on the biggest sum over all arg max ($\sum_{F1,F2} T_{F1,F2}^{ic}$) the terms can be sorted. The cumulative sum of the sorted relevances were examined, relevance threshold was reduced from 100% to top 70% (sum of relevances = 0.7), then further reduced in increments of 5% to top 50% (sum of relevances = 0.5) and features extracted up to that threshold. This equated to examining 496 features (100% relevance) compared to 152 features (70% relevance threshold), 126 features (65% relevance threshold), 103 features (60% relevance threshold), 83 features (55% relevance threshold) and 67 features (50% relevance threshold). These indices were then used for retraining the “reduced” model. Effect on sensitivity for diagnosis of any

disease, sensitivity for diagnosis of correct disease (also known as macro-averaged accuracy of disease, MAA(disease)) and specificity were examined. Results are shown for 70% relevance threshold and 55% relevance threshold for the individual, ensembled and geodesic averaged approaches.

5.8. Results

5.8.1. Probabilistic Models with Full Features

All results are displayed as a mean and standard deviation (std) over 5 folds of cross validation. Training performance was consistently very high for all performance metrics for all iterations of ALVQ. A summary of the test performance for the different probabilistic ALVQ models using all 496 features is shown in **Table 5.1.** This shows performance for the individual model approach, taking the mean and standard deviation of the performance from 100 individual models, for the ensembled model approach using 100 models each applied to 5 folds of cross validation, and for the geodesic averaged model with 100 models averaged and applied to five folds of cross validation. The top three rows of the table show the results for the three model approaches using an oversampling approach in the form of geodesic SMOTE, oversampling the minority classes to the majority class size. The bottom three rows of the table show the results for the three model approaches using domain knowledge informed cost weighting as described above. Despite there being an overlap between the standard deviations of the performances obtained by the different approaches, the fact that the cost weighted approach takes into account a clinical perspective gives an additional advantage to this approach. Both group sensitivities to diagnose a patient

with any disease and to diagnose a patient with the correct disease improved with cost weighting above oversampling, but the specificity was negatively impacted. Maximum sensitivity was seen with the ensembled and geodesic averaged models with cost weighting (mean 98.9% (std 0.02) sensitivity (any disease), mean 92.7% (std 0.04) sensitivity (correct disease)). Despite priority being given to the correct assignment of 21OHD, class wise sensitivity of this class was optimum at (mean 92% (std 0.18)) with the ensemble approach with geodesic oversampling. 11βOHD was consistently well identified by all models, but particularly by the ensembled and geodesic averaged models (mean 100% (std 0)). PORD diagnoses were improved with cost weighting as opposed to geodesic SMOTE oversampling, and further improved by combining multiple learners through geodesic average and ensemble approaches (mean 88.9% (std 0.12)). 17αOHD diagnoses were optimum with the ensemble approach and geodesic oversampling (mean 96.7% (std 0.07)) but ensemble and geodesic average models with cost weighting performed well (mean 93.3% (std 0.09). The different strategies shown here did not significantly improve diagnostic performance of 3βHSD2D with mean class sensitivity varying from 85.6% (std 0.23) to 87.8% (std 0.17). 5αRD was consistently well diagnosed by all model types but identified with 100% class sensitivity (std 0) for the ensembled and geodesic averaged models with cost weighting.

Table 5.2 shows the comparative training and test performance (mean and standard deviation over 5 runs of cross validation) for all three model approaches (individual, ensembled and geodesic averaged) with cost weighting, using all 496 features. Training performance was high, with 100% sensitivity to diagnose any disease for all 3 models, sensitivity to diagnose the correct disease >98%, specificity >90% and class

wise sensitivity varying from 94.2% (17αOHD) to 100% (21OHD, 11βOHD, 3βHSD2D and 5αRD). Test performance was reduced compared to training for 3βHSD2D and 21OHD, showing the lowest class sensitivities despite having 100% training performance for both the ensembled and geodesic averaged model (mean 86% (std 0.22) and mean 88% (std 0.11) respectively). Performance for ensembled approach and geodesic averaged approach was identical apart from specificity where the averaged approach was superior (mean 92.5% (std 0.04), compared to mean 90% (std 0.03)).

Table 5.3 shows the detailed test results for the cost weighted geodesic averaged model with full features in the form of a confusion matrix. All performance metrics are shown as an average and standard deviation over 5 folds of cross validation for test data only. 11βOHD and 5αRD were the best performing classes (mean 100% (std 0)). POR and 3βHSD2D classes showed some misclassifications as healthy despite the cost weighting but no other patient classes had healthy misclassifications. There were patients with 17αOHD who were misdiagnosed with PORD. Along with the healthy misclassifications, POR patients were also misdiagnosed with 21OHD, 17αOHD and 5αRD. The 21OHD class showed misclassifications with PORD, 3βHSD2D and 5αRD. In addition to healthy, 3βHSD2D patients were mistaken for 5αRD. The healthy class showed a spread of misdiagnoses across the patient groups, the most misdiagnoses observed with the 5αRD group.

Table 5.1: Summary of Probabilistic ALVQ model variant performances, full features, test results only

Oversampling is with geodesic SMOTE, oversampling the minority classes to the majority class size. Cost weighting is according to domain knowledge. Results shown are the mean and standard deviation over 5 folds of cross validation. Individual model results presented are calculated from the average of 100 individual models over 5 folds of cross validation.

	Sensitivity (Any disease)		Sensitivity (Correct disease)		Specificity (Healthy class)		21OHD		11βOHD		PORD		17αOHD		3βHSD2D		5αRD	
	Mean	std	Mean	std	Mean	std	Mean	std	Mean	std	Mean	std	Mean	std	Mean	std	Mean	std
Individual, 100 models, oversampling	94.7%	0.02	88.9%	0.04	96.9%	0.02	81%	0.2	99.6%	0.01	86.4%	0.15	83.7%	0.05	85.6%	0.23	97.1%	0.05
Ensemble, 100 models, oversampling	95.5%	0.03	92.1%	0.05	98.8%	0.01	92%	0.18	100%	0	83.6%	0.15	96.7%	0.07	86%	0.22	94.4%	0.08
Average, 100 models, oversampling	95.5%	0.02	90.4%	0.03	97.4%	0.02	88%	0.18	100%	0	86.8%	0.15	83.3%	0	86%	0.22	98.2%	0.04
Individual 100 models, Cost-weighted	98.3%	0.02	92.5%	0.05	88.3%	0.04	90.1%	0.13	99.7%	0.05	86.4%	0.11	91.6%	0.09	87.8%	0.17	99.5%	0.02
Ensemble, 100 models, Cost-weighted	98.9%	0.02	92.7%	0.04	90.0%	0.03	88%	0.11	100.0%	0.00	88.9%	0.12	93.3%	0.09	86.0%	0.22	100%	0.00
Average, 100 models, Cost-weighted	98.9%	0.02	92.7%	0.04	92.5%	0.04	88%	0.11	100.0%	0.00	88.9%	0.12	93.3%	0.09	86.0%	0.22	100%	0.00

Table 5.2: Training and test performance for cost weighted probabilistic ALVQ models, full features

		Sensitivity (Any disease)		Sensitivity (Correct disease)		Specificity (Healthy class)		21OHD		11βOHD		PORD		17αOHD		3βHSD2D		5αRD	
		Mean	std	Mean	std	Mean	std	Mean	std	Mean	std	Mean	std	Mean	std	Mean	std	Mean	std
Individual	train	100.0%	0.00	98.5%	0.03	90.9%	0.03	99.9%	0.02	99.8%	0.04	97.6%	0.02	94.2%	0.03	99.8%	0.03	99.9%	0.03
Individual	test	98.3%	0.02	92.5%	0.05	88.3%	0.04	90.1%	0.13	99.7%	0.05	86.4%	0.11	91.6%	0.09	87.8%	0.17	99.5%	0.02
Ensemble	train	100.0%	0.00	98.7%	0.00	92.2%	0.03	100.0%	0.00	100.0%	0.00	97.9%	0.02	94.2%	0.02	100.0%	0.00	100.0%	0.00
Ensemble	test	98.9%	0.02	92.7%	0.04	90.0%	0.03	88.0%	0.11	100.0%	0.00	88.9%	0.12	93.3%	0.09	86.0%	0.22	100.0%	0.00
Average	train	100.0%	0.00	98.8%	0.01	94.9%	0.01	100.0%	0.00	100.0%	0.00	97.9%	0.02	95.0%	0.02	100.0%	0.00	100.0%	0.00
Average	test	98.9%	0.02	92.7%	0.04	92.5%	0.04	88.0%	0.11	100.0%	0.00	88.9%	0.12	93.3%	0.09	86.0%	0.22	100.0%	0.00

Table 5.3: Confusion matrix- Cost Weighted, Probabilistic ALVQ, Geodesic averaged model, full features

Average test performance over 5 folds of cross validation. Diagonal (shaded grey) are correctly identified per class.

	Predicted Healthy	Predicted 21OHD	Predicted 11βOHD	Predicted PORD	Predicted 17αOHD	Predicted 3βHSD2D	Predicted 5αRD	Total per class
true Healthy	153.4	1.2	0.2	3.0	0.8	1.8	5.4	165.8
true 21OHD	0.0	4.8	0.0	0.2	0.0	0.2	0.2	5.4
true 11βOHD	0.0	0.0	2.4	0.0	0.0	0.0	0.0	2.4
true PORD	0.2	0.2	0.0	6.4	0.2	0.0	0.2	7.2
true 17αOHD	0.0	0.0	0.0	0.4	5.6	0.0	0.0	6
true 3βHSD2D	0.2	0.0	0.0	0.0	0.0	3.8	0.4	4.4
true 5αRD	0.0	0.0	0.0	0.0	0.0	0.0	10.2	10.2

5.8.2. Probabilistic ALVQ Models, with Reduced Features

Figure 5.7 shows how the sensitivity to identify a patient with any disease (sens), sensitivity to identify a patient with the correct disease (MAA(disease)) and specificity (spec) varied with the number of features used for the geodesic averaged model. This figure shows that the two markers of sensitivity remained relatively static throughout training, meaning feature reduction did not have a negative impact on training performance, whereas there was a detrimental effect on specificity with feature reduction. From the testing data, we can see that feature reduction leads to less overfitting of the data. Sensitivity to diagnose a patient with any disease increased with a relevance threshold set to include only features in the top 60% of relevance (103 ratios). Sensitivity to diagnose a patient with the correct disease increased to a peak at a relevance threshold of 55% (top 83 ratios in order of relevance). The rate of decline of the specificity increased with a relevance threshold set to smaller than the top 55%. A sharp decline was noted between 55% (top 83 features) and 50% (top 67 features). Only 152 of the 496 ratios constituted the top 70% of the relevances, meaning the majority of features did not contribute significantly to the classification task.

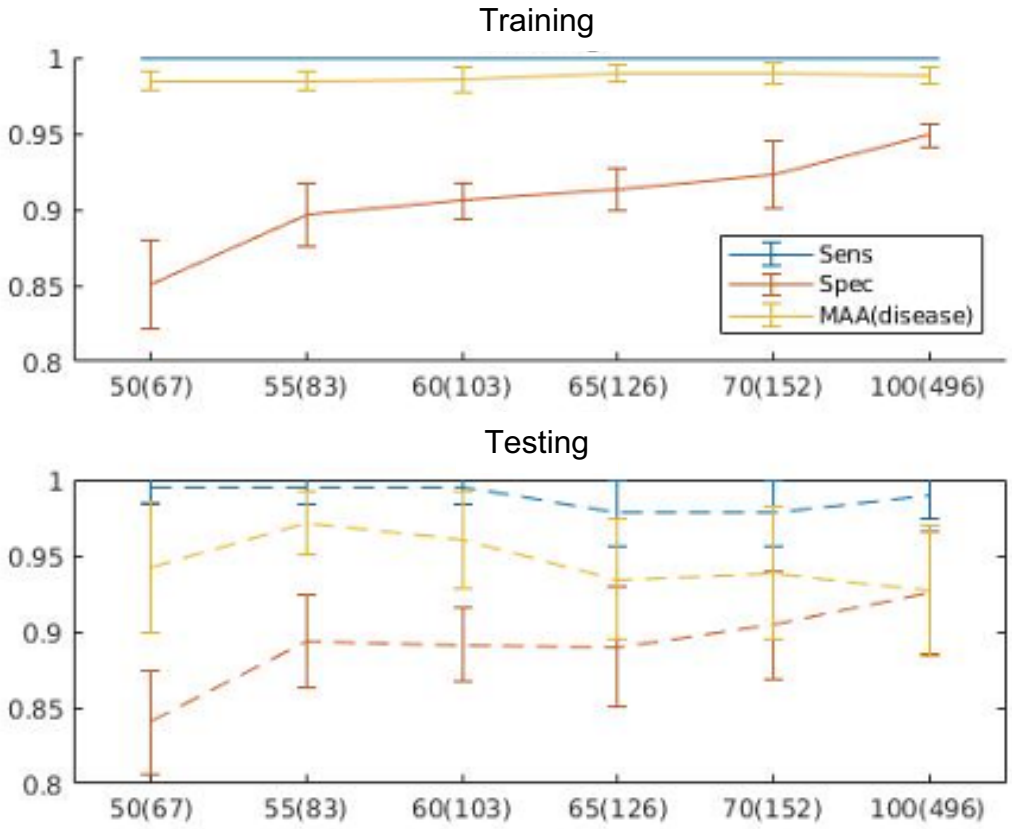


Figure 5.7: Variability in training and testing performance with feature reduction

Results shown are an average of 5 folds of cross validation for the geodesic averaged model, with 100 models averaged. Whiskers represent the standard deviation. Sens = sensitivity to diagnose a patient with any condition, Spec = specificity, MAA(disease) = "macro-averaged accuracy" or Sensitivity to diagnose a patient with the correct disease. X-axis shows relevance threshold with corresponding number of features in brackets. Y – axis denotes percentage, 1 = 100%. The top panel shows average results from training, the bottom panel shows results from testing.

In **Table 5.4**, training and test performance of the three model strategies, using cost weighting and a relevance threshold of 70% (top 152 ratios, in descending order of relevance) is shown. If we compare this table to **Table 5.2**, there was a reduction in specificity performance for all three model types. There was a reduction in performance of the 17αOHD class, with class sensitivity reducing from 93.3% (std 0.09) to 90% (std 0.15) for the geodesic averaged model. Classification of the 3βHSD2D class remained a challenge, with test sensitivity of 86% (std 0.22) for the ensembled model and the geodesic averaged model, which was unchanged. Other class sensitives were improved or remain high. Of note, 21OHD class test performance increased from 88% (std 0.11) to 92% (std 0.11). and POR class test performance increased from 88.9% (std 0.12) to 92% (std 0.11). **Table 5.5** shows more detailed performance results for the ensembled model in the form of a confusion matrix and **Table 5.6** shows more detailed performance results for the geodesic averaged model in the form of a confusion matrix. In **Table 5.7**, training and test performance of the three model strategies, using cost weighting and a relevance threshold of 55% (top 83 features in descending order of relevance) is shown. Comparing this table to **Table 5.2** and **Table 5.4**, specificity was further reduced in test sets to 86% (std 0.06), 88% (std 0.03) and 89% (std 0.03) for the individual, ensembled and geodesic averaged models respectively. Very little overfitting was observed with very similar or even identical values noted for training and test performance. Sensitivity to diagnose a patient with any disease was similarly near perfect as seen for the 70% reduced features model. Sensitivity to diagnose a patient with the correct disease, however, was improved at 97% (std 0.02) for both the ensembled and geodesic averaged models. Class wise performance was improved with 17αOHD showing the lowest class sensitivity at 93% (std 0.15) for the ensembled

and geodesic averaged models. Notably, performance of the 3βHSD2D class was improved with a class sensitivity of 96% (std 0.09) seen for both the ensembled and geodesic averaged models. The geodesic averaged model showed superior performance for specificity (89% (std 0.03)), and for the 5αRD class (100% (std 0)). Confusion matrices for the ensembled model and the geodesic averaged model, with cost weighting and top 55% relevance threshold are shown in **Table 5.8**. and **Table 5.9** respectively. There was a persistent misclassification of a 3βHSD2D patient for a healthy control in all models, but all other false negative cases (patient mistaken for healthy) were no longer seen with the cost weighted geodesic averaged probabilistic ALVQ model with features reduced to the top 55% cumulative relevance. There remained a patient with 21OHD misclassified for 5αRD, a patient with PORD misclassified for 21OHD, and patients with 17αOHD misclassified for PORD and 5αRD. There is a detailed review of patients frequently misclassified further in this chapter.

In the absence of a prospective validation cohort or a holdout set, a synthetic database was constructed to approximate some of the characteristics of the real-life dataset, so we could perform a controlled experiment where there is a known ground truth, using the ideal conditions of having training, validation, and generalisation datasets. Using this synthetic dataset, performance was compared for the various ALVQ models presented here, along with random forest, Linear Discriminant Analysis and kNN. These results are not presented here but can be viewed at (Ghosh et al., 2022; Ghosh, 2021).

Table 5.4: Cost weighted probabilistic ALVQ models, reduced features- top 70% relevance, training, and test performance

		Sensitivity (Any disease)		Sensitivity (Correct disease)		Specificity (Healthy class)		21OHD		11βOHD		PORD		17αOHD		3βHSD2D		5αRD	
		Mean	std	Mean	std	Mean	std	Mean	std	Mean	std	Mean	std	Mean	std	Mean	std	Mean	std
individual	train	100.0%	0.00	98.4%	0.04	90.0%	0.04	99.7%	0.03	99.6%	0.05	97.4%	0.04	94.6%	0.04	99.5%	0.04	99.6%	0.04
individual	test	97.9%	0.02	91.6%	0.05	87.2%	0.05	90.9%	0.13	99.8%	0.04	90.7%	0.10	86.4%	0.13	83.6%	0.20	98.4%	0.06
Ensemble	train	100.0%	0.00	99.1%	0.01	91.5%	0.01	100.0%	0.00	100.0%	0.00	97.9%	0.02	96.7%	0.02	100.0%	0.00	100.0%	0.00
Ensemble	test	98.3%	0.03	92.8%	0.03	88.9%	0.04	92.0%	0.11	100.0%	0.00	92.1%	0.11	86.7%	0.14	86.0%	0.22	100.0%	0.00
Average	train	100.0%	0.00	99.1%	0.01	92.3%	0.01	100.0%	0.00	100.0%	0.00	97.9%	0.02	96.7%	0.02	100.0%	0.00	100.0%	0.00
Average	test	98.9%	0.02	93.4%	0.04	89.9%	0.04	92.0%	0.11	100.0%	0.00	92.1%	0.11	90.0%	0.15	86.0%	0.22	100.0%	0.00

Table 5.5: Cost weighted probabilistic ALVQ, reduced features 70%, ensemble model, test performance.

Average performance over 5 folds of cross validation. Diagonal (shaded grey) are correctly identified individuals in each class.

	Predicted Healthy	Predicted 21OHD	Predicted 11βOHD	Predicted PORD	Predicted 17αOHD	Predicted 3βHSD2D	Predicted 5αRD	Total per class
true Healthy	147.4	0.4	0.4	5.6	1.2	3.6	7.2	165.8
true 21OHD	0	5	0	0.2	0	0	0.2	5.4
true 11βOHD	0	0	2.4	0	0	0	0	2.4
true PORD	0.2	0.4	0	6.6	0	0	0	7.2
true 17αOHD	0.2	0.2	0	0.4	5.2	0	0	6
true 3βHSD2D	0.2	0	0	0	0	3.8	0.4	4.4
true 5αRD	0	0	0	0	0	0	10.2	10.2

Table 5.6: Cost weighted probabilistic ALVQ, reduced features >70% relevance, geodesic averaged model, test performance

Average performance over 5 folds of cross validation. Diagonal (shaded grey) are correctly identified individuals in each class.

	Predicted Healthy	Predicted 21OHD	Predicted 11βOHD	Predicted PORD	Predicted 17αOHD	Predicted 3βHSD2D	Predicted 5αRD	Total per class
true Healthy	149	0.6	0.4	4.2	1.2	3.2	7.2	165.8
true 21OHD	0	5	0	0.2	0	0	0.2	5.4
true 11βOHD	0	0	2.4	0	0	0	0	2.4
true PORD	0.2	0.4	0	6.6	0	0	0	7.2
true 17αOHD	0	0.2	0	0.4	5.4	0	0	6
true 3βHSD2D	0.2	0	0	0	0	3.8	0.4	4.4
true 5αRD	0	0	0	0	0	0	10.2	10.2

Table 5.7: Cost weighted probabilistic models, reduced features top 55% relevance, training, and test performance

		Sensitivity (Any disease)		Sensitivity (Correct disease)		Specificity (Healthy class)		21OHD		11βOHD		PORD		17αOHD		3βHSD2D		5αRD	
		Mean	std	Mean	std	Mean	std	Mean	std	Mean	std	Mean	std	Mean	std	Mean	std	Mean	std
Individual	train	99.9%	0.01	97.3%	0.07	86.7%	0.04	99.1%	0.06	99.4%	0.04	95.7%	0.09	92.3%	0.11	98.9%	0.07	98.6%	0.10
Individual	test	99.0%	0.02	95.2%	0.08	85.5%	0.06	94.1%	0.12	99.4%	0.05	95.0%	0.11	91.7%	0.17	94.3%	0.12	96.8%	0.11
Ensemble	train	100.0%	0.00	98.4%	0.01	88.1%	0.03	100.0%	0.00	100.0%	0.00	97.2%	0.02	93.3%	0.04	100.0%	0.00	100.0%	0.00
Ensemble	test	98.9%	0.02	96.8%	0.02	87.8%	0.04	96.0%	0.09	100.0%	0.00	97.5%	0.06	93.3%	0.15	96.0%	0.09	98.0%	0.04
Average	train	100.0%	0.00	98.4%	0.01	89.7%	0.02	100.0%	0.00	100.0%	0.00	97.2%	0.02	93.3%	0.04	100.0%	0.00	100.0%	0.00
Average	test	99.5%	0.01	97.1%	0.02	89.4%	0.03	96.0%	0.09	100.0%	0.00	97.5%	0.06	93.3%	0.15	96.0%	0.09	100.0%	0.00

Table 5.8: Cost weighted probabilistic ALVQ, reduced features top 55% relevance, ensembled model, test performance.

Average performance over 5 folds of cross validation. Diagonal (shaded grey) are correctly identified individuals in each class.

	Predicted Healthy	Predicted 21OHD	Predicted 11βOHD	Predicted PORD	Predicted 17αOHD	Predicted 3βHSD2D	Predicted 5αRD	Total per class
true Healthy	145.6	0.4	0.6	5.8	3.4	6.2	3.8	165.8
true 21OHD	0	5.2	0	0	0	0	0.2	5.4
true 11βOHD	0	0	2.4	0	0	0	0	2.4
true PORD	0	0.2	0	7	0	0	0	7.2
true 17αOHD	0	0	0	0.2	5.6	0	0.2	6
true 3βHSD2D	0.2	0	0	0	0	4.2	0	4.4
true 5αRD	0.2	0	0	0	0	0	10	10.2

Table 5.9: Cost weighted probabilistic ALVQ, reduced features top 55% relevance, geodesic averaged model, test performance.

Average performance over 5 folds of cross validation. Diagonal (shaded grey) are correctly identified individuals in each class.

	Predicted Healthy	Predicted 21OHD	Predicted 11βOHD	Predicted PORD	Predicted 17αOHD	Predicted 3βHSD2D	Predicted 5αRD	Total per class
true Healthy	148.2	0.4	0.4	4.4	3	5	4.4	165.8
true 21OHD	0	5.2	0	0	0	0	0.2	5.4
true 11βOHD	0	0	2.4	0	0	0	0	2.4
true PORD	0	0.2	0	7	0	0	0	7.2
true 17αOHD	0	0	0	0.2	5.6	0	0.2	6
true 3βHSD2D	0.2	0	0	0	0	4.2	0	4.4
true 5αRD	0	0	0	0	0	0	10.2	10.2

5.8.3. Analysis of the Misclassified Participants

As each test fold contains a unique subset of the cohort, looking at each test fold result in isolation shows how the whole cohort is classified by the algorithm. For the cost weighted probabilistic ALVQ geodesic averaged model using all features**,** examining across all 5 test folds, we saw 62 healthy control misclassifications. The most frequent age group to be misclassified was the over 16 years age group (n=26 across 5 test folds). The most frequent misclassification overall for all age groups was 5αRD (n = 27, 9 misclassifications in each of the 3 age groups), followed by PORD (n=15, <6m n = 2, 6m – 16yr n = 9, >16yr n = 4). For 49 of the 62 healthy controls misclassified as patients by the most likely diagnosis, the second most likely diagnosis was correct as

healthy. Of the 13 healthy controls who remained misclassified as patients, 6 were classified at 21OHD, 1 classified as 11βOHD, 2 classified as PORD, 2 classified as 3βHSD2D and 2 classified as 5αRD. These results are summarised in **Figure 5.8**. Looking at the patients who were misclassified by this model, there were 12 misclassifications. 3 patients with 21OHD are misclassified, 1 is assigned PORD, 1 is assigned 3βHSD2D, and 1 assigned 5αRD. All 3 were correctly identified by the second most likely diagnosis. All 3 were female, 2 were young babies and 1 was 18 months old. 4 patients with PORD were misclassified, 2 females of 12 and 17 years, and 2 male, one baby and one child of 11 years. 1 was assigned healthy, 1 was assigned 21OHD, 1 assigned 17αOHD and 1 assigned 5αRD. The patient with PORD who was assigned as having 17αOHD was correctly classified by the second most likely diagnosis. 2 patients with 17αOHD are misclassified, both are young adult males of 20 and 23 years of age, both assigned PORD, and both were correctly identified as having 17αOHD by the second most likely diagnosis. Three patients with 3βHSD2D are misclassified, one female and two males, ages 9 years, 10 years, and 10 years. One was assigned healthy and two were assigned 5αRD. All three patients were correctly assigned as having 3βHSD2D by the second most likely diagnosis. Probabilities for the most likely diagnosis ranged from 0.22 to 0.68. Probability of the second most likely diagnosis ranged from 0.07 to 0.38. These results are summarised in **Table 5.10**.

3 patients were misclassified in training and two are highlighted here as being particularly difficult to correctly diagnose. One patient with 17αOHD (Record Nb 1012) was misclassified as PORD in all 4 training folds. One patient with PORD (Record Nb 977) was misclassified as 21OHD in 3 out of 4 training folds. One patient with 17αOHD

(Record Nb 858) was misclassified as PORD in 2 out of 4 training folds. The second most likely diagnosis was correct for both patients with 17αOHD in all incorrect rounds of training. The patient with PORD was correctly identified by the second most likely diagnosis in only one of the incorrect training folds. Patients 1012 (17αOHD) and 977 (PORD) were equivalently misclassified in training for the geodesic averaged model with full features, with reduced features using a 70% relevance threshold and with reduced features using a 55% relevance threshold. Interestingly, these patients were also misclassified by the biochemical ratios approach (1012 is also classified as having PORD and 977 is also classified as having 21OHD). Patients 1012, 977 and 858 were all male, 1012 and 858 were adults and 977 was a child of 11 years.

For the cost weighted geodesic averaged model with reduced features using a 55% relevance threshold, again looking at the cumulative sum of test fold results, the results are presented in **Figure 5.9** and **Table 5.11**. There were 88 healthy control misclassifications (<6m n=34, 6m-16yr n=12, >16yr n=42). The most frequent misclassifications were made with 3βHSD2D (n=25), PORD (n=22) and 5αRD (n=22). Of the 88 healthy controls who were misclassified, 72 were identified as healthy by the 2nd most likely diagnosis. The 16 healthy controls who persisted to be misclassified were assigned following diagnoses: 21OHD n=2, PORD n=6, 17αOHD n=1, 3βHSD2D n=4 and 5αRD n=3. 8 of these were less than 6 months of age, 1 was 6 months to 16 years and 7 were over 16 years of age. There were 5 patient misclassifications. 1 patient with 21OHD, a baby girl, was classified as 5αRD but the diagnosis was correct for the 2nd most likely diagnosis (probability 0.58 vs 0.16). 1 patient with PORD, an 11yr male, was classified as 21OHD (probability 0.68). 2 patients with 17αOHD, both adult males, were misclassified, 1 for 5αRD (probability

0.33) and 1 for PORD (probability 0.6). The latter was correctly assigned by the second most likely diagnosis (probability 0.13). Finally, 1 patient with 3βHSD2D, a 9-year-old female, was classified as healthy (probability 0.63). Patients 1012 and 977, mentioned above, were misclassified in all 4 training rounds. In addition, for this model, patient 961, an adult male with 17αOHD, was also misclassified in all 4 training rounds.

While there is a reasonably even distribution of male and female patients misclassified in testing (6 of each sex for the full features model described, and 2 females and 3 males for the reduced features model described), the four particularly difficult patients misclassified in training across these two models were all male. These are small numbers but sex-specific challenges in differential diagnosis warrants further investigation once more data becomes available to support this analysis.

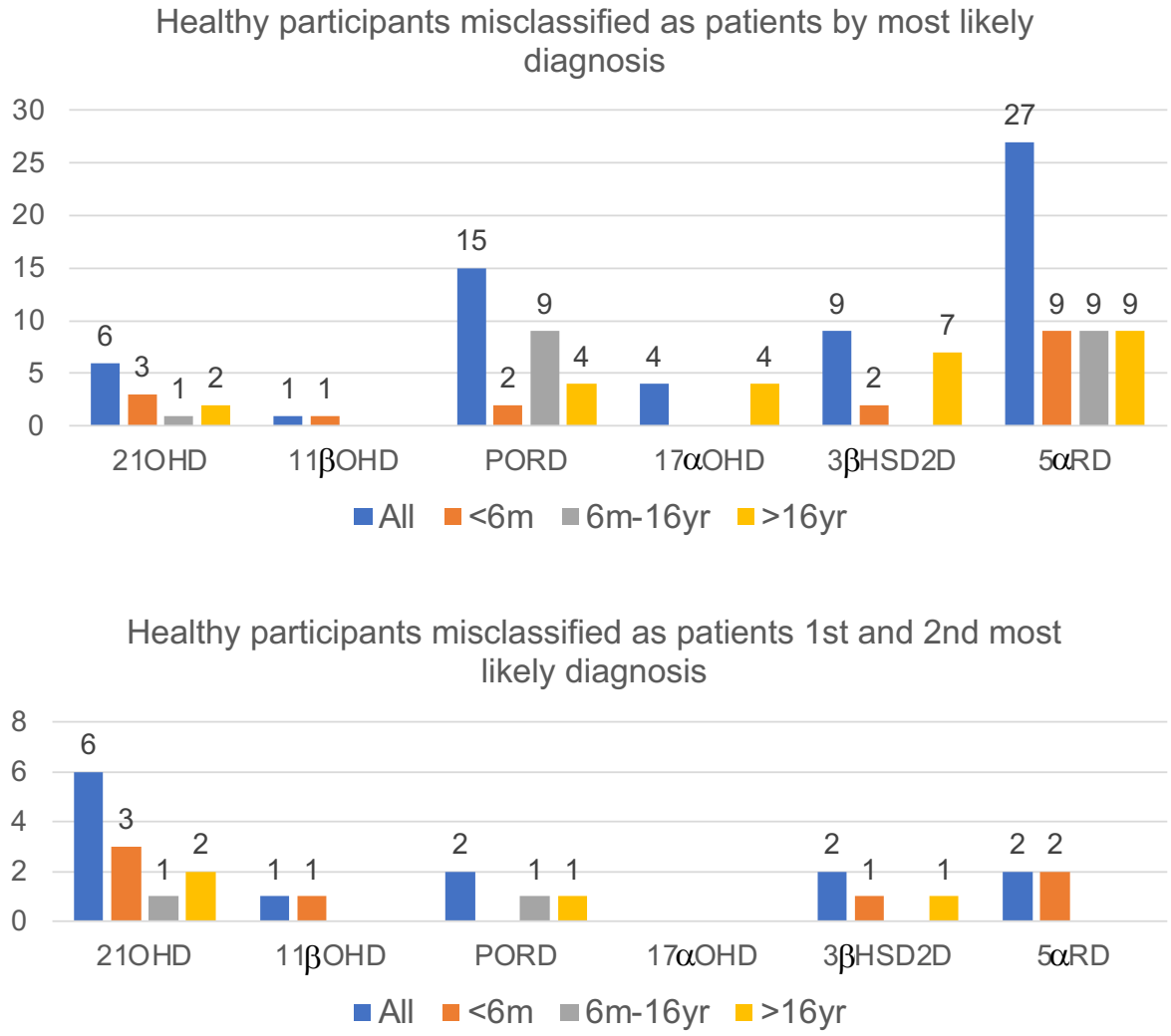


Figure 5.8: Healthy participants misclassified as patients

Cost weighted probabilistic ALVQ, geodesic averaged model, all features used. Cumulative results across 5 test folds shown only. Top panel: healthy controls misclassified as patients looking only at the most likely diagnosis. Bottom panel: healthy controls who remain misclassified as patients by the first and second most likely diagnosis, diagnosis assigned is most likely diagnosis.

Table 5.10: Patient misclassifications

Cost weighted probabilistic ALVQ, geodesic averaged model, all features used. Cumulative results across 5 test folds shown only. Probability of diagnosis is shown in brackets. Grey background indicates patients who are correctly identified by the second most likely diagnosis and for these patients the probability of first vs second diagnosis is shown in brackets.

	Predicted Healthy	Predicted 21OHD	Predicted PORD	Predicted 17αOHD	Predicted 3βHSD2D	Predicted 5αRD
True 21OHD			1(0.47 vs 0.26)		1(0.52 vs 0.21)	1(0.39 vs 0.33)
True PORD	1(0.4)	1(0.68)		1(0.4 vs 0.35)		1(0.38)
True 17αOHD			1(0.59 vs 0.15) 1(0.63 vs 0.09)			
True 3βHSD2D	1(0.59 vs 0.1)					1(0.4 vs 0.29) 1(0.52 vs 0.17)

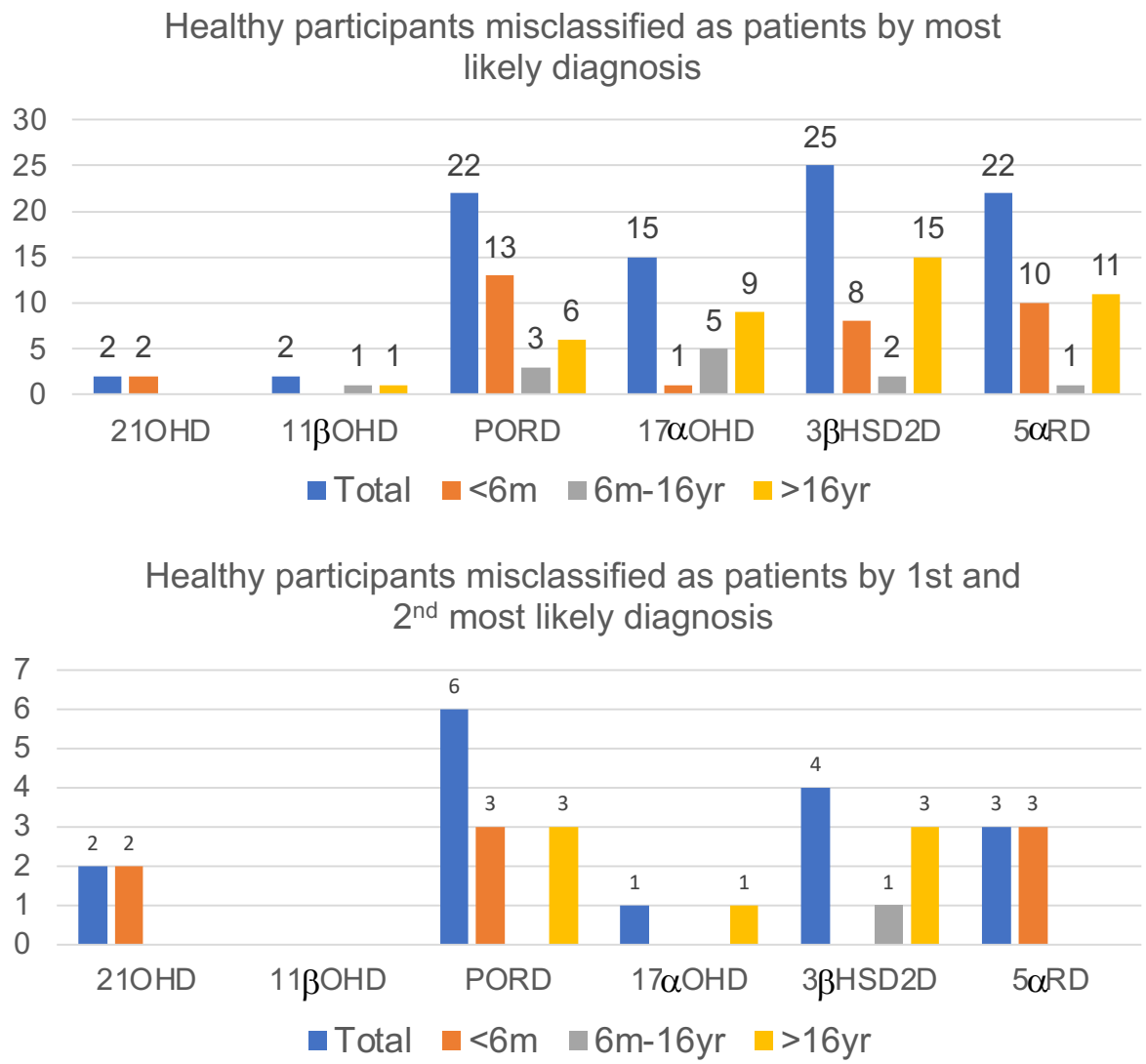


Figure 5.9: Healthy participants misclassified as patients (reduced features)

Cost weighted probabilistic ALVQ, geodesic averaged model, reduced features with 55% relevance threshold. Cumulative results across 5 test folds shown only. Top panel: healthy controls misclassified as patients looking only at the most likely diagnosis. Bottom panel: healthy controls who remain misclassified as patients by the first and second most likely diagnosis, diagnosis assigned is most likely diagnosis.

Table 5.11: Patient misclassifications (reduced features model)

Cost weighted probabilistic ALVQ, geodesic averaged model, reduced features with 55% relevance threshold. Cumulative results across 5 test folds shown only. Probability of diagnosis is shown in brackets. Grey background indicates patients who are correctly identified by the second most likely diagnosis and for these patients the probability of first vs second diagnosis is shown in brackets.

	Predicted Healthy	Predicted 21OHD	Predicted PORD	Predicted 5αRD
True 21OHD				1(0.58 vs 0.16)
True PORD		1(0.68)		
True 17αOHD			1(0.6 vs 0.13)	1(0.33)
True 3βHSD2D	1(0.63)			

5.8.4. Visual Interpretation Tools

Creating visualisations to assist non-experts in understanding how the classification method works is straightforward for ALVQ, as with other forms of LVQ, as the classification occurs on the hypersphere meaning that vectors can be displayed as datapoints via an embedding procedure. Through informative visualisations, clinicians could appreciate not only the general working of ALVQ, but also how their unclassified sample relates to samples from each of the known conditions and highlight if their patient sample lies close to a decision boundary indicating some degree of diagnostic uncertainty. For a preliminary 4-class problem, a 2-dimensional visualization of the relative sample and prototype projected vectors with approximate class boundaries superimposed is illustrated in **Figure 5.10.** In **Figure 5.11**, we show an approximation of how the classification is being made for the 7-class problem. This figure shows geodesic distance on the hypersphere from samples to their respective prototypes, each symbol representing a single participant and the position of the sample dictated by the 496 unique pairwise interactions of the 32 measured urinary steroid metabolites. While this isn't the distance measure employed by our classification method, it closely approximates and allows us to create a clear and interpretable visualisation tool which displays sample vectors clustering closest to their respective class prototype and other class sample vectors clustering further away. To further aid with user interpretation, we can overlay a single "unknown" participant to show how they relate to each of the different class prototypes to establish a diagnosis. This is illustrated in **Figure 5.12** where we show a single unknown patient sample and the relative position to each of the class prototypes. The individual in this example has 5αRD and we can see that they would be correctly identified through this method.

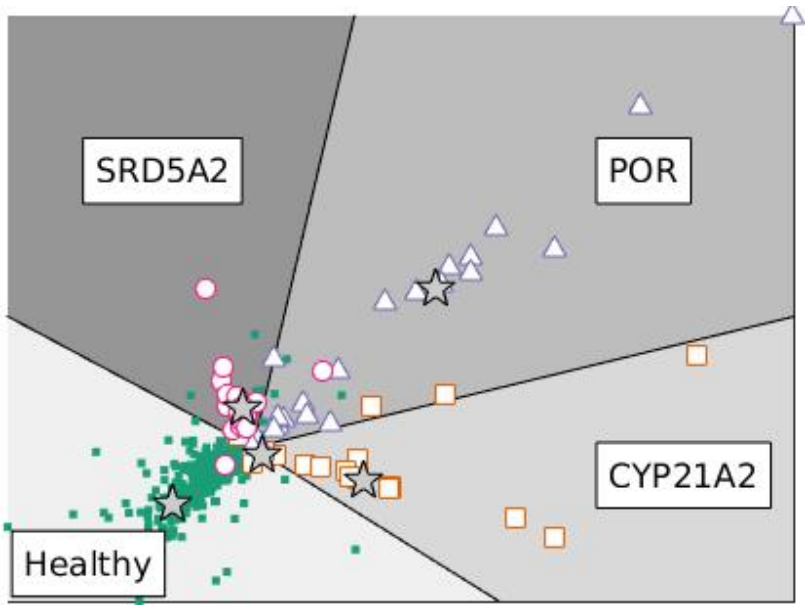


Figure 5.10: ALVQ, deterministic version, 4-class visualisation

Green dots represent healthy controls, pink circles represent patients with SRD5A2 (5αRD), purple triangles represent patient with POR (PORD) and orange squares represent patients with CYP21A2 (21OHD). Each symbol represents an individual participant, the position of the symbol determined by the interaction of 496 pairwise urinary steroid metabolite ratios. Stars show the position of class prototypes. Class boundaries are overlayed.

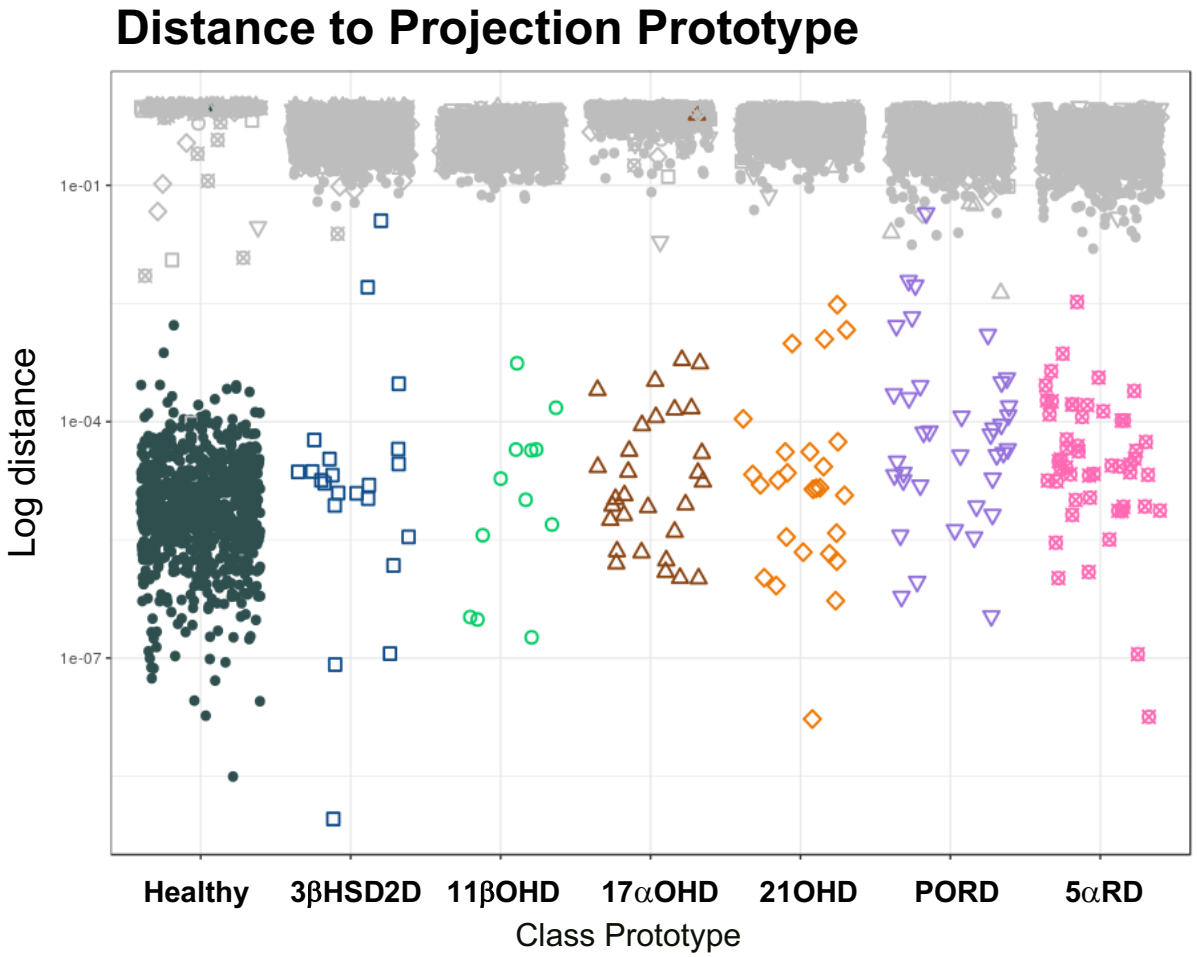


Figure 5.11: Illustration approximating the classification technique

Illustration shows geodesic distance on the hypersphere from samples to respective prototypes. Each symbol represents a single individual. All samples for all classes are displayed relative to each prototype. The closer the sample is to the x-axis, the closer to that prototype. Samples are displayed clustering closest to their own prototype.

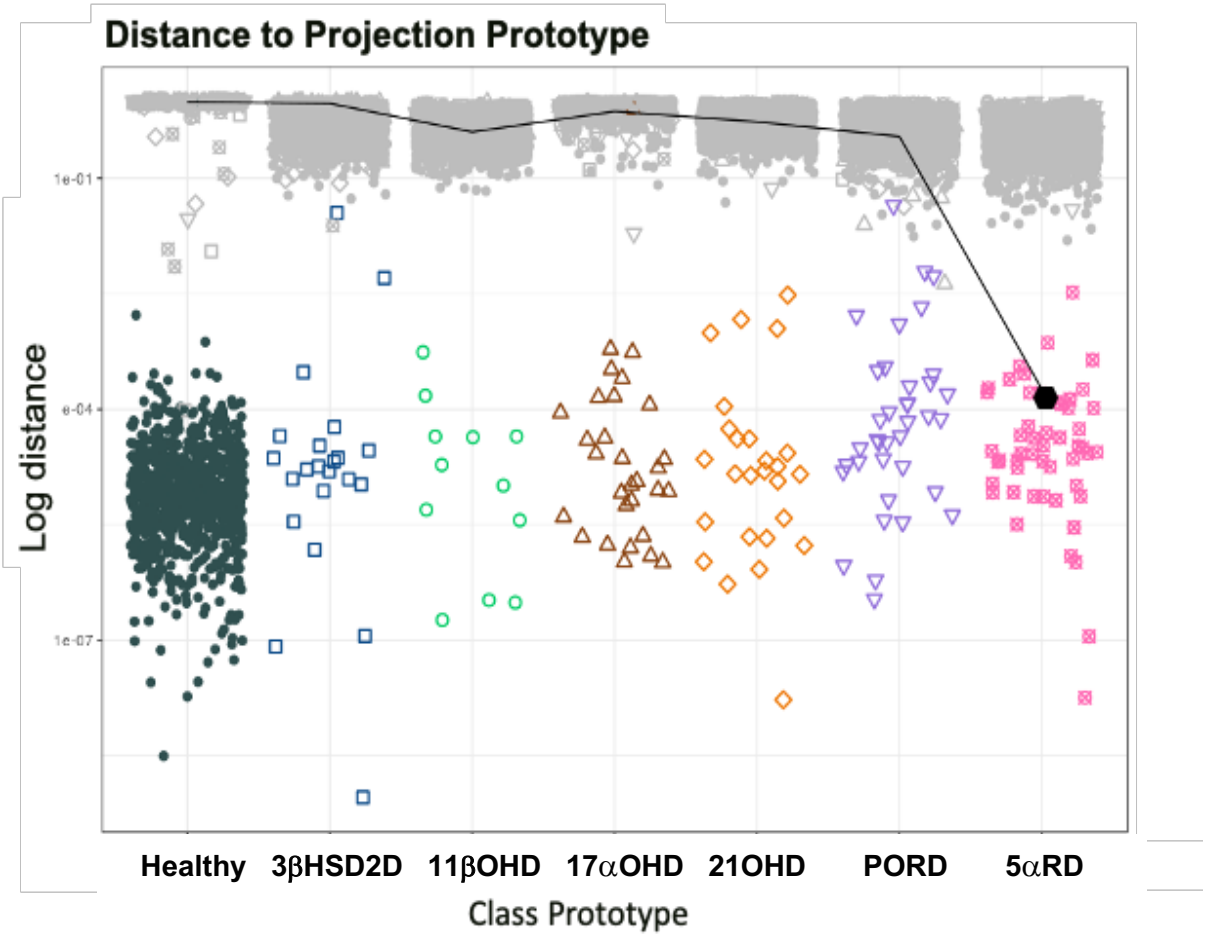


Figure 5.12: Illustration approximating the classification technique with the relative position of a single unidentified sample relative to class prototypes indicated

Illustration shows geodesic distance on the hypersphere from samples to respective class prototypes. Each symbol represents a single individual. All samples for all classes are displayed relative to each prototype. The closer the sample is to the x-axis, the closer to that prototype. Samples are displayed clustering closest to their own prototype.

For the 7-class problem, a representative 2-dimensional visualisation is more complex. 1 model out of 5 from the cost weighted probabilistic geodesic averaged ALVQ was used to inform a helpful visualisation. The 496-dimensional hypersphere was reduced to 3 dimensions and then transformed into its Mollweide projection for ease of visualisation. The Molleweide projection is a method commonly used to produce maps of the world, transforming a 3-dimensional globe into a 2-dimensional map. It maintains accuracy in the proportions of area for the cost of accuracy of shape. It is a pseudocylindrical map projection, where central meridian and equator are perpendicular, lines of latitude are straight and parallel, and lines of longitude are curved. Traditionally, it takes an ellipsoid shape, but can be projected onto a disc to take a circular shape, as in our example (Lapaine, 2011). This is shown in **Figure 5.13**. A 3-dimensional rotating sphere can also be constructed by taking single images of slightly different projections of the 3-dimensional sphere and sequentially showing the images in a gif form.

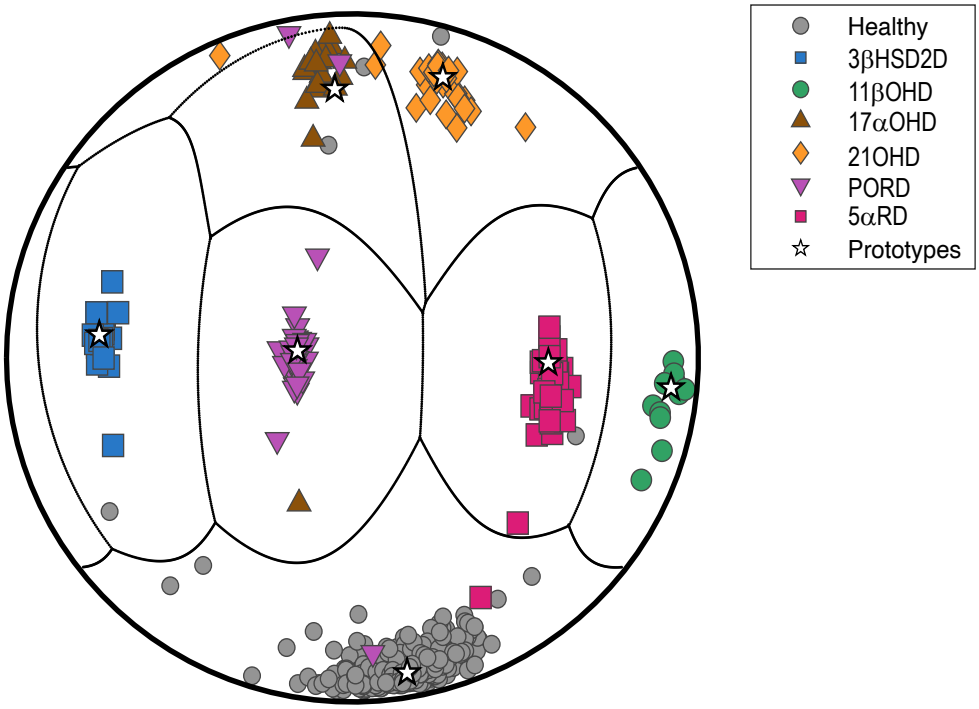


Figure 5.13: Illustration of classification on the hypersphere

Mollweide projection of the decision boundaries of the probabilistic ALVQ classifier, prototype positions, and where the data-points lie in relation to the prototypes and decision boundaries. Each symbol is a single individual, the position of the symbol is determined by the interaction of all pairwise combinations of measured urinary steroid metabolites. Bold stars are prototypes- computationally derived typical representative examples of each disease class. Unseen samples are classified according to which prototype the sample projects closest to on the hypersphere. Distance to prototypes is determined by the cosine of the angle between projected vectors.

5.8.5. Knowledge Extraction from ALVQ Variants

The LVQs are fully transparent machine learning methods, meaning we can easily analyse the way decisions are made by the classifier. For our problem, we can look in detail at the prototypes constructed during training to see what the algorithm learned to be the typical representative pattern of metabolites for each disease class.

For these experiments, information gained from the cost-weighted, averaged probabilistic ALVQ model with all 496 features was used. The classification terms for each condition were examined, including both diagonal and off-diagonal elements in the relevance matrix constructed for each disease prototype. Features (ratios) were ranked in descending order of relevance to identify the most discriminatory. Across the 5 averaged models constructed, while the exact feature order varied slightly between models, each model displayed very little variation in the terms deemed overall most discriminatory for each condition. Therefore, we took the cumulative sum of the classification terms for each prototype across the 5 averaged models to determine the stereotypical profile for each condition. As we used the cumulative sum across all 5 folds of training, this examines the whole dataset simultaneously therefore we present a descriptive analysis that does not equate to test performance or generalisability. By ranking features by relevance rather than by their absolute magnitude, we are not only able to pick out which ratios are significantly raised, but also importantly which are the more unique or discriminatory for each class.

To determine the relative complexity of the classification task for each of the classes, starting from the highest relevance, features were systematically added in one at a

time to observe the effect increasing the number of features (ratios) used has on discriminative ability. For each class in question, the true positive rate, or sensitivity of each class was calculated with each additional feature, aiming to discriminate the class from all other classes combined. **Figure 5.14** shows these results. All classes are shown simultaneously on this graph. We can observe from this figure that the top 60 or so ratios are critical in order to pick up the maximal number of patients. Discrimination of 11βOHD is by far the simplest classification task, with all patients identified with only 3 ratios. Of the patient classes, PORD was the most complex to diagnose, requiring 113 ratios to be able to identify 97.2% of PORD patients. We were also unable to correctly identify all patients with 17αOHD, 20 ratios were required for the maximum identification of 96.7% of patients. The healthy group was by far the most complex classification task to identify, requiring 358 ratios for maximum pick up, with 5.3% still being missed. 100% separation was achievable in 4 classes: 21OHD, 11βOHD, 3βHSD2D and 5αRD.

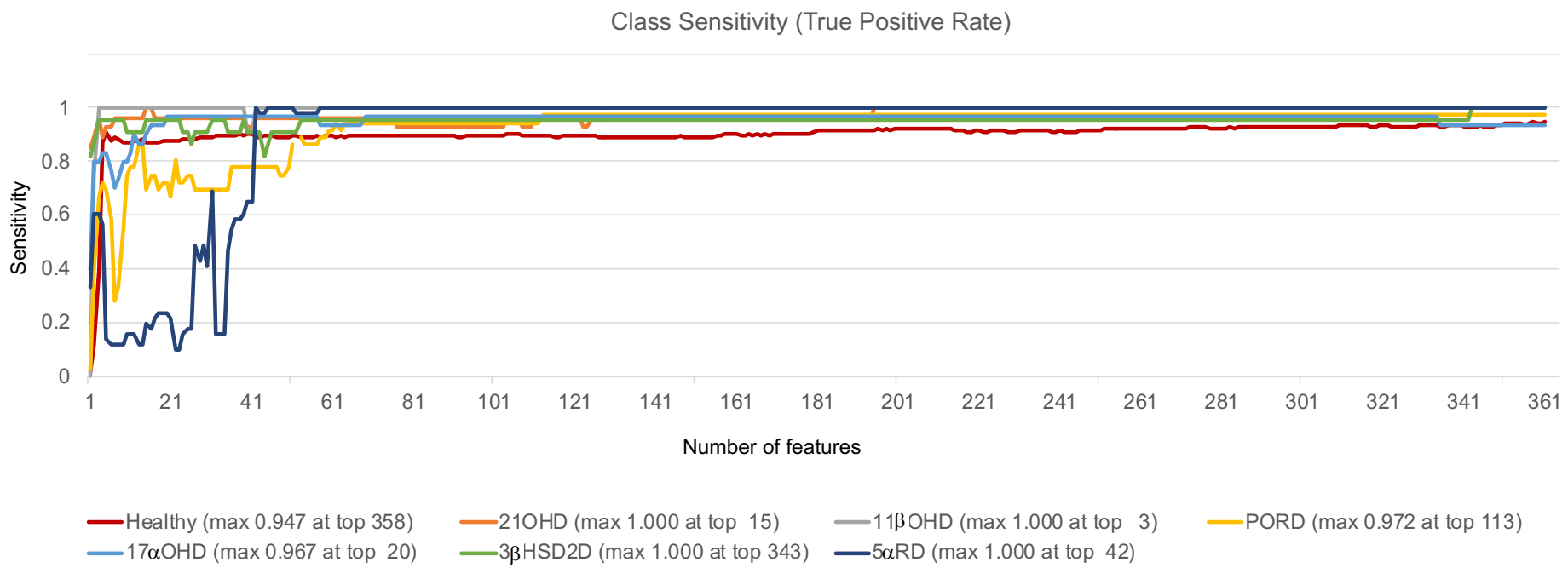


Figure 5.14: Sensitivity per the number of features (ratios, in descending order of relevance) used for the classification of a given class against all other classes combined

Prototypic classification terms were examined, and features ranked in order of relevance. The graph shows how many of the most relevant ratios are required to discriminate a given class from all others combined. Results for all classes including healthy are shown according to colour.

To better ascertain exactly which metabolites were deemed characteristically important for discrimination of each class we took two approaches. First, we examined the top 50 most relevant features (ratios) for each class. For each ratio we determined which was the higher value and which was the lower value (by comparing to healthy group class average for each metabolite). Then we ascertained the frequency in which individual metabolites occurred as a numerator (high value) or denominator (low value) in these top 50 ratios. The result is a graphic which represents the data driven, unique “urine steroid metabolite fingerprint” for each condition as determined by the probabilistic ALVQ model. This is shown in **Figure 5.15**. **Figure 5.16** and **Figure 5.17** show the same information but split into the metabolites that are characteristically raised and the metabolites that are characteristically reduced displayed separately. Metabolites are colour coded according to their origin (dark blue- androgens, light blue- androgen precursors, green – mineralocorticoids, yellow – glucocorticoid precursors, orange – glucocorticoids).

The second approach that we took was to find the number of features that resulted in the optimum balanced accuracy measure, the mean of sensitivity and specificity, for each class. and produce the same figures using these most discriminatory features. By using this approach, we looked at only the most relevant features per class and no more. This is less informative to identify subtle changes in the wider urine steroid metabolome but instead highlights two other important characteristics- first, the relative complexity of each class to accurately discriminate, and second, which are the key metabolites within the urine steroid metabolome for each condition. The results for this are shown in **Figure 5.18**, **Figure 5.19** and **Figure 5.20**.

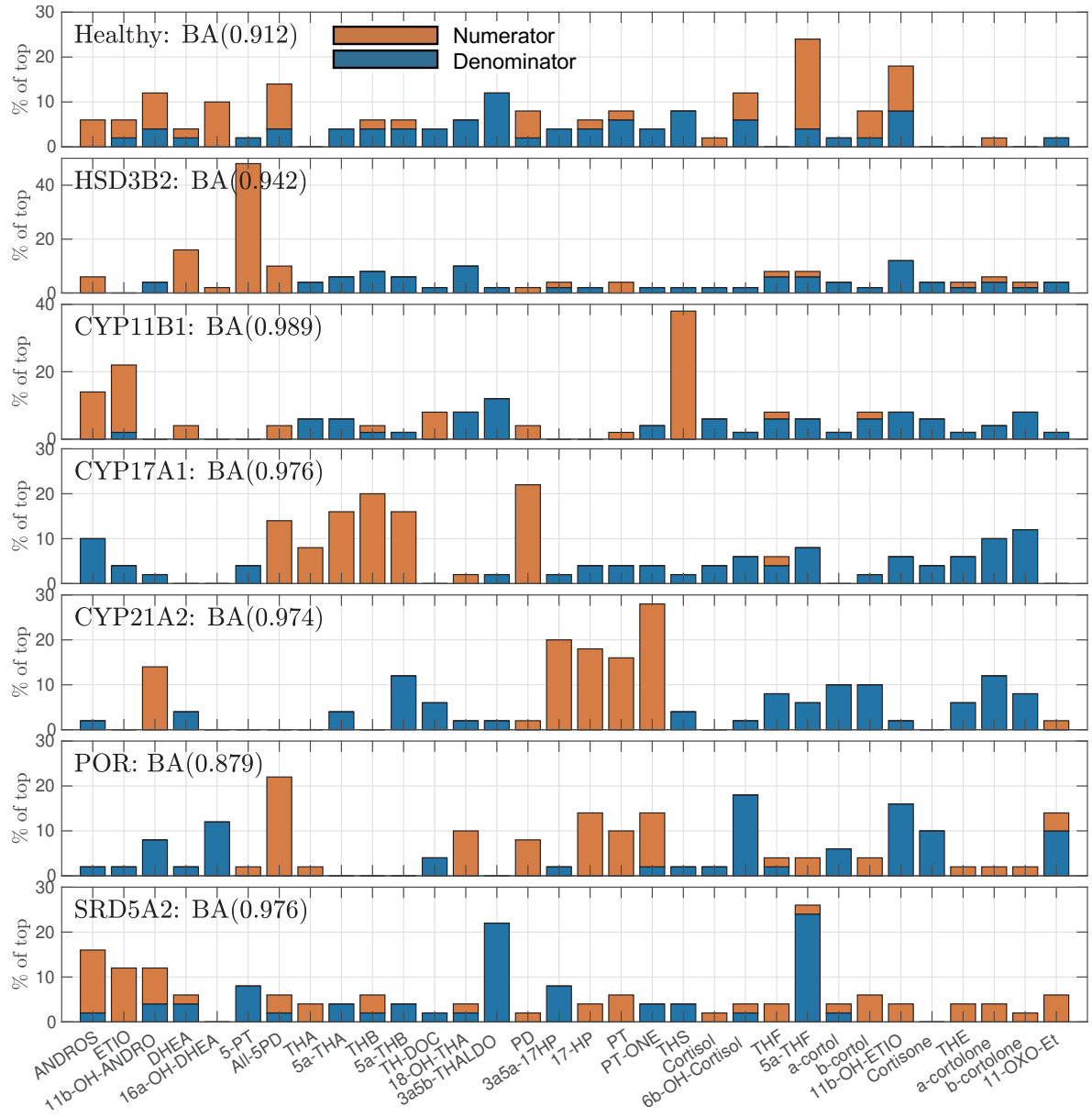


Figure 5.15: The characteristic urinary steroid metabolome changes observed for each condition – all important metabolites

Bars indicate the % frequency a metabolite (across the x-axis) occurs in the top 50 discriminatory ratios for each condition. Red indicates that the metabolite concentration is observed to be high (numerator), and blue indicates the metabolite concentration is observed to be low (denominator). BA = balanced accuracy, the mean of sensitivity and specificity for each class when the top 50 ratios are used to discriminate the named class from all other classes combined.

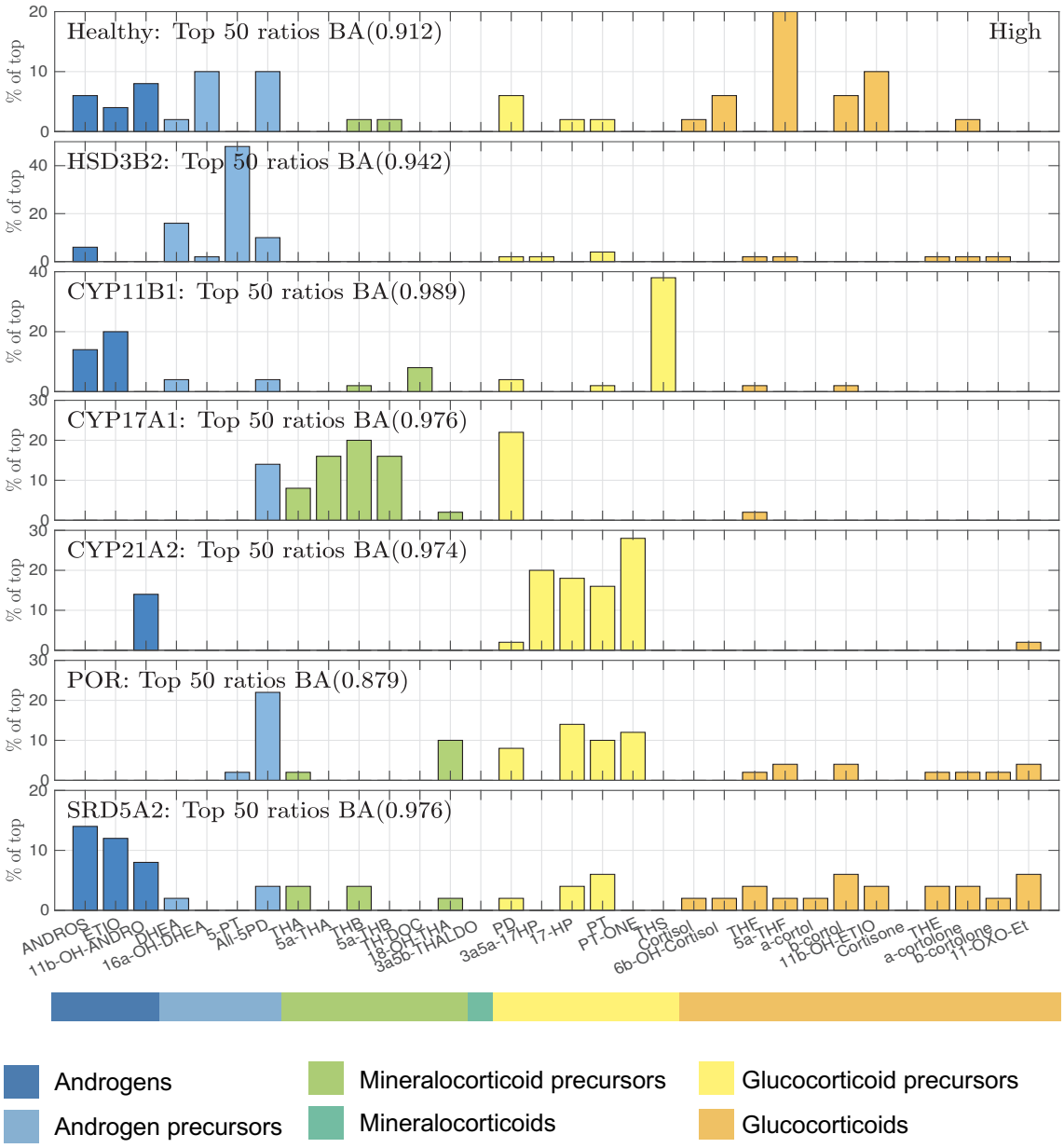


Figure 5.16: The characteristic urinary steroid metabolome changes observed for each condition – only metabolites observed to be raised

Bars indicate the % frequency of a metabolite (across the x-axis) occurring as the numerator (i.e., a high value) in the top 50 discriminatory ratios for each condition. BA = balanced accuracy, the mean of sensitivity and specificity for each class when the top 50 ratios are used to discriminate the named class from all other classes combined. Colours indicate the origin of the metabolite in the steroidogenesis pathway.

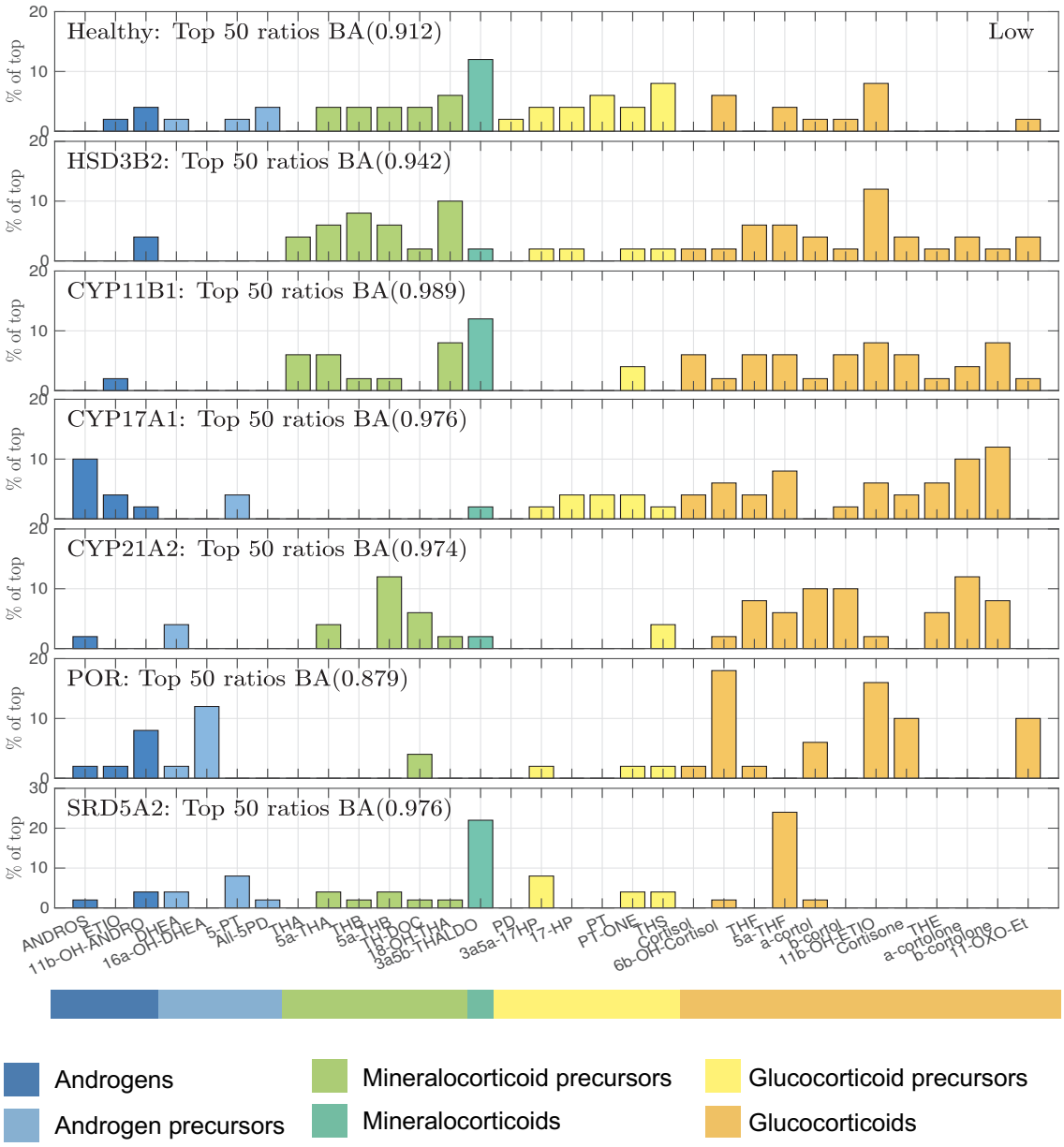


Figure 5.17: The characteristic urinary steroid metabolome changes observed for each condition – only metabolites observed to be reduced

Bars indicate the % frequency of a metabolite (across the x-axis) occurring as the denominator (i.e., a low value) in the top 50 discriminatory ratios for each condition. BA = balanced accuracy, the mean of sensitivity and specificity for each class when the top 50 ratios are used to discriminate the named class from all other classes combined. Colours indicate the origin of the metabolite in the steroidogenesis pathway.

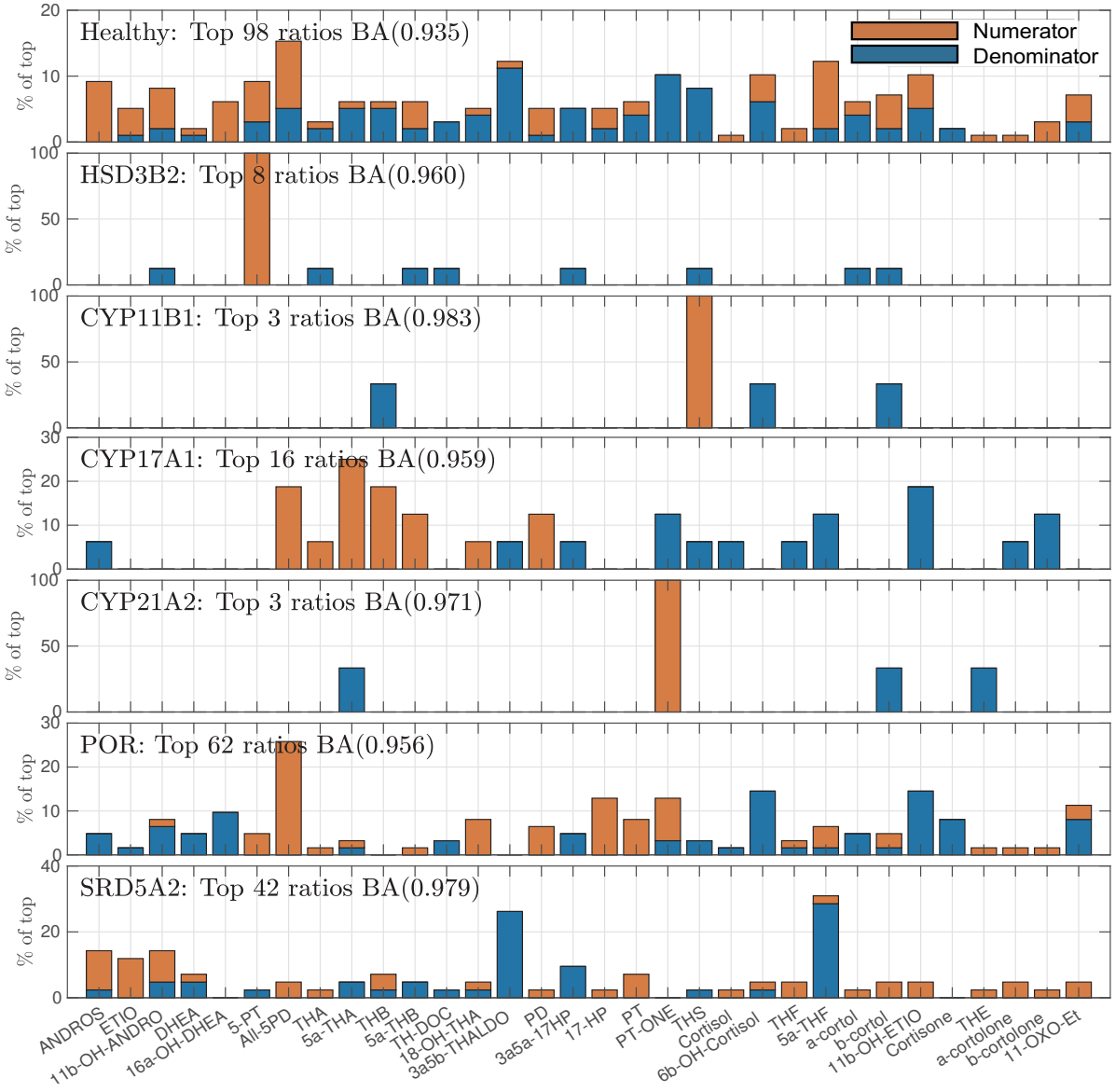


Figure 5.18: The characteristic urinary steroid metabolome changes observed for each condition – only the most discriminatory features for each class

Bars indicate the % frequency a metabolite (across the x-axis) occurs in the most discriminatory ratios for each condition. Red indicates that the metabolite concentration is observed to be high (numerator), and blue indicates the metabolite concentration is observed to be low (denominator). BA = balanced accuracy (top left of each panel), the mean of sensitivity and specificity to discriminate the named class from all other classes combined. The number of ratios required to reach the maximum balanced accuracy is also indicated at the top left of each panel.

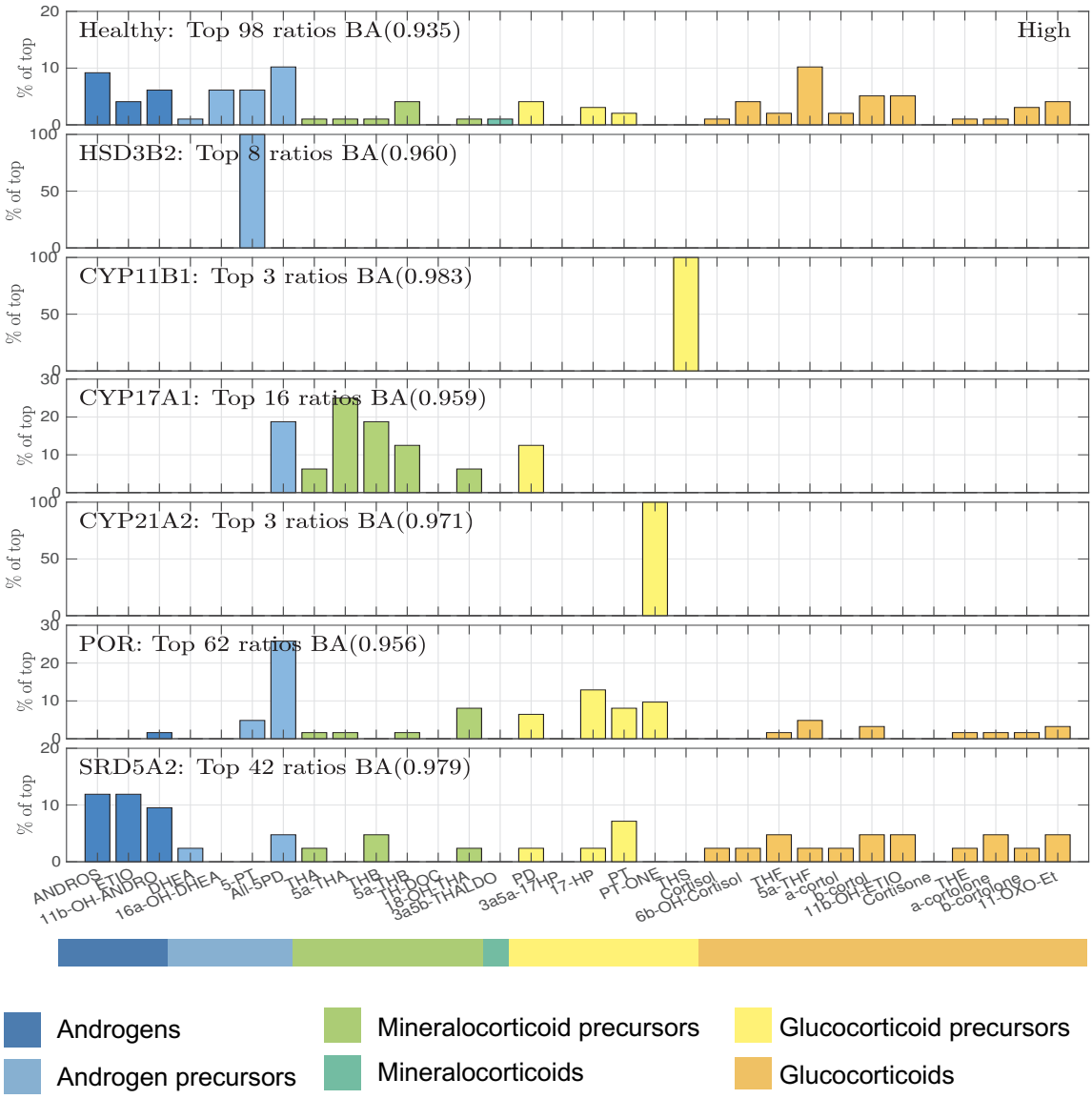


Figure 5.19: The characteristic raised metabolites observed for each condition – the most discriminatory features for each class

Bars indicate the % frequency a metabolite (across the x-axis) occurs as the numerator (i.e., a high value) in the most discriminatory ratios for each condition. Colours indicate the origin of the metabolite in the steroidogenesis pathway. BA = balanced accuracy (top left of each panel), the mean of sensitivity and specificity to discriminate the named class from all other classes combined. The number of ratios required to reach the maximum balanced accuracy is indicated at the top left of each panel.

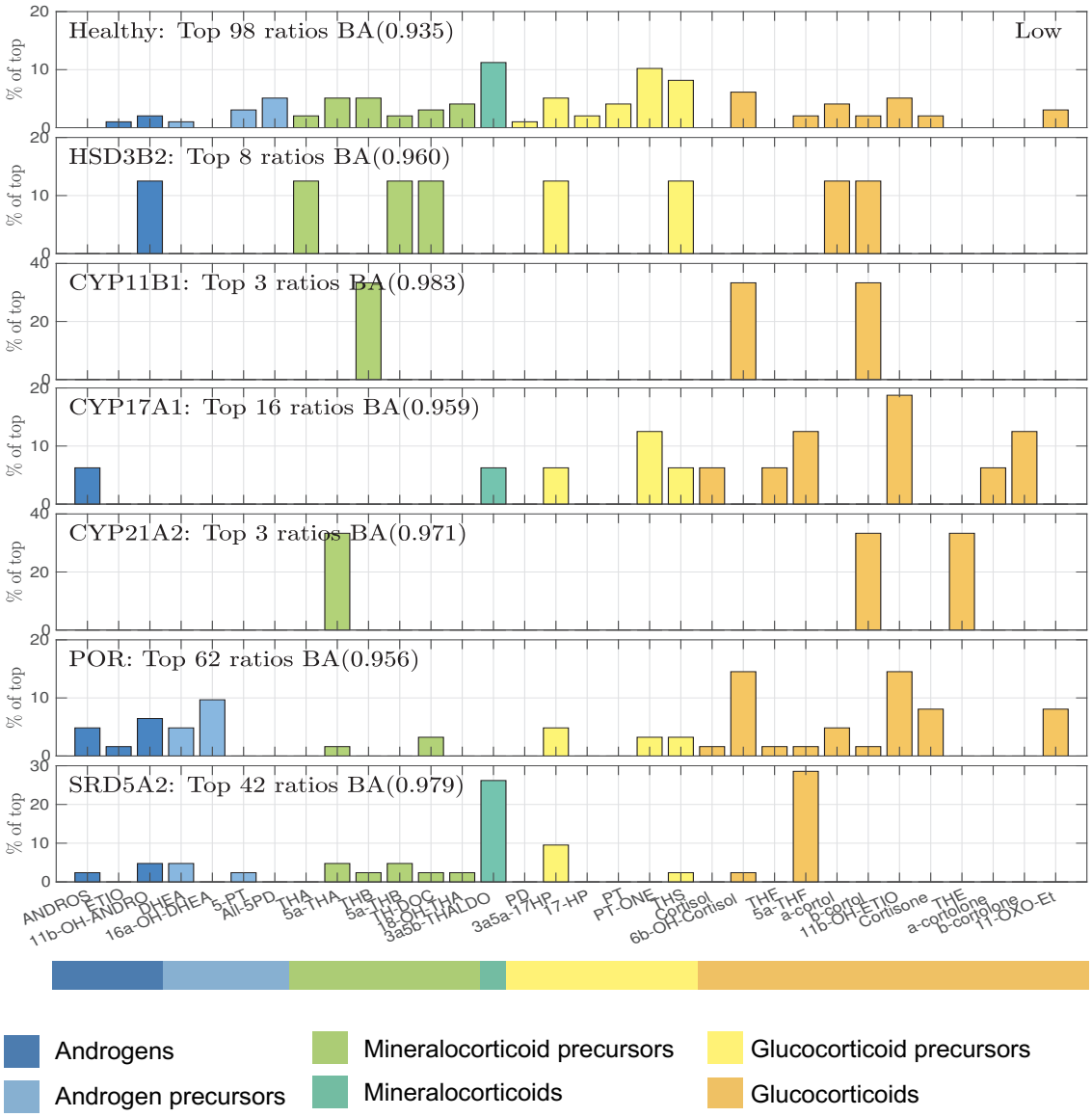


Figure 5.20: The characteristic reduced metabolites observed for each condition – the most discriminatory features for each class

Bars indicate the % frequency of a metabolite (across the x-axis) occurring as the denominator (i.e., a low value) in the most discriminatory ratios for each condition. Colours indicate the origin of the metabolite in the steroidogenesis pathway. BA = balanced accuracy (top left of each panel), the mean of sensitivity and specificity to discriminate the named class from all other classes combined. The number of ratios required to reach the maximum balanced accuracy is indicated at the top left of each panel.

From these figures, we can reassuringly see a great deal of agreement with existing domain knowledge, with the emergence of ratios from the list of established biochemical ratios within these results. Overall, the low values are seen to be common between different conditions, whereas the high values show more distinct patterns unique to each condition. Common to all CAH conditions and not observed in 5αRD, glucocorticoid metabolites were predictably low.

For 21OHD, glucocorticoid precursor metabolites 17HP, 3α5α17HP, PT and PTONE frequently occurred as high concentration metabolites, with downstream glucocorticoid metabolites appearing in low concentrations. Metabolites of the 11-oxygenated androgen pathway and the alternative pathway to DHT synthesis were particularly highlighted as high magnitude (11βOHAn and 3α5α17HP), whereas classical androgen pathway metabolites were not deemed to be discriminatory. Looking at only the most relevant features to maximise balanced accuracy, only 3 ratios were required for class discrimination, and of those PTONE was the only key raised metabolite.

For 11βOHD, 11-deoxycortisol metabolite THS and classical pathway androgen metabolites An and Et were the predominant high concentration metabolites, with reduced downstream glucocorticoid metabolites, consistent with existing domain knowledge. Looking at only the most relevant features to maximise balanced accuracy, only 3 ratios were required for class discrimination, and THS was the numerator in all 3 ratios. This agrees with existing domain knowledge, that THS is the most important metabolite for identification of this condition.

For PORD, the high concentration metabolites were as expected- 5PD, PD, 17HP, PT, PTONE, THA, 18-OH THA. Looking at the consistently low metabolites for PORD, androgens and glucocorticoid metabolites were low as expected but the most frequently observed low concentration metabolite was 6βOH-cortisol. This is a relatively overlooked metabolite in standard analysis, but this observation is logical from a biochemical perspective. Not only is this a metabolite of cortisol, which would be reduced in PORD, but it is produced from cortisol through the action of CYP3A4, which also requires POR as an electron donor for its function. Looking at only the most relevant features, 62 ratios were required to achieve the maximum balanced accuracy which reflects the known complexity of this condition compared to the other causes of CAH. The key raised metabolite was 5PD.

For 17αOHD, androgen and glucocorticoid metabolites were expectedly reduced while mineralocorticoid metabolites were expectedly raised. Progesterone metabolite PD was the most frequently occurring high concentration metabolite in the top 50 ratios. The observation that PD was particularly raised in this condition is also made in the biochemical ratios chapter. Inclusion of the “PD/THE+THF+5aTHF” ratio in the physiologically informed approach, described in the previous chapter, was a frequent cause for these patients to be misclassified as PORD. Looking at the most discriminatory features, 16 ratios were required to maximise balanced accuracy and the key metabolites mirror what is observed in the top 50 ratios. The key raised metabolites were PD, the mineralocorticoid metabolites (particularly THB) and 5PD.

For 3βHSD2D, by far the most important high concentration metabolite observed was 5PT which is entirely consistent with existing domain knowledge. 8 ratios were

required to maximise balanced accuracy and all 8 contained 5PT as the numerator. 5PD and DHEA also occur as high metabolites in the top 50 magnitude ratios and all downstream metabolites are expectedly low. Low magnitude metabolites occurred in similar frequency, representing the high block in the steroidogenesis pathway leading to a reduction in metabolites from all three adrenal hormone pathways.

For 5αRD, more significance is assigned to the absent metabolites, as opposed to the raised metabolites. 5αTHF is the most frequently occurring reduced or low metabolite, consistent with domain knowledge. Looking at the ratios produced from the machine learning method, it is reassuring “THF/5αTHF” and “Et/An”, which perform very well for this condition in the biochemical ratios approach, did appear in the top discriminatory ratios using this data driven approach. However, 5αTHF was deemed much more important than THF for accurate discrimination, and more commonly observed were ratios where androgen metabolites were the raised numerator. Other metabolites which frequently occurred in low concentrations were 3α5α17HP and 3α5βTHAldo. 5α-reductase is an important enzyme for the alternative pathway from 17OHP to DHT synthesis, and 3α5α17HP is a metabolite arising from this alternative pathway, so this observation makes biochemical sense. 3α5βTHAldo is a metabolite of Aldosterone which is apparently produced through action of 5β–reductase as opposed to 5α-reductase, so it is very interesting that this metabolite is identified by the data-driven approach as being as significant as 5αTHF. 42 features were required to maximise balanced accuracy. This was surprising as this condition is very straightforward to identify using 3 single metabolite ratios as per the findings in chapter 4. However, we do not include age information in the machine learning method which is particularly important for the key metabolites in this condition, as described in

chapter 3. There is also likely to be a high degree of overlap with healthy control subjects for the majority of the observed steroid metabolome which remains unaffected despite this enzymatic defect.

5.9. Discussion

In this chapter we present the creation of a machine learning method which is an extension of learning vector quantization, designed to discriminate patients with inborn steroidogenesis disorders from each other and from healthy controls. We show how we developed this method from a deterministic to probabilistic approach to provide users with useful information about the certainty of diagnosis. We refined our method to improve performance through domain knowledge driven cost weighting to handle bias which occurs secondary to the imbalanced classes while also highlighting the key unfortunate misclassifications which should incur the maximum penalty during training. We further refined our method by combining the power of multiple learners through an ensembled and a geodesic averaged approach. Our final approach to improve performance was through feature reduction. We maximise on the transparency of the classification method employed by our model to produce figures which effectively display for users how their patient's sample is assigned to a given class. We were also able to display the characteristic observed steroid metabolome fingerprint that has been gathered through training the class prototypes.

Our classifier achieved excellent sensitivity results while maintaining acceptable specificity results, superior to the performance achieved through use of established biochemical ratios as described in chapter 4 (see **Table 5.12** for a comparison of the

results for the 3 approaches described in this thesis). Because 3 different methods have been used for the creation of these 3 approaches, the performance results in this table are not directly comparable. However, for the biochemical ratios approach using 97.5th percentile as the upper limit of normal, a cross validation approach was also used to enable a more valid comparison of performance with the machine learning techniques- these results with the final machine learning results are shown in **Table 5.13**.

For ease of construction of the geodesic average model and maintenance of interpretability, we have restricted the number of prototypes to one per class. It is possible that specificity could be improved without sacrificing sensitivity by the inclusion of a second prototype for the healthy class. This hypothesis has not been explored in the work presented in this thesis but construction of an approach able to utilise multiple prototypes per class, ideally maintaining interpretability, would warrant exploration in future work. While we improved the model performance with each method of model refinement, there appeared to be certain individual patients that were consistently difficult to correctly diagnose across all model variants. These individuals were also misclassified by the biochemical ratios approach. Due to the natural variation observed in the human metabolome, there is unlikely to be a perfect classifier. This highlights previous publications indicating the importance that any machine learning method operating in the healthcare domain is completely transparent in order to maintain accountability for any misclassifications (Holzinger et al., 2019; Seneviratne et al., 2020; Vayena et al., 2018; Watson et al., 2019).

Table 5.12: Comparison of results obtained from the 3 described classifiers.

Shown are the results from the classifiers constructed using biochemical ratios with 97.5th percentile as upper limit of normal (whole cohort analysis), biochemical ratios with an optimised multiple of the median approach (whole cohort analysis) and the cost weighted, reduced features (55%), geodesic averaged probabilistic ALVQ model (5 folds of cross validation).

	Biochemical Ratios, 97.5th percentile ULN [95%CI]	**Biochemical Ratios, Multiple of Median ULN** [95%CI]	**Angle Learning Vector Quantization** Mean, (SD)
Sensitivity (Any disease)	100% [97.9 - 100]	100% [97.9 - 100]	99.5% (0.01)
Sensitivity (Correct disease)	89.9% [84.5 - 93.5]	94.4% [92.1 - 98.1]	97.1% (0.02)
Healthy (Specificity)	80% [77.1 - 82.6]	79.1% [76.2 - 81.8]	89.4% (0.03)
21-hydroxylase deficiency	81.5% [63.3 - 91.8]	96.3% [81.7 - 99.3]	96% (0.09)
11β-hydroxylase deficiency	83.3% [55.2 - 95.3]	100% [75.8 - 100]	100% (0)
P450 oxidoreductase deficiency	77.8% [61.9 - 88.3]	91.7% [78.2 - 97.1]	97.5% (0.06)
17α-hydroxylase deficiency	90.0% [74.4 - 96.5]	90.0% [74.4 - 96.5]	93.3% (0.15)
3β-hydroxysteroid dehydrogenase type 2 deficiency	100.0% [85.1 - 100]	100.0% [85.1 - 100]	96% (0.09)
5α-reductase type 2 deficiency	100.0% [93.0 - 100]	100.0% [93.0 - 100]	100% (0)

Table 5.13: Comparison of results from classifiers constructed using cross validation

Results shown are from the classifier constructed using the biochemical ratios and 97.5th percentile as upper limit of normal compared to the cost weighted, reduced features (55%), geodesic averaged probabilistic ALVQ model. For both approaches, results shown here were obtained using 5 folds of cross validation, with identical separation of subjects into training and testing folds.

	Biochemical Ratios, 97.5th percentile ULN Mean, (SD)	**Angle Learning Vector Quantization** Mean, (SD)
Sensitivity (Any disease)	100% (0)	99.5% (0.01)
Sensitivity (Correct disease)	90.4% (0.05)	97.1% (0.02)
Healthy (Specificity)	72.7% (0.04)	89.4% (0.03)
21-hydroxylase deficiency	81.3% (0.16)	96% (0.09)
11β-hydroxylase deficiency	83.3% (0.21)	100% (0)
P450 oxidoreductase deficiency	83.6% (0.04)	97.5% (0.06)
17α-hydroxylase deficiency	85.3% (0.15)	93.3% (0.15)
3β-hydroxysteroid dehydrogenase type 2 deficiency	100.0% (0)	96% (0.09)
5α-reductase type 2 deficiency	100.0% (0)	100% (0)

Learning Vector Quantization was an excellent base method to build our classifier on given its inherent transparency (Ghosh et al., 2020). Our extension can intrinsically handle some of the key issues which are found in most human data databases – heterogeneity, missing values, imbalanced classes, and small class sizes which do not represent the relative importance of a given class. The final model developed was more sophisticated than previous proposed LVQ extensions in order to tackle these detrimental intrinsic factors. We did not need to include any patient demographic data or clinical features within the classification tool as we achieved satisfactory results without this information. A previous similar study aimed to design a model able to replicate domain expert analysis (Wilkes et al., 2018). As the diagnosis was genetically confirmed in all patient participants, our designed model has the potential to improve upon human analysis alone.

We have further demonstrated that the intrinsic interpretability of the method allows the ability to gain mechanistic insight through thorough investigation of learned prototypic information. It is reassuring that known information about these conditions, gathered over decades of study, is reflected in what the machine learning method has deemed discriminatory information. Certainly, this will aid acceptability of the method by the medical and biomedical communities. Probabilistic ALVQ has the potential to add to the knowledge landscape for inborn steroidogenic disorders or other conditions it is applied to. All 3 pathways to active androgen synthesis (classical, alternative, and 11-oxygenated androgens) have all been found to be raised in 21OHD in different studies (Kamrath et al., 2012; Turcu et al., 2016). Our model assigned more discriminatory weight to the alternative pathway and 11-oxygenated androgens for accurate discrimination of 21OHD than the classical pathway to DHT synthesis.

Overlooked metabolites might be important for classification, such as 6α–hydroxycortisol being significant for accurate discrimination of PORD and 3α5αTHAldo being a significantly affected metabolite in patients with 5αRD.

5.10. Conclusions

In conclusion, we present a collaboratively designed, fully interpretable machine learning method able to discriminate patients with inborn steroidogenesis disorders from each other and from healthy controls. We have shown that domain knowledge approaches and data driven approaches can be integrative rather than competitive. Probabilistic ALVQ performed with excellent sensitivity and specificity and is particularly well suited to rare disease research. The transparency of the method allows for mechanistic insight into these rare diseases, will aid with wider acceptability and support implementation in routine clinical practice.

6. Final Conclusions and Future Directions

We have presented in this thesis two methods for assisted interpretation of urinary steroid metabolome data for the differential diagnosis of inborn disorders of steroidogenesis. The first method took a physiologically informed approach, built on previously published biochemical steroid precursor to product metabolite ratios, informed by decades of previous study into these disorders. We presented two methods for determining reference intervals for this approach. The second method employed a Urine Steroid Metabolomics approach. This was a data-driven approach initially, with discriminatory patterns found in the data informing classification. We proceeded to create a more collaborative model by including domain knowledge led cost weighting. Finally, we designed a series of visualisations to display results and present the inner working of the model to users.

6.1. Comparison of Methods

The biochemical ratios with the clinical decision tool had acceptable performance. Given the method is founded in established methodology (Caulfield et al., 2002; Shackleton et al., 1980; Storbeck et al., 2019), this approach has high clinical translation potential. As previously noted, these biochemical ratios show a lack of specificity for the target disease (Lucas-Herald et al., 2015). We showed the importance of scaling ratio values and having a clinical decision algorithm or tool to aid their interpretation. We found it was particularly difficult to discriminate PORD using only ratio values. It is likely the clinical decision tool we have presented could be refined and improved with increased sophistication through inclusion of further domain knowledge evidenced to help discrimination of these disorders (Caulfield et al., 2002;

Koyama et al., 2012; Koyama et al., 2016). A balance is required though between performance and useability.

Using a multiple of the median approach for determination of reference intervals for steroid metabolome profiling is promising. We have shown it is a more stable marker than 97.5th percentile for skewed steroid metabolome data. However, care needs to be taken for how best to establish reference intervals using this approach. The benefit of using a percentile is that limits are constructed solely from the larger healthy class and therefore, in the case of rare diseases, determined thresholds are comparable between centres and more generalisable.

Our refined machine learning algorithm showed excellent training and test performance, with superior performance to the biochemical ratio approach. Despite small class sizes and a high degree of heterogeneity, no significant overfitting was observed. Feature reduction improved performance. A high proportion of features were not contributing to the classification task, which provided a strong rationale for feature reduction to reduce any noise component. Noise reduction reduces overfitting and increases generalisable potential (Ying, 2019). We examined in detail the decision-making rationale for ALVQ, and it was reassuring that the features determined to be important by the model reflected known characteristics of these diseases.

There is a suggestion that integration of machine learning in healthcare deskills practitioners (Cabitza et al., 2017). Through development of a probabilistic approach, not only do we provide an additional layer of information to clinicians by being able to

identify outlying subjects, but this also highlights that it is a tool to assist with interpretation, not a replacement.

6.2. Strengths and Limitations

The greatest strength of this work was the close collaborative working across multiple disciplines, particularly for the development of the machine learning model. This ensured our constructed output was not only novel from a computational perspective, but also biologically relevant and useful for clinicians. This supports translatability into routine clinical practice.

The main limitation of this study was the lack of a prospective test validation cohort to quantify generalisability of either method. This is an inherent limitation of rare disease research. Further strengths and limitations of this work both derive from the cohort used. We did not intend to design a whole population screening test, but we use “healthy” subjects as our disease-free class which is not representative of the negative class the test is intended for. Our class size for 21OHD was relatively small given the proportionally higher prevalence of the disease in comparison to other classes. However, when compared with existing literature, this work still represents one of the more comprehensive comparative analyses.

6.3. Translatability and Future Directions

Combining biological knowledge with advanced computational analysis has the potential to expand our knowledge of disorders of adrenal steroid synthesis. The ALVQ

method and its variations created also have the potential for translatability external to steroid metabolome research. The transparency of the model is appealing, particularly in the healthcare domain, and the intrinsic interpretability makes it highly suited for other “-omics” studies. The ability to combine the power of multiple learners while remaining computationally efficient and retaining model transparency means that the model can rival other powerful ensemble methods such as random forests.

For differential diagnosis of CAH and DSD, it is possible to combine both presented approaches. Either introduce greater domain knowledge or clinical information into the ALVQ classifier or use knowledge gained from Urine Steroid Metabolomics to construct new ratios to assist with biochemical discrimination of each disease. It is likely that improved performance can be achieved, and generalisability maximised by incorporating clinical phenotypic features and more existing domain knowledge into either presented clinical decision tool. We have shown the complexity of the early neonatal period for steroid metabolome analysis, in keeping with previously published studies (Caulfield et al., 2002; Shackleton et al., 1979). CAH and DSD commonly will present in this challenging early neonatal period. Therefore, a specific neonatal algorithm benefitting from additional recognised important steroid markers would be beneficial to develop.

The ability to visualise “adrenal age” gives physiological insight into adrenal developmental processes which may warrant further investigation. It also provides the potential for adapting this tool for treatment monitoring. Taking a cohort of pre-treated or untreated patients and a cohort of healthy controls, an optimally treated patient will sit somewhere between those two entities. If we know what the metabolome of an

optimally treated patient looks like we can then assess “distance” from this profile to quantify the degree of suboptimal treatment. This information could be used to inform treatment monitoring decisions such as dose titration or method of hydrocortisone replacement.

LC-MS/MS is much more widely available than GC-MS, is much better suited to high throughput use and sample turnaround time is rapid in comparison. Using a feature reduction approach, it is possible to extract a minimum number of the most important steroids required for differential diagnosis of these disorders. Work could then be done to move this method onto an LC-MS/MS platform, enabling preliminary automated urinary steroid profiling on a local level. For treatment monitoring, a method developed on a less costly platform, both from a monetary and a time perspective, would be advantageous.

The predictable challenge to overcome with research into inborn disorders of steroidogenesis is collection of suitable numbers of patients. Creation of translatable machine learning models for application in this healthcare domain therefore requires multi-centre international collaboration between clinicians to obtain meaningful numbers of subjects. The emergence of worldwide clinical disease networks certainly paves the way for larger scale multi-national studies. While machine learning is a powerful technique with seemingly limitless potential to improve healthcare on an individual patient level, it is important that clinicians and computer scientists also collaborate closely to ensure relevance and acceptability of any created model and its suggested outputs.

[2] *“Never trust anything that can think for itself if you can’t see where it keeps its brain.”*

J.K. Rowling, Harry Potter and the Chamber of Secrets

References

Abraham, G. E. (1974). Ovarian and adrenal contribution to peripheral androgens during the menstrual cycle. *J Clin Endocrinol Metab*, *39*(2), 340-346. https://doi.org/10.1210/jcem-39-2-340

Ackermann, D., Groessl, M., Pruijm, M., Ponte, B., Escher, G., d'Uscio, C. H., Guessous, I., Ehret, G., Pechere-Bertschi, A., Martin, P. Y., Burnier, M., Dick, B., Vogt, B., Bochud, M., Rousson, V., & Dhayat, N. A. (2019). Reference intervals for the urinary steroid metabolome: The impact of sex, age, day and night time on human adult steroidogenesis. *PLoS One*, *14*(3), e0214549. https://doi.org/10.1371/journal.pone.0214549

Adachi, M., Tachibana, K., Asakura, Y., Yamamoto, T., Hanaki, K., & Oka, A. (2004). Compound heterozygous mutations of cytochrome P450 oxidoreductase gene (POR) in two patients with Antley-Bixler syndrome. *Am J Med Genet A*, *128A*(4), 333-339. https://doi.org/10.1002/ajmg.a.30169

Akhtar, M. K., Kelly, S. L., & Kaderbhai, M. A. (2005). Cytochrome b(5) modulation of 17{alpha} hydroxylase and 17-20 lyase (CYP17) activities in steroidogenesis. *J Endocrinol*, *187*(2), 267-274. https://doi.org/10.1677/joe.1.06375

Al Alawi, A. M., Nordenstrom, A., & Falhammar, H. (2019). Clinical perspectives in congenital adrenal hyperplasia due to 3beta-hydroxysteroid dehydrogenase type 2 deficiency. *Endocrine*, *63*(3), 407-421. https://doi.org/10.1007/s12020-018-01835-3

Altman, N. (1992). The american statistician. *An introduction to kernel and nearest-neighbor nonparametric regression*, *46*(3), 175-185.

Arlt, W., Biehl, M., Taylor, A. E., Hahner, S., Libe, R., Hughes, B. A., Schneider, P., Smith, D. J., Stiekema, H., Krone, N., Porfiri, E., Opocher, G., Bertherat, J., Mantero, F., Allolio, B., Terzolo, M., Nightingale, P., Shackleton, C. H., Bertagna, X., . . . Stewart, P. M. (2011). Urine steroid metabolomics as a biomarker tool for detecting malignancy in adrenal tumors. *J Clin Endocrinol Metab*, *96*(12), 3775-3784. https://doi.org/10.1210/jc.2011-1565

Arlt, W., Callies, F., van Vlijmen, J. C., Koehler, I., Reincke, M., Bidlingmaier, M., Huebler, D., Oettel, M., Ernst, M., Schulte, H. M., & Allolio, B. (1999). Dehydroepiandrosterone replacement in women with adrenal insufficiency. *N Engl J Med*, *341*(14), 1013-1020. https://doi.org/10.1056/NEJM199909303411401

Arlt, W., Justl, H. G., Callies, F., Reincke, M., Hubler, D., Oettel, M., Ernst, M., Schulte, H. M., & Allolio, B. (1998). Oral dehydroepiandrosterone for adrenal androgen replacement: pharmacokinetics and peripheral conversion to androgens and estrogens in young healthy females after dexamethasone suppression. *J Clin Endocrinol Metab*, *83*(6), 1928-1934. https://doi.org/10.1210/jcem.83.6.4850

Arlt, W., Walker, E. A., Draper, N., Ivison, H. E., Ride, J. P., Hammer, F., Chalder, S. M., Borucka-Mankiewicz, M., Hauffa, B. P., Malunowicz, E. M., Stewart, P. M., & Shackleton, C. H. (2004). Congenital adrenal hyperplasia caused by mutant P450 oxidoreductase and human androgen synthesis: analytical study.

Lancet, *363*(9427), 2128-2135. https://doi.org/10.1016/S0140-6736(04)16503-3

Arlt, W., Willis, D. S., Wild, S. H., Krone, N., Doherty, E. J., Hahner, S., Han, T. S., Carroll, P. V., Conway, G. S., Rees, D. A., Stimson, R. H., Walker, B. R., Connell, J. M., Ross, R. J., & United Kingdom Congenital Adrenal Hyperplasia Adult Study, E. (2010). Health status of adults with congenital adrenal hyperplasia: a cohort study of 203 patients. *J Clin Endocrinol Metab*, *95*(11), 5110-5121. https://doi.org/10.1210/jc.2010-0917

Auchus, R. J. (2017). Steroid 17-hydroxylase and 17,20-lyase deficiencies, genetic and pharmacologic. *J Steroid Biochem Mol Biol*, *165*(Pt A), 71-78. https://doi.org/10.1016/j.jsbmb.2016.02.002

Auchus, R. J., Lee, T. C., & Miller, W. L. (1998). Cytochrome b5 augments the 17,20-lyase activity of human P450c17 without direct electron transfer. *J Biol Chem*, *273*(6), 3158-3165. https://doi.org/10.1074/jbc.273.6.3158

Auchus, R. J., & Rainey, W. E. (2004). Adrenarche - physiology, biochemistry and human disease. *Clin Endocrinol (Oxf)*, *60*(3), 288-296. https://doi.org/10.1046/j.1365-2265.2003.01858.x

Audi, L., Ahmed, S. F., Krone, N., Cools, M., McElreavey, K., Holterhus, P. M., Greenfield, A., Bashamboo, A., Hiort, O., Wudy, S. A., McGowan, R., & The, E. U. C. A. (2018). GENETICS IN ENDOCRINOLOGY: Approaches to molecular genetic diagnosis in the management of differences/disorders of sex development (DSD): position paper of EU COST Action BM 1303 'DSDnet'. *Eur J Endocrinol*, *179*(4), R197-R206. https://doi.org/10.1530/EJE-18-0256

Balsamo, A., Baronio, F., Ortolano, R., Menabo, S., Baldazzi, L., Di Natale, V., Vissani, S., & Cassio, A. (2020). Congenital Adrenal Hyperplasias Presenting in the Newborn and Young Infant. *Front Pediatr*, *8*, 593315. https://doi.org/10.3389/fped.2020.593315

Bancos, I., Taylor, A. E., Chortis, V., Sitch, A. J., Jenkinson, C., Davidge-Pitts, C. J., Lang, K., Tsagarakis, S., Macech, M., Riester, A., Deutschbein, T., Pupovac, I. D., Kienitz, T., Prete, A., Papathomas, T. G., Gilligan, L. C., Bancos, C., Reimondo, G., Haissaguerre, M., . . . Investigators, E. E.-A. (2020). Urine steroid metabolomics for the differential diagnosis of adrenal incidentalomas in the EURINE-ACT study: a prospective test validation study. *Lancet Diabetes Endocrinol*, *8*(9), 773-781. https://doi.org/10.1016/S2213-8587(20)30218-7

Baranowski, E. S., Arlt, W., & Idkowiak, J. (2018). Monogenic Disorders of Adrenal Steroidogenesis. *Horm Res Paediatr*, *89*(5), 292-310. https://doi.org/10.1159/000488034

Barnard, M., Mostaghel, E. A., Auchus, R. J., & Storbeck, K. H. (2020). The role of adrenal derived androgens in castration resistant prostate cancer. *J Steroid Biochem Mol Biol*, *197*, 105506. https://doi.org/10.1016/j.jsbmb.2019.105506

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82-115. https://doi.org/https://doi.org/10.1016/j.inffus.2019.12.012

Bart, A. G., & Scott, E. E. (2017). Structural and functional effects of cytochrome b5 interactions with human cytochrome P450 enzymes. *J Biol Chem*, *292*(51), 20818-20833. https://doi.org/10.1074/jbc.RA117.000220

Batista, R. L., & Mendonca, B. B. (2020). Integrative and Analytical Review of the 5-Alpha-Reductase Type 2 Deficiency Worldwide. *Appl Clin Genet*, *13*, 83-96. https://doi.org/10.2147/TACG.S198178

Beck, A. H., Sangoi, A. R., Leung, S., Marinelli, R. J., Nielsen, T. O., van de Vijver, M. J., West, R. B., van de Rijn, M., & Koller, D. (2011). Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci Transl Med*, *3*(108), 108ra113. https://doi.org/10.1126/scitranslmed.3002564

Becker, M., & Hesse, V. (2020). Minipuberty: Why Does it Happen? *Horm Res Paediatr*, *93*(2), 76-84. https://doi.org/10.1159/000508329

Beischer, N. A., Brown, J. B., Smith, M. A., & Townsend, L. (1969). Studies in prolonged pregnancy. II. Clinical results and urinary estriol excretion in prolonged pregnancy. *Am J Obstet Gynecol*, *103*(4), 483-495. https://www.ncbi.nlm.nih.gov/pubmed/5764177

Bernstein, R. M., Sterner, K. N., & Wildman, D. E. (2012). Adrenal androgen production in catarrhine primates and the evolution of adrenarche. *Am J Phys Anthropol*, *147*(3), 389-400. https://doi.org/10.1002/ajpa.22001

Berrar, D. (2019). Cross-Validation. In.

Bertholf, R. L. (2006). Statistical methods for establishing and validating reference intervals. *Laboratory Medicine*, *37*(5), 306-310.

Biehl, M., Ghosh, A., & Hammer, B. (2007). Dynamics and Generalization Ability of LVQ Algorithms. *Journal of machine learning research*, *8*(2).

Biehl, M., Hammer, B., & Villmann, T. (2013). Distance measures for prototype based classification. International Workshop on Brain-Inspired Computing,

Biehl, M., Hammer, B., & Villmann, T. (2016). Prototype-based models in machine learning. *Wiley Interdisciplinary Reviews: Cognitive Science*, *7*(2), 92-111.

Biehl, M., Schneider, P., Smith, D., Stiekema, H., Taylor, A., Hughes, B., Shackleton, C., Stewart, P., & Arlt, W. (2012). Matrix relevance LVQ in steroid metabolomics based classification of adrenal tumors. ESANN,

Biglieri, E. G., Herron, M. A., & Brust, N. (1966). 17-hydroxylation deficiency in man. *J Clin Invest*, *45*(12), 1946-1954. https://doi.org/10.1172/JCI105499

Blacklaws, C. (2018). Algorithms: transparency and accountability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2128), 20170351.

Bocian-Sobkowska, J. (2000). Morphometric study of the human suprarenal gland in the first postnatal year. *Folia Morphol (Warsz)*, *58*(4), 275-284. https://www.ncbi.nlm.nih.gov/pubmed/11000884

Bongiovanni, A. M. (1962). The adrenogenital syndrome with deficiency of 3 beta-hydroxysteroid dehydrogenase. *J Clin Invest*, *41*, 2086-2092. https://doi.org/10.1172/JCI104666

Bongiovanni, A. M., & Eberlein, W. R. (1956). Plasma and urinary corticosteroids in the hypertensive form of congenital adrenal hyperplasia. *J Biol Chem*, *223*(1), 85-94. https://www.ncbi.nlm.nih.gov/pubmed/13376579

Bose, H. S., Sugawara, T., Strauss, J. F., 3rd, Miller, W. L., & International Congenital Lipoid Adrenal Hyperplasia, C. (1996). The pathophysiology and genetics of congenital lipoid adrenal hyperplasia. *N Engl J Med*, *335*(25), 1870-1878. https://doi.org/10.1056/NEJM199612193352503

Bose, H. S., Whittal, R. M., Huang, M. C., Baldwin, M. A., & Miller, W. L. (2000). N-218 MLN64, a protein with StAR-like steroidogenic activity, is folded and

cleaved similarly to StAR. *Biochemistry*, *39*(38), 11722-11731. https://doi.org/10.1021/bi000911l

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L., Lijmer, J. G., Moher, D., Rennie, D., de Vet, H. C., Kressel, H. Y., Rifai, N., Golub, R. M., Altman, D. G., Hooft, L., Korevaar, D. A., Cohen, J. F., & Group, S. (2015). STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ*, *351*, h5527. https://doi.org/10.1136/bmj.h5527

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140.

Bulsari, K., & Falhammar, H. (2017). Clinical perspectives in congenital adrenal hyperplasia due to 11beta-hydroxylase deficiency. *Endocrine*, *55*(1), 19-36. https://doi.org/10.1007/s12020-016-1189-x

Bunte, K., Baranowski, E. S., Arlt, W., & Tino, P. (2016). Relevance learning vector quantization in variable dimensional spaces. Workshop of the GI-Fachgruppe Neuronale Netze and the German Neural Networks Society in connection to GCPR 2016,

Burkhard, F. Z., Parween, S., Udhane, S. S., Fluck, C. E., & Pandey, A. V. (2017). P450 Oxidoreductase deficiency: Analysis of mutations and polymorphisms. *J Steroid Biochem Mol Biol*, *165*(Pt A), 38-50. https://doi.org/10.1016/j.jsbmb.2016.04.003

Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended Consequences of Machine Learning in Medicine. *JAMA*, *318*(6), 517-518. https://doi.org/10.1001/jama.2017.7797

Carlson, P. B., J.R.; Engel, K.; Fried, J.; Kircher, H.W.; Loaning, K.L.; Moss, G.P.; Popják, G.; Uskokovic, M.R. (1989). IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN). The nomenclature of steroids. Recommendations 1989. *Eur J Biochem*, *186*(3), 429-458. https://www.ncbi.nlm.nih.gov/pubmed/2606099

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). *Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission* Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia. https://doi.org/10.1145/2783258.2788613

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, *8*(8), 832. https://www.mdpi.com/2079-9292/8/8/832

Caulfield, M. P., Lynn, T., Gottschalk, M. E., Jones, K. L., Taylor, N. F., Malunowicz, E. M., Shackleton, C. H., Reitz, R. E., & Fisher, D. A. (2002). The diagnosis of congenital adrenal hyperplasia in the newborn by gas chromatography/mass spectrometry analysis of random urine specimens. *J Clin Endocrinol Metab*, *87*(8), 3682-3690. https://doi.org/10.1210/jcem.87.8.8712

Chan, A. O., Taylor, N. F., Tiu, S. C., & Shek, C. C. (2008). Reference intervals of urinary steroid metabolites using gas chromatography-mass spectrometry in Chinese adults. *Steroids*, *73*(8), 828-837. https://doi.org/10.1016/j.steroids.2008.03.004

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

Chortis, V., Bancos, I., Nijman, T., Gilligan, L. C., Taylor, A. E., Ronchi, C. L., O'Reilly, M. W., Schreiner, J., Asia, M., Riester, A., Perotti, P., Libe, R.,

Quinkler, M., Canu, L., Paiva, I., Bugalho, M. J., Kastelan, D., Dennedy, M. C., Sherlock, M., . . . Arlt, W. (2020). Urine Steroid Metabolomics as a Novel Tool for Detection of Recurrent Adrenocortical Carcinoma. *J Clin Endocrinol Metab*, *105*(3). https://doi.org/10.1210/clinem/dgz141

Chung, B. C., Matteson, K. J., Voutilainen, R., Mohandas, T. K., & Miller, W. L. (1986). Human cholesterol side-chain cleavage enzyme, P450scc: cDNA cloning, assignment of the gene to chromosome 15, and expression in the placenta. *Proc Natl Acad Sci U S A*, *83*(23), 8962-8966. https://doi.org/10.1073/pnas.83.23.8962

Chung, B. C., Picado-Leonard, J., Haniu, M., Bienkowski, M., Hall, P. F., Shively, J. E., & Miller, W. L. (1987). Cytochrome P450c17 (steroid 17 alpha-hydroxylase/17,20 lyase): cloning of human adrenal and testis cDNAs indicates the same gene is expressed in both tissues. *Proc Natl Acad Sci U S A*, *84*(2), 407-411. https://doi.org/10.1073/pnas.84.2.407

Cios, K. J., & Moore, G. W. (2002). Uniqueness of medical data mining. *Artif Intell Med*, *26*(1-2), 1-24. https://doi.org/10.1016/s0933-3657(02)00049-0

Claahsen-van der Grinten, H. L., Speiser, P. W., Ahmed, S. F., Arlt, W., Auchus, R. J., Falhammar, H., Fluck, C. E., Guasti, L., Huebner, A., Kortmann, B. B. M., Krone, N., Merke, D. P., Miller, W. L., Nordenstrom, A., Reisch, N., Sandberg, D. E., Stikkelbroeck, N., Touraine, P., Utari, A., . . . White, P. C. (2022). Congenital Adrenal Hyperplasia-Current Insights in Pathophysiology, Diagnostics, and Management. *Endocr Rev*, *43*(1), 91-159. https://doi.org/10.1210/endrev/bnab016

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, *26*(4), 404-413.

Cohen, J. F., Korevaar, D. A., Altman, D. G., Bruns, D. E., Gatsonis, C. A., Hooft, L., Irwig, L., Levine, D., Reitsma, J. B., de Vet, H. C., & Bossuyt, P. M. (2016). STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open*, *6*(11), e012799. https://doi.org/10.1136/bmjopen-2016-012799

Costa-Santos, M., Kater, C. E., Auchus, R. J., & Brazilian Congenital Adrenal Hyperplasia Multicenter Study, G. (2004). Two prevalent CYP17 mutations and genotype-phenotype correlations in 24 Brazilian patients with 17-hydroxylase deficiency. *J Clin Endocrinol Metab*, *89*(1), 49-60. https://doi.org/10.1210/jc.2003-031021

Coulter, C. L., & Jaffe, R. B. (1998). Functional maturation of the primate fetal adrenal in vivo: 3. Specific zonal localization and developmental regulation of CYP21A2 (P450c21) and CYP11B1/CYP11B2 (P450c11/aldosterone synthase) lead to integrated concept of zonal and temporal steroid biosynthesis. *Endocrinology*, *139*(12), 5144-5150. https://doi.org/10.1210/endo.139.12.6333

Cruz-Topete, D., & Cidlowski, J. A. (2015). One hormone, two actions: anti- and pro-inflammatory effects of glucocorticoids. *Neuroimmunomodulation*, *22*(1-2), 20-32. https://doi.org/10.1159/000362724

Davio, A., Woolcock, H., Nanba, A. T., Rege, J., O'Day, P., Ren, J., Zhao, L., Ebina, H., Auchus, R., Rainey, W. E., & Turcu, A. F. (2020). Sex Differences in 11-Oxygenated Androgen Patterns Across Adulthood. *J Clin Endocrinol Metab*, *105*(8). https://doi.org/10.1210/clinem/dgaa343

de Jong, W. H., Buitenwerf, E., Pranger, A. T., Riphagen, I. J., Wolffenbuttel, B. H., Kerstens, M. N., & Kema, I. P. (2018). Determination of reference intervals for

urinary steroid profiling using a newly validated GC-MS/MS method. *Clinical Chemistry and Laboratory Medicine (CCLM)*, *56*(1), 103-112.

Deo, R. C. (2015). Machine Learning in Medicine. *Circulation*, *132*(20), 1920-1930. https://doi.org/10.1161/CIRCULATIONAHA.115.001593

Derks, H. J., & Drayer, N. M. (1978). Polar corticosteroids in human neonatal urine; synthesis and gas chromatography-mass spectrometry of ring A reduced 6-hydroxylated corticosteroids. *Steroids*, *31*(1), 9-22. https://doi.org/10.1016/0039-128x(78)90016-8

Dhayat, N. A., Frey, A. C., Frey, B. M., d'Uscio, C. H., Vogt, B., Rousson, V., Dick, B., & Fluck, C. E. (2015). Estimation of reference curves for the urinary steroid metabolome in the first year of life in healthy children: Tracing the complexity of human postnatal steroidogenesis. *J Steroid Biochem Mol Biol*, *154*, 226-236. https://doi.org/10.1016/j.jsbmb.2015.07.024

Diczfalusy, E. (1964). Endocrine Functions of the Human Fetoplacental Unit. *Fed Proc*, *23*, 791-798. https://www.ncbi.nlm.nih.gov/pubmed/14195464

Dietterich, T. G. (2002). Ensemble learning. *The handbook of brain theory and neural networks*, *2*(1), 110-125.

Duggal, R., Liu, Y., Gregory, M. C., Denisov, I. G., Kincaid, J. R., & Sligar, S. G. (2016). Evidence that cytochrome b5 acts as a redox donor in CYP17A1 mediated androgen synthesis. *Biochem Biophys Res Commun*, *477*(2), 202-208. https://doi.org/10.1016/j.bbrc.2016.06.043

Eisenhofer, G., Peitzsch, M., Kaden, D., Langton, K., Pamporaki, C., Masjkur, J., Tsatsaronis, G., Mangelis, A., Williams, T. A., Reincke, M., Lenders, J. W. M., & Bornstein, S. R. (2017). Reference intervals for plasma concentrations of adrenal steroids measured by LC-MS/MS: Impact of gender, age, oral contraceptives, body mass index and blood pressure status. *Clin Chim Acta*, *470*, 115-124. https://doi.org/10.1016/j.cca.2017.05.002

El-Maouche, D., Arlt, W., & Merke, D. P. (2017). Congenital adrenal hyperplasia. *Lancet*, *390*(10108), 2194-2210. https://doi.org/10.1016/S0140-6736(17)31431-9

Engels, M., Gehrmann, K., Falhammar, H., Webb, E. A., Nordenstrom, A., Sweep, F. C., Span, P. N., van Herwaarden, A. E., Rohayem, J., Richter-Unruh, A., Bouvattier, C., Kohler, B., Kortmann, B. B., Arlt, W., Roeleveld, N., Reisch, N., Stikkelbroeck, N., Claahsen-van der Grinten, H. L., & dsd, L. g. (2018). Gonadal function in adult male patients with congenital adrenal hyperplasia. *Eur J Endocrinol*, *178*(3), 285-294. https://doi.org/10.1530/EJE-17-0862

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115-118. https://doi.org/10.1038/nature21056

Estrada, D. F., Skinner, A. L., Laurence, J. S., & Scott, E. E. (2014). Human cytochrome P450 17A1 conformational selection: modulation by ligand and cytochrome b5. *J Biol Chem*, *289*(20), 14310-14320. https://doi.org/10.1074/jbc.M114.560144

Falhammar, H., Frisen, L., Norrby, C., Hirschberg, A. L., Almqvist, C., Nordenskjold, A., & Nordenstrom, A. (2014). Increased mortality in patients with congenital adrenal hyperplasia due to 21-hydroxylase deficiency. *J Clin Endocrinol Metab*, *99*(12), E2715-2721. https://doi.org/10.1210/jc.2014-2957

Fanelli, F., Cantu, M., Temchenko, A., Mezzullo, M., Lindner, J. M., Peitzsch, M., Hawley, J. M., Bruce, S., Binz, P. A., Ackermans, M. T., Heijboer, A. C., Van den Ouweland, J., Koeppl, D., Nardi, E., MacKenzie, F., Rauh, M.,

Eisenhofer, G., Keevil, B. G., Vogeser, M., & Pagotto, U. (2022). Report from the HarmoSter study: impact of calibration on comparability of LC-MS/MS measurement of circulating cortisol, 17OH-progesterone and aldosterone. *Clin Chem Lab Med*, *60*(5), 726-739. https://doi.org/10.1515/cclm-2021-1028

Feldman, H. A., Longcope, C., Derby, C. A., Johannes, C. B., Araujo, A. B., Coviello, A. D., Bremner, W. J., & McKinlay, J. B. (2002). Age trends in the level of serum testosterone and other hormones in middle-aged men: longitudinal results from the Massachusetts male aging study. *J Clin Endocrinol Metab*, *87*(2), 589-598. https://doi.org/10.1210/jcem.87.2.8201

Ferdinandy, B., Gerencsér, L., Corrieri, L., Perez, P., Újváry, D., Csizmadia, G., & Miklósi, Á. (2020). Challenges of machine learning model validation using correlated behaviour data: evaluation of cross-validation strategies and accuracy measures. *PLoS One*, *15*(7), e0236092.

Fluck, C. E., Miller, W. L., & Auchus, R. J. (2003). The 17, 20-lyase activity of cytochrome p450c17 from human fetal testis favors the delta5 steroidogenic pathway. *J Clin Endocrinol Metab*, *88*(8), 3762-3766. https://doi.org/10.1210/jc.2003-030143

Fluck, C. E., Tajima, T., Pandey, A. V., Arlt, W., Okuhara, K., Verge, C. F., Jabs, E. W., Mendonca, B. B., Fujieda, K., & Miller, W. L. (2004). Mutant P450 oxidoreductase causes disordered steroidogenesis with and without Antley-Bixler syndrome. *Nat Genet*, *36*(3), 228-230. https://doi.org/10.1038/ng1300

Forest, M. G. (1983). Role of androgens in fetal and pubertal development. *Horm Res*, *18*(1-3), 69-83. https://doi.org/10.1159/000179780

Forest, M. G., Cathiard, A. M., & Bertrand, J. A. (1973). Evidence of testicular activity in early infancy. *J Clin Endocrinol Metab*, *37*(1), 148-151. https://doi.org/10.1210/jcem-37-1-148

Fukami, M., Homma, K., Hasegawa, T., & Ogata, T. (2013). Backdoor pathway for dihydrotestosterone biosynthesis: implications for normal and abnormal human sex development. *Dev Dyn*, *242*(4), 320-329. https://doi.org/10.1002/dvdy.23892

Fukami, M., Nishimura, G., Homma, K., Nagai, T., Hanaki, K., Uematsu, A., Ishii, T., Numakura, C., Sawada, H., Nakacho, M., Kowase, T., Motomura, K., Haruna, H., Nakamura, M., Ohishi, A., Adachi, M., Tajima, T., Hasegawa, Y., Hasegawa, T., . . . Ogata, T. (2009). Cytochrome P450 oxidoreductase deficiency: identification and characterization of biallelic mutations and genotype-phenotype correlations in 35 Japanese patients. *J Clin Endocrinol Metab*, *94*(5), 1723-1731. https://doi.org/10.1210/jc.2008-2816

Fuller, P. J., & Young, M. J. (2005). Mechanisms of mineralocorticoid action. *Hypertension*, *46*(6), 1227-1235. https://doi.org/10.1161/01.HYP.0000193502.77417.17

Garagorri, J. M., Rodriguez, G., Lario-Elboj, A. J., Olivares, J. L., Lario-Munoz, A., & Orden, I. (2008). Reference levels for 17-hydroxyprogesterone, 11-desoxycortisol, cortisol, testosterone, dehydroepiandrosterone sulfate and androstenedione in infants from birth to six months of age. *Eur J Pediatr*, *167*(6), 647-653. https://doi.org/10.1007/s00431-007-0565-1

García-Laencina, P. J., Sancho-Gómez, J.-L., & Figueiras-Vidal, A. R. (2010). Pattern classification with missing data: a review. *Neural Computing and Applications*, *19*(2), 263-282. https://doi.org/10.1007/s00521-009-0295-6

Geissler, W. M., Davis, D. L., Wu, L., Bradshaw, K. D., Patel, S., Mendonca, B. B., Elliston, K. O., Wilson, J. D., Russell, D. W., & Andersson, S. (1994). Male

pseudohermaphroditism caused by mutations of testicular 17 beta-hydroxysteroid dehydrogenase 3. *Nat Genet*, *7*(1), 34-39. https://doi.org/10.1038/ng0594-34

Gell, J. S., Atkins, B., Margraf, L., Mason, J. I., Sasano, H., Rainey, W. E., & Carr, B. R. (1996). Adrenarche is associated with decreased 3 beta-hydroxysteroid dehydrogenase expression in the adrenal reticularis. *Endocr Res*, *22*(4), 723-728. https://doi.org/10.1080/07435809609043768

Geller, D. H., Auchus, R. J., Mendonca, B. B., & Miller, W. L. (1997). The genetic and functional basis of isolated 17,20-lyase deficiency. *Nat Genet*, *17*(2), 201-205. https://doi.org/10.1038/ng1097-201

Ghosh, S., Baranowski, E. S., Biehl, M., Arlt, W., Tino, P., & Bunte, K. (2022). Interpretable Models Capable of Handling Systematic Missingness in Imbalanced Classes and Heterogeneous Datasets. arXiv:2206.02056. Retrieved June 01, 2022, from https://ui.adsabs.harvard.edu/abs/2022arXiv220602056G

Ghosh, S., Baranowski, E. S., van Veen, R., de Vries, G.-J., Biehl, M., Arlt, W., Tino, P., & Bunte, K. (2017). Comparison of strategies to learn from imbalanced classes for computer aided diagnosis of inborn steroidogenic disorders. ESANN,

Ghosh, S., Tino, P., & Bunte, K. (2020). Visualisation and knowledge discovery from interpretable models. 2020 International Joint Conference on Neural Networks (IJCNN),

Ghosh, S. S. (2021). *Intrinsically Interpretable Machine Learning In Computer Aided Diagnosis* University of Groningen].

Giordano, S. J., Kaftory, A., & Steggles, A. W. (1994). A splicing mutation in the cytochrome b5 gene from a patient with congenital methemoglobinemia and pseudohermaphrodism. *Hum Genet*, *93*(5), 568-570. https://doi.org/10.1007/BF00202825

Goto, M., Piper Hanley, K., Marcos, J., Wood, P. J., Wright, S., Postle, A. D., Cameron, I. T., Mason, J. I., Wilson, D. I., & Hanley, N. A. (2006). In humans, early cortisol biosynthesis provides a mechanism to safeguard female sexual development. *J Clin Invest*, *116*(4), 953-960. https://doi.org/10.1172/JCI25091

Greaves, R. F. (2017). The central role of external quality assurance in harmonisation and standardisation for laboratory medicine. *Clin Chem Lab Med*, *55*(4), 471-473. https://doi.org/10.1515/cclm-2016-0782

Grueters, A., & Korth-Schutz, S. (1982). Longitudinal study of plasma dehydroepiandrosterone sulfate in preterm and fullterm infants. *J Clin Endocrinol Metab*, *55*(2), 314-320. https://doi.org/10.1210/jcem-55-2-314

Guerami, A., Varner, M. W., Shackleton, C. H., MacDonald, P. C., & Casey, M. L. (1988). Origin of deoxycorticosterone and deoxycorticosterone sulfate in human pregnancy: absence of steroid 21-sulfatase activity in sulfatase-deficient placenta. *J Steroid Biochem*, *29*(1), 57-62. https://doi.org/10.1016/0022-4731(88)90376-7

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, *316*(22), 2402-2410. https://doi.org/10.1001/jama.2016.17216

Gwynne, J. T., & Strauss, J. F., 3rd. (1982). The role of lipoproteins in steroidogenesis and cholesterol metabolism in steroidogenic glands. *Endocr Rev*, *3*(3), 299-329. https://doi.org/10.1210/edrv-3-3-299

Hammer, B., Mokbel, B., Schleif, F.-M., & Zhu, X. (2011). Prototype-based classification of dissimilarity data. International Symposium on Intelligent Data Analysis,

Hammer, B., & Villmann, T. (2002). Generalized relevance learning vector quantization. *Neural Networks*, *15*(8-9), 1059-1068.

Haning, R. V., Jr., Carlson, I. H., Flood, C. A., Hackett, R. J., & Longcope, C. (1991). Metabolism of dehydroepiandrosterone sulfate (DS) in normal women and women with high DS concentrations. *J Clin Endocrinol Metab*, *73*(6), 1210-1215. https://doi.org/10.1210/jcem-73-6-1210

Hargitai, G., Solyom, J., Battelino, T., Lebl, J., Pribilincova, Z., Hauspie, R., Kovacs, J., Waldhauser, F., Frisch, H., & Group, M.-C. S. (2001). Growth patterns and final height in congenital adrenal hyperplasia due to classical 21-hydroxylase deficiency. Results of a multicenter study. *Horm Res*, *55*(4), 161-171. https://doi.org/10.1159/000049990

Hata, K., Hata, T., & Kitao, M. (1988). Ultrasonographic identification and measurement of the human fetal adrenal gland in utero: clinical application. *Gynecol Obstet Invest*, *25*(1), 16-22. https://doi.org/10.1159/000293739

Hata, K., Nagata, H., Nishigaki, A., Aoki, S., Hata, T., Murao, F., & Kitao, M. (1988). Ultrasonographic evaluation of adrenal involution during antenatal and neonatal periods. *Gynecol Obstet Invest*, *26*(1), 29-32. https://doi.org/10.1159/000293668

Hegesh, E., Hegesh, J., & Kaftory, A. (1986). Congenital methemoglobinemia with a deficiency of cytochrome b5. *N Engl J Med*, *314*(12), 757-761. https://doi.org/10.1056/NEJM198603203141206

Held, P. K., Bialk, E. R., Lasarev, M. R., & Allen, D. B. (2022). 21-Deoxycortisol is a Key Screening Marker for 21-Hydroxylase Deficiency. *J Pediatr*, *242*, 213-219 e211. https://doi.org/10.1016/j.jpeds.2021.10.063

Hey, T., Butler, K., Jackson, S., & Thiyagalingam, J. (2020). Machine learning and big scientific data. *Philosophical Transactions of the Royal Society A*, *378*(2166), 20190054.

Hill, M., Parizek, A., Cibula, D., Kancheva, R., Jirasek, J. E., Jirkovska, M., Velikova, M., Kubatova, J., Klimkova, M., Paskova, A., Zizka, Z., Kancheva, L., Kazihnitkova, H., Zamrazilova, L., & Starka, L. (2010). Steroid metabolome in fetal and maternal body fluids in human late pregnancy. *J Steroid Biochem Mol Biol*, *122*(4), 114-132. https://doi.org/10.1016/j.jsbmb.2010.05.007

Hill, M., Paskova, A., Kanceva, R., Velikova, M., Kubatova, J., Kancheva, L., Adamcova, K., Mikesova, M., Zizka, Z., Koucky, M., Sarapatkova, H., Kacer, V., Matucha, P., Meloun, M., & Parizek, A. (2014). Steroid profiling in pregnancy: a focus on the human fetus. *J Steroid Biochem Mol Biol*, *139*, 201-222. https://doi.org/10.1016/j.jsbmb.2013.03.008

Hogan, H., Healey, F., Neale, G., Thomson, R., Vincent, C., & Black, N. (2012). Preventable deaths due to problems in care in English acute hospitals: a retrospective case record review study. *BMJ Qual Saf*, *21*(9), 737-745. https://doi.org/10.1136/bmjqs-2011-001159

Holcombe, J. H., Keenan, B. S., Nichols, B. L., Kirkland, R. T., & Clayton, G. W. (1980). Neonatal salt loss in the hypertensive form of congenital adrenal

hyperplasia. *Pediatrics*, *65*(4), 777-781. https://www.ncbi.nlm.nih.gov/pubmed/6966049

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery*, *9*(4), e1312. https://doi.org/https://doi.org/10.1002/widm.1312

Homma, K., Hasegawa, T., Masumoto, M., Takeshita, E., Watanabe, K., Chiba, H., Kurosawa, T., Takahashi, T., & Matsuo, N. (2003). Reference values for urinary steroids in Japanese newborn infants: gas chromatography/mass spectrometry in selected ion monitoring. *Endocr J*, *50*(6), 783-792. https://doi.org/10.1507/endocrj.50.783

Homoki, J., & Teller, W. M. (1982). Increased urinary excretion of total 16 alpha-hydroxypregnenolone in newborn infants with congenital adrenal hyperplasia due to 21-hydroxylase deficiency. *Klin Wochenschr*, *60*(8), 407-410. https://doi.org/10.1007/BF01735932

Honour, J., & Brook, C. (1997). Clinical indications for the use of urinary steroid profiles in neonates and children. *Annals of clinical biochemistry*, *34*(1), 45-54.

Honour, J. W. (2018). Urinary Steroid Profiling in the Diagnosis of Adrenal Disorders. *Clin Chem*, *64*(8), 1257-1258. https://doi.org/10.1373/clinchem.2017.285890

Honour, J. W., Conway, E., Hodkinson, R., & Lam, F. (2018). The evolution of methods for urinary steroid metabolomics in clinical investigations particularly in childhood. *J Steroid Biochem Mol Biol*, *181*, 28-51. https://doi.org/10.1016/j.jsbmb.2018.02.013

Honour, J. W., & Rumsby, G. (1993). Problems in diagnosis and management of congenital adrenal hyperplasia due to 21-hydroxylase deficiency. *J Steroid Biochem Mol Biol*, *45*(1-3), 69-74. https://doi.org/10.1016/0960-0760(93)90124-f

Honour, J. W., Tourniaire, J., Biglieri, E. G., & Shackleton, C. H. (1978). Urinary steroid excretion in 17 alpha-hydroxylase deficiency. *J Steroid Biochem*, *9*(6), 495-505. https://doi.org/10.1016/0022-4731(78)90115-2

Idkowiak, J., Lavery, G. G., Dhir, V., Barrett, T. G., Stewart, P. M., Krone, N., & Arlt, W. (2011). Premature adrenarche: novel lessons from early onset androgen excess. *Eur J Endocrinol*, *165*(2), 189-207. https://doi.org/10.1530/EJE-11-0223

Idkowiak, J., O'Riordan, S., Reisch, N., Malunowicz, E. M., Collins, F., Kerstens, M. N., Kohler, B., Graul-Neumann, L. M., Szarras-Czapnik, M., Dattani, M., Silink, M., Shackleton, C. H., Maiter, D., Krone, N., & Arlt, W. (2011). Pubertal presentation in seven patients with congenital adrenal hyperplasia due to P450 oxidoreductase deficiency. *J Clin Endocrinol Metab*, *96*(3), E453-462. https://doi.org/10.1210/jc.2010-1607

Idkowiak, J., Randell, T., Dhir, V., Patel, P., Shackleton, C. H., Taylor, N. F., Krone, N., & Arlt, W. (2012). A missense mutation in the human cytochrome b5 gene causes 46,XY disorder of sex development due to true isolated 17,20 lyase deficiency. *J Clin Endocrinol Metab*, *97*(3), E465-475. https://doi.org/10.1210/jc.2011-2413

Imperato-McGinley, J., Gautier, T., Peterson, R. E., & Shackleton, C. (1986). The prevalence of 5 alpha-reductase deficiency in children with ambiguous genitalia in the Dominican Republic. *J Urol*, *136*(4), 867-873. https://doi.org/10.1016/s0022-5347(17)45108-1

Imperato-McGinley, J., Gautier, T., Pichardo, M., & Shackleton, C. (1986). The diagnosis of 5 alpha-reductase deficiency in infancy. *J Clin Endocrinol Metab*, *63*(6), 1313-1318. https://doi.org/10.1210/jcem-63-6-1313

Imperato-McGinley, J., Guerrero, L., Gautier, T., & Peterson, R. E. (1974). Steroid 5alpha-reductase deficiency in man: an inherited form of male pseudohermaphroditism. *Science*, *186*(4170), 1213-1215. https://doi.org/10.1126/science.186.4170.1213

Ipsos, M. (2017). Public views of machine learning. *Report title*, *92*.

Ishimoto, H., & Jaffe, R. B. (2011). Development and function of the human fetal adrenal cortex: a key component in the feto-placental unit. *Endocr Rev*, *32*(3), 317-355. https://doi.org/10.1210/er.2010-0001

Jaeken, J., Lefeber, D. J., & Matthijs, G. (2020). SRD5A3 defective congenital disorder of glycosylation: clinical utility gene card. *Eur J Hum Genet*, *28*(9), 1297-1300. https://doi.org/10.1038/s41431-020-0647-3

Jha, S., Turcu, A. F., Sinaii, N., Brookner, B., Auchus, R. J., & Merke, D. P. (2021). 11-Oxygenated Androgens Useful in the Setting of Discrepant Conventional Biomarkers in 21-Hydroxylase Deficiency. *J Endocr Soc*, *5*(2), bvaa192. https://doi.org/10.1210/jendso/bvaa192

Joannou, G. E. (1981). Identification of 15 beta-hydroxylated C21 steroids in the neo-natal period: the role of 3 alpha,15 beta,17 alpha-trihydroxy-5 beta-pregnan-20-one in the perinatal diagnosis of congenital adrenal hyperplasia (CAH) due to a 21-hydroxylase deficiency. *J Steroid Biochem*, *14*(9), 901-912. https://doi.org/10.1016/0022-4731(81)90239-9

Joehrer, K., Geley, S., Strasser-Wozak, E. M., Azziz, R., Wollmann, H. A., Schmitt, K., Kofler, R., & White, P. C. (1997). CYP11B1 mutations causing non-classic adrenal hyperplasia due to 11 beta-hydroxylase deficiency. *Hum Mol Genet*, *6*(11), 1829-1834. https://doi.org/10.1093/hmg/6.11.1829

Johnston, Z. C., Bellingham, M., Filis, P., Soffientini, U., Hough, D., Bhattacharya, S., Simard, M., Hammond, G. L., King, P., O'Shaughnessy, P. J., & Fowler, P. A. (2018). The human fetal adrenal produces cortisol but no detectable aldosterone throughout the second trimester. *BMC Med*, *16*(1), 23. https://doi.org/10.1186/s12916-018-1009-7

Jones, C. M., Mallappa, A., Reisch, N., Nikolaou, N., Krone, N., Hughes, B. A., O'Neil, D. M., Whitaker, M. J., Tomlinson, J. W., Storbeck, K. H., Merke, D. P., Ross, R. J., & Arlt, W. (2017). Modified-Release and Conventional Glucocorticoids and Diurnal Androgen Excretion in Congenital Adrenal Hyperplasia. *J Clin Endocrinol Metab*, *102*(6), 1797-1806. https://doi.org/10.1210/jc.2016-2855

Jones, G., & Barker, A. (2008). Reference intervals. *Clin Biochem Rev*, *29 Suppl 1*, S93-97. https://www.ncbi.nlm.nih.gov/pubmed/18852866

Judd, H. L., Judd, G. E., Lucas, W. E., & Yen, S. S. (1974). Endocrine function of the postmenopausal ovary: concentration of androgens and estrogens in ovarian and peripheral vein blood. *J Clin Endocrinol Metab*, *39*(6), 1020-1024. https://doi.org/10.1210/jcem-39-6-1020

Kaden, M., Lange, M., Nebel, D., Riedel, M., Geweniger, T., & Villmann, T. (2014). Aspects in Classification Learning - Review of Recent Developments in Learning Vector Quantization. *Foundations of Computing and Decision Sciences*, *39*(2), 79-105. https://doi.org/doi:10.2478/fcds-2014-0006

Kamrath, C., Hartmann, M. F., Pons-Kuhnemann, J., & Wudy, S. A. (2020). Urinary GC-MS steroid metabotyping in treated children with congenital adrenal

hyperplasia. *Metabolism*, *112*, 154354. https://doi.org/10.1016/j.metabol.2020.154354

Kamrath, C., Hochberg, Z., Hartmann, M. F., Remer, T., & Wudy, S. A. (2012). Increased activation of the alternative "backdoor" pathway in patients with 21-hydroxylase deficiency: evidence from urinary steroid hormone analysis. *J Clin Endocrinol Metab*, *97*(3), E367-375. https://doi.org/10.1210/jc.2011-1997

Kamrath, C., Wettstaedt, L., Boettcher, C., Hartmann, M. F., & Wudy, S. A. (2017). The urinary steroidome of treated children with classic 21-hydroxylase deficiency. *J Steroid Biochem Mol Biol*, *165*(Pt B), 396-406. https://doi.org/10.1016/j.jsbmb.2016.08.006

Katayev, A., Balciza, C., & Seccombe, D. W. (2010). Establishing reference intervals for clinical laboratory test results: is there a better way? *Am J Clin Pathol*, *133*(2), 180-186. https://doi.org/10.1309/AJCPN5BMTSF1CDYP

Kawamoto, T., Mitsuuchi, Y., Toda, K., Yokoyama, Y., Miyahara, K., Miura, S., Ohnishi, T., Ichikawa, Y., Nakao, K., Imura, H., & et al. (1992). Role of steroid 11 beta-hydroxylase and steroid 18-hydroxylase in the biosynthesis of glucocorticoids and mineralocorticoids in humans. *Proc Natl Acad Sci U S A*, *89*(4), 1458-1462. https://doi.org/10.1073/pnas.89.4.1458

Keene, M. F. (1927). Observations on the Development of the Human Suprarenal Gland. *J Anat*, *61*(Pt 3), 302-324. https://www.ncbi.nlm.nih.gov/pubmed/17104143

Kelnar, C. J., & Brook, C. G. (1983). A mixed longitudinal study of adrenal steroid excretion in childhood and the mechanism of adrenarche. *Clin Endocrinol (Oxf)*, *19*(1), 117-129. https://doi.org/10.1111/j.1365-2265.1983.tb00750.x

Kim, J. H., Lee, Y. A., Lim, Y. H., Lee, K., Kim, B. N., Kim, J. I., Hong, Y. C., Yang, S. W., Song, J., & Shin, C. H. (2020). Changes in Adrenal Androgens and Steroidogenic Enzyme Activities From Ages 2, 4, to 6 Years: A Prospective Cohort Study. *J Clin Endocrinol Metab*, *105*(10). https://doi.org/10.1210/clinem/dgaa498

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464-1480.

Kojima, S., Yanaihara, T., & Nakayama, T. (1981). Serum steroid levels in children at birth and in early neonatal period. *Am J Obstet Gynecol*, *140*(8), 961-965. https://doi.org/10.1016/0002-9378(81)90092-2

Kok, R. C., Timmerman, M. A., Wolffenbuttel, K. P., Drop, S. L., & de Jong, F. H. (2010). Isolated 17,20-lyase deficiency due to the cytochrome b5 mutation W27X. *J Clin Endocrinol Metab*, *95*(3), 994-999. https://doi.org/10.1210/jc.2008-1745

Koyama, Y., Homma, K., Fukami, M., Miwa, M., Ikeda, K., Ogata, T., Hasegawa, T., & Murata, M. (2012). Two-step biochemical differential diagnosis of classic 21-hydroxylase deficiency and cytochrome P450 oxidoreductase deficiency in Japanese infants by GC-MS measurement of urinary pregnanetriolone/ tetrahydroxycortisone ratio and 11β-hydroxyandrosterone. *Clin Chem*, *58*(4), 741-747. https://doi.org/10.1373/clinchem.2011.173286

Koyama, Y., Homma, K., Fukami, M., Miwa, M., Ikeda, K., Ogata, T., Murata, M., & Hasegawa, T. (2016). Classic and non-classic 21-hydroxylase deficiency can be discriminated from P450 oxidoreductase deficiency in Japanese infants by urinary steroid metabolites. *Clin Pediatr Endocrinol*, *25*(2), 37-44. https://doi.org/10.1297/cpe.25.37

Kraan, G. P., Wolthers, B. G., van der Molen, J. C., Nagel, G. T., Drayer, N. M., & Joannou, G. E. (1993). New identified 15 beta-hydroxylated 21-deoxy-pregnanes in congenital adrenal hyperplasia due to 21-hydroxylase deficiency. *J Steroid Biochem Mol Biol*, *45*(5), 421-434. https://doi.org/10.1016/0960-0760(93)90011-k

Krone, N., & Arlt, W. (2009). Genetics of congenital adrenal hyperplasia. *Best Pract Res Clin Endocrinol Metab*, *23*(2), 181-192. https://doi.org/10.1016/j.beem.2008.10.014

Krone, N., Hughes, B. A., Lavery, G. G., Stewart, P. M., Arlt, W., & Shackleton, C. H. (2010). Gas chromatography/mass spectrometry (GC/MS) remains a pre-eminent discovery tool in clinical steroid investigations even in the era of fast liquid chromatography tandem mass spectrometry (LC/MS/MS). *J Steroid Biochem Mol Biol*, *121*(3-5), 496-504. https://doi.org/10.1016/j.jsbmb.2010.04.010

Krone, N., Reisch, N., Idkowiak, J., Dhir, V., Ivison, H. E., Hughes, B. A., Rose, I. T., O'Neil, D. M., Vijzelaar, R., Smith, M. J., MacDonald, F., Cole, T. R., Adolphs, N., Barton, J. S., Blair, E. M., Braddock, S. R., Collins, F., Cragun, D. L., Dattani, M. T., . . . Arlt, W. (2012). Genotype-phenotype analysis in congenital adrenal hyperplasia due to P450 oxidoreductase deficiency. *J Clin Endocrinol Metab*, *97*(2), E257-267. https://doi.org/10.1210/jc.2011-0640

Kroon, D.-J. (2022). *FMINLBFGS: Fast Limited Memory Optimizer* MATLAB Central File Exchange. Retrieved 10 June from https://www.mathworks.com/matlabcentral/fileexchange/23245-fminlbfgs-fast-limited-memory-optimizer

Kubat, M. (2015). *An Introduction to Machine Learning*. Springer Publishing Company, Incorporated.

Kulle, A., Krone, N., Holterhus, P. M., Schuler, G., Greaves, R. F., Juul, A., de Rijke, Y. B., Hartmann, M. F., Saba, A., Hiort, O., Wudy, S. A., & Action, E. C. (2017). Steroid hormone analysis in diagnosis and treatment of DSD: position paper of EU COST Action BM 1303 'DSDnet'. *Eur J Endocrinol*, *176*(5), P1-P9. https://doi.org/10.1530/EJE-16-0953

Labrie, F., Belanger, A., Cusan, L., Gomez, J. L., & Candas, B. (1997). Marked decline in serum concentrations of adrenal C19 sex steroid precursors and conjugated androgen metabolites during aging. *J Clin Endocrinol Metab*, *82*(8), 2396-2402. https://doi.org/10.1210/jcem.82.8.4160

Labrie, F., Luu-The, V., Labrie, C., Belanger, A., Simard, J., Lin, S. X., & Pelletier, G. (2003). Endocrine and intracrine sources of androgens in women: inhibition of breast cancer and other roles of androgens and their precursor dehydroepiandrosterone. *Endocr Rev*, *24*(2), 152-182. https://doi.org/10.1210/er.2001-0031

Labrie, F., Simard, J., Luu-The, V., Pelletier, G., Belghmi, K., & Belanger, A. (1994). Structure, regulation and role of 3 beta-hydroxysteroid dehydrogenase, 17 beta-hydroxysteroid dehydrogenase and aromatase enzymes in the formation of sex steroids in classical and peripheral intracrine tissues. *Baillieres Clin Endocrinol Metab*, *8*(2), 451-474. https://doi.org/10.1016/s0950-351x(05)80261-7

Lachance, Y., Luu-The, V., Labrie, C., Simard, J., Dumont, M., de Launoit, Y., Guerin, S., Leblanc, G., & Labrie, F. (1990). Characterization of human 3 beta-hydroxysteroid dehydrogenase/delta 5-delta 4-isomerase gene and its

expression in mammalian cells. *J Biol Chem*, *265*(33), 20469-20475. https://www.ncbi.nlm.nih.gov/pubmed/2243100

Lahti, A., Hyltoft Petersen, P., Boyd, J. C., Fraser, C. G., & Jorgensen, N. (2002). Objective criteria for partitioning Gaussian-distributed reference values into subgroups. *Clin Chem*, *48*(2), 338-352. https://www.ncbi.nlm.nih.gov/pubmed/11805016

Lahti, A., Petersen, P. H., Boyd, J. C., Rustad, P., Laake, P., & Solberg, H. E. (2004). Partitioning of nongaussian-distributed biochemical reference data into subgroups. *Clin Chem*, *50*(5), 891-900. https://doi.org/10.1373/clinchem.2003.027953

Lam, K. Y., Chan, A. C., & Lo, C. Y. (2001). Morphological analysis of adrenal glands: a prospective analysis. *Endocr Pathol*, *12*(1), 33-38. https://doi.org/10.1385/ep:12:1:33

Lapaine, M. (2011). Mollweide map projection. *KoG*, *15*(15.), 7-16.

Lee, P. A., Nordenstrom, A., Houk, C. P., Ahmed, S. F., Auchus, R., Baratz, A., Baratz Dalke, K., Liao, L. M., Lin-Su, K., Looijenga, L. H., 3rd, Mazur, T., Meyer-Bahlburg, H. F., Mouriquand, P., Quigley, C. A., Sandberg, D. E., Vilain, E., Witchel, S., & Global, D. S. D. U. C. (2016). Global Disorders of Sex Development Update since 2006: Perceptions, Approach and Care. *Horm Res Paediatr*, *85*(3), 158-180. https://doi.org/10.1159/000442975

Lee, T. C., Miller, W. L., & Auchus, R. J. (1999). Medroxyprogesterone acetate and dexamethasone are competitive inhibitors of different human steroidogenic enzymes. *J Clin Endocrinol Metab*, *84*(6), 2104-2110. https://doi.org/10.1210/jcem.84.6.5646

Lee-Robichaud, P., Wright, J. N., Akhtar, M. E., & Akhtar, M. (1995). Modulation of the activity of human 17 alpha-hydroxylase-17,20-lyase (CYP17) by cytochrome b5: endocrinological and mechanistic implications. *Biochem J*, *308 (Pt 3)*, 901-908. https://doi.org/10.1042/bj3080901

Little, R. J. A. (1988). A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association*, *83*(404), 1198-1202. https://doi.org/10.1080/01621459.1988.10478722

Little, R. J. A. R., D.B. (2019). *Statistical Analysis with Missing Data* (3rd Edition ed., Vol. Volume 793). John Wiley & Sons.

Loch Batista, R., Inacio, M., Prado Arnhold, I. J., Gomes, N. L., Diniz Faria, J. A., Rodrigues de Moraes, D., Frade Costa, E. M., Domenice, S., & Bilharinho Mendonca, B. (2019). Psychosexual Aspects, Effects of Prenatal Androgen Exposure, and Gender Change in 46,XY Disorders of Sex Development. *J Clin Endocrinol Metab*, *104*(4), 1160-1170. https://doi.org/10.1210/jc.2018-01866

Lucas-Herald, A. K., Rodie, M., Lucaccioni, L., Shapiro, D., McNeilly, J., Shaikh, M. G., & Ahmed, S. F. (2015). The pitfalls associated with urinary steroid metabolite ratios in children undergoing investigations for suspected disorders of steroid synthesis. *Int J Pediatr Endocrinol*, *2015*(1), 10. https://doi.org/10.1186/s13633-015-0007-1

Lupien, S., Lecours, A. R., Schwartz, G., Sharma, S., Hauger, R. L., Meaney, M. J., & Nair, N. P. (1996). Longitudinal study of basal cortisol levels in healthy elderly subjects: evidence for subgroups. *Neurobiol Aging*, *17*(1), 95-105. https://doi.org/10.1016/0197-4580(95)02005-5

Lutfallah, C., Wang, W., Mason, J. I., Chang, Y. T., Haider, A., Rich, B., Castro-Magana, M., Copeland, K. C., David, R., & Pang, S. (2002). Newly proposed

hormonal criteria via genotypic proof for type II 3beta-hydroxysteroid dehydrogenase deficiency. *J Clin Endocrinol Metab*, *87*(6), 2611-2622. https://doi.org/10.1210/jcem.87.6.8615

Luu-The, V., Takahashi, M., de Launoit, Y., Dumont, M., Lachance, Y., & Labrie, F. (1991). Evidence for distinct dehydrogenase and isomerase sites within a single 3 beta-hydroxysteroid dehydrogenase/5-ene-4-ene isomerase protein. *Biochemistry*, *30*(36), 8861-8865. https://doi.org/10.1021/bi00100a019

Lynch, K. L. (2018). Accreditation and Quality Assurance for Clinical Liquid Chromatography-Mass Spectrometry Laboratories. *Clin Lab Med*, *38*(3), 515-526. https://doi.org/10.1016/j.cll.2018.05.002

Maimoun, L., Philibert, P., Cammas, B., Audran, F., Bouchard, P., Fenichel, P., Cartigny, M., Pienkowski, C., Polak, M., Skordis, N., Mazen, I., Ocal, G., Berberoglu, M., Reynaud, R., Baumann, C., Cabrol, S., Simon, D., Kayemba-Kay's, K., De Kerdanet, M., . . . Sultan, C. (2011). Phenotypical, biological, and molecular heterogeneity of 5alpha-reductase deficiency: an extensive international experience of 55 patients. *J Clin Endocrinol Metab*, *96*(2), 296-307. https://doi.org/10.1210/jc.2010-1024

Malunowicz, E. M., Mitkowska, Z., Bal, K., Nizankowska-Blaz, T., Moszczynska, E., Iwanicka, Z., & Romer, T. E. (1997). Definitive diagnosis of enzymatic deficiencies of steroidogenesis in at-risk newborns and infants by urinary marker analysis using GC/MS-SIM. *Horm Res*, *48*(6), 243-251. https://doi.org/10.1159/000185529

Martel, C., Melner, M. H., Gagne, D., Simard, J., & Labrie, F. (1994). Widespread tissue distribution of steroid sulfatase, 3 beta-hydroxysteroid dehydrogenase/delta 5-delta 4 isomerase (3 beta-HSD), 17 beta-HSD 5 alpha-reductase and aromatase activities in the rhesus monkey. *Mol Cell Endocrinol*, *104*(1), 103-111. https://doi.org/10.1016/0303-7207(94)90056-6

Mason, J. I., & Rainey, W. E. (1987). Steroidogenesis in the human fetal adrenal: a role for cholesterol synthesized de novo. *J Clin Endocrinol Metab*, *64*(1), 140-147. https://doi.org/10.1210/jcem-64-1-140

Mendonca, B. B., Batista, R. L., Domenice, S., Costa, E. M., Arnhold, I. J., Russell, D. W., & Wilson, J. D. (2016). Steroid 5alpha-reductase 2 deficiency. *J Steroid Biochem Mol Biol*, *163*, 206-211. https://doi.org/10.1016/j.jsbmb.2016.05.020

Mendonca, B. B., Gomes, N. L., Costa, E. M., Inacio, M., Martin, R. M., Nishi, M. Y., Carvalho, F. M., Tibor, F. D., & Domenice, S. (2017). 46,XY disorder of sex development (DSD) due to 17beta-hydroxysteroid dehydrogenase type 3 deficiency. *J Steroid Biochem Mol Biol*, *165*(Pt A), 79-85. https://doi.org/10.1016/j.jsbmb.2016.05.002

Merke, D. P., Chrousos, G. P., Eisenhofer, G., Weise, M., Keil, M. F., Rogol, A. D., Van Wyk, J. J., & Bornstein, S. R. (2000). Adrenomedullary dysplasia and hypofunction in patients with classic 21-hydroxylase deficiency. *N Engl J Med*, *343*(19), 1362-1368. https://doi.org/10.1056/NEJM200011093431903

Mermejo, L. M., Elias, L. L., Marui, S., Moreira, A. C., Mendonca, B. B., & de Castro, M. (2005). Refining hormonal diagnosis of type II 3beta-hydroxysteroid dehydrogenase deficiency in patients with premature pubarche and hirsutism based on HSD3B2 genotyping. *J Clin Endocrinol Metab*, *90*(3), 1287-1293. https://doi.org/10.1210/jc.2004-1552

Meunier, B., de Visser, S. P., & Shaik, S. (2004). Mechanism of oxidation reactions catalyzed by cytochrome p450 enzymes. *Chem Rev*, *104*(9), 3947-3980. https://doi.org/10.1021/cr020443g

Midgley, P. C., Russell, K., Oates, N., Holownia, P., Shaw, J. C., & Honour, J. W. (1998). Adrenal function in preterm infants: ACTH may not be the sole regulator of the fetal zone. *Pediatr Res*, *44*(6), 887-893. https://doi.org/10.1203/00006450-199812000-00011

Midgley, P. C., Russell, K., Oates, N., Shaw, J. C., & Honour, J. W. (1996). Activity of the adrenal fetal zone in preterm infants continues to term. *Endocr Res*, *22*(4), 729-733. https://doi.org/10.1080/07435809609043769

Miller, W. L. (1998). Steroid hormone biosynthesis and actions in the materno-feto-placental unit. *Clin Perinatol*, *25*(4), 799-817, v. https://www.ncbi.nlm.nih.gov/pubmed/9891616

Miller, W. L. (2002). Androgen biosynthesis from cholesterol to DHEA. *Mol Cell Endocrinol*, *198*(1-2), 7-14. https://doi.org/10.1016/s0303-7207(02)00363-5

Miller, W. L. (2017). Disorders in the initial steps of steroid hormone synthesis. *J Steroid Biochem Mol Biol*, *165*(Pt A), 18-37. https://doi.org/10.1016/j.jsbmb.2016.03.009

Miller, W. L. (2019). Congenital adrenal hyperplasia: time to replace 17OHP with 21-deoxycortisol. *Hormone Research in Paediatrics*, *91*(6), 416-420.

Miller, W. L., & Auchus, R. J. (2011). The molecular biology, biochemistry, and physiology of human steroidogenesis and its disorders. *Endocr Rev*, *32*(1), 81-151. https://doi.org/10.1210/er.2010-0013

Miller, W. L., & Auchus, R. J. (2019). The "backdoor pathway" of androgen synthesis in human male sexual development. *PLoS Biol*, *17*(4), e3000198. https://doi.org/10.1371/journal.pbio.3000198

Miller, W. L., Huang, N., Fluck, C. E., & Pandey, A. V. (2004). P450 oxidoreductase deficiency. *Lancet*, *364*(9446), 1663. https://doi.org/10.1016/S0140-6736(04)17344-3

Moffat, S. D., An, Y., Resnick, S. M., Diamond, M. P., & Ferrucci, L. (2020). Longitudinal Change in Cortisol Levels Across the Adult Life Span. *J Gerontol A Biol Sci Med Sci*, *75*(2), 394-400. https://doi.org/10.1093/gerona/gly279

Morel, Y., Roucher, F., Plotton, I., Goursaud, C., Tardy, V., & Mallet, D. (2016). Evolution of steroids during pregnancy: Maternal, placental and fetal synthesis. *Ann Endocrinol (Paris)*, *77*(2), 82-89. https://doi.org/10.1016/j.ando.2016.04.023

Mueller, J. W., Gilligan, L. C., Idkowiak, J., Arlt, W., & Foster, P. A. (2015). The Regulation of Steroid Action by Sulfation and Desulfation. *Endocr Rev*, *36*(5), 526-563. https://doi.org/10.1210/er.2015-1036

Mulatero, P., Curnow, K. M., Aupetit-Faisant, B., Foekling, M., Gomez-Sanchez, C., Veglio, F., Jeunemaitre, X., Corvol, P., & Pascoe, L. (1998). Recombinant CYP11B genes encode enzymes that can catalyze conversion of 11-deoxycortisol to cortisol, 18-hydroxycortisol, and 18-oxocortisol. *J Clin Endocrinol Metab*, *83*(11), 3996-4001. https://doi.org/10.1210/jcem.83.11.5237

Naccache, A., Louiset, E., Duparc, C., Laquerriere, A., Patrier, S., Renouf, S., Gomez-Sanchez, C. E., Mukai, K., Lefebvre, H., & Castanet, M. (2016). Temporal and spatial distribution of mast cells and steroidogenic enzymes in the human fetal adrenal. *Mol Cell Endocrinol*, *434*, 69-80. https://doi.org/10.1016/j.mce.2016.06.015

Nakamura, Y., Gang, H. X., Suzuki, T., Sasano, H., & Rainey, W. E. (2009). Adrenal changes associated with adrenarche. *Rev Endocr Metab Disord*, *10*(1), 19-26. https://doi.org/10.1007/s11154-008-9092-2

Nakamura, Y., Hornsby, P. J., Casson, P., Morimoto, R., Satoh, F., Xing, Y., Kennedy, M. R., Sasano, H., & Rainey, W. E. (2009). Type 5 17beta-hydroxysteroid dehydrogenase (AKR1C3) contributes to testosterone production in the adrenal reticularis. *J Clin Endocrinol Metab*, *94*(6), 2192-2198. https://doi.org/10.1210/jc.2008-2374

Nakamura, Y., Xing, Y., Hui, X. G., Kurotaki, Y., Ono, K., Cohen, T., Sasano, H., & Rainey, W. E. (2011). Human adrenal cells that express both 3beta-hydroxysteroid dehydrogenase type 2 (HSD3B2) and cytochrome b5 (CYB5A) contribute to adrenal androstenedione production. *J Steroid Biochem Mol Biol*, *123*(3-5), 122-126. https://doi.org/10.1016/j.jsbmb.2010.12.001

Nanba, A. T., Rege, J., Ren, J., Auchus, R. J., Rainey, W. E., & Turcu, A. F. (2019). 11-Oxygenated C19 Steroids Do Not Decline With Age in Women. *J Clin Endocrinol Metab*, *104*(7), 2615-2622. https://doi.org/10.1210/jc.2018-02527

Nanba, K., Vaidya, A., Williams, G. H., Zheng, I., Else, T., & Rainey, W. E. (2017). Age-Related Autonomous Aldosteronism. *Circulation*, *136*(4), 347-355. https://doi.org/10.1161/CIRCULATIONAHA.117.028201

Nebert, D. W., & Russell, D. W. (2002). Clinical importance of the cytochromes P450. *Lancet*, *360*(9340), 1155-1162. https://doi.org/10.1016/S0140-6736(02)11203-7

Neres, M. S., Auchus, R. J., Shackleton, C. H., & Kater, C. E. (2010). Distinctive profile of the 17-hydroxylase and 17,20-lyase activities revealed by urinary steroid metabolomes of patients with CYP17 deficiency. *Arq Bras Endocrinol Metabol*, *54*(9), 826-832. https://doi.org/10.1590/s0004-27302010000900009

Nonaka, K., Aida, J., Takubo, K., Yamazaki, Y., Gao, X., Komatsu, A., Takakuma, S., Kakizaki, M., Inoshita, N., Gomi, F., Ishiwata, T., Chong, J. M., Arai, T., & Sasano, H. (2020). Correlation Between Telomere Attrition of Zona Fasciculata and Adrenal Weight Reduction in Older Men. *J Clin Endocrinol Metab*, *105*(3). https://doi.org/10.1210/clinem/dgz214

Nova, D., & Estévez, P. A. (2014). A review of learning vector quantization classifiers. *Neural Computing and Applications*, *25*(3), 511-524.

Nowakowski, H., & Lenz, W. (1961). Genetic aspects in male hypogonadism. *Recent Prog Horm Res*, *17*, 53-95. https://www.ncbi.nlm.nih.gov/pubmed/13729828

O'Reilly, M. W., Kempegowda, P., Jenkinson, C., Taylor, A. E., Quanson, J. L., Storbeck, K. H., & Arlt, W. (2017). 11-Oxygenated C19 Steroids Are the Predominant Androgens in Polycystic Ovary Syndrome. *J Clin Endocrinol Metab*, *102*(3), 840-848. https://doi.org/10.1210/jc.2016-3285

O'Shaughnessy, P. J., Monteiro, A., Bhattacharya, S., Fraser, M. J., & Fowler, P. A. (2013). Steroidogenic enzyme expression in the human fetal liver and potential role in the endocrinology of pregnancy. *Mol Hum Reprod*, *19*(3), 177-187. https://doi.org/10.1093/molehr/gas059

Odame, I., Donaldson, M. D., Wallace, A. M., Cochran, W., & Smith, P. J. (1992). Early diagnosis and management of 5 alpha-reductase deficiency. *Arch Dis Child*, *67*(6), 720-723. https://doi.org/10.1136/adc.67.6.720

Okeigwe, I., & Kuohung, W. (2014). 5-Alpha reductase deficiency: a 40-year retrospective review. *Curr Opin Endocrinol Diabetes Obes*, *21*(6), 483-487. https://doi.org/10.1097/MED.0000000000000116

OpenStaxCollege. The Adrenal Glands. In F. T. A. Glands.jpg" (Ed.), *Creative Commons* (pp. "File:1818 The Adrenal Glands.jpg" by OpenStax College is marked with CC BY 1813.1810. To view the terms, visit https://creativecommons.org/licenses/by/1813.1810?ref=openverse).

Ozarda, Y., Sikaris, K., Streichert, T., Macri, J., intervals, I. C. o. R., & Decision, L. (2018). Distinguishing reference intervals and clinical decision limits - A review by the IFCC Committee on Reference Intervals and Decision Limits. *Crit Rev Clin Lab Sci*, *55*(6), 420-431. https://doi.org/10.1080/10408363.2018.1482256

Pandey, A. V., & Sproll, P. (2014). Pharmacogenomics of human P450 oxidoreductase. *Front Pharmacol*, *5*, 103. https://doi.org/10.3389/fphar.2014.00103

Penning, T. M. (1997). Molecular endocrinology of hydroxysteroid dehydrogenases. *Endocr Rev*, *18*(3), 281-305. https://doi.org/10.1210/edrv.18.3.0302

Perry, R. J., Novikova, E., Wallace, A. M., & Donaldson, M. D. (2011). Pitfalls in the diagnosis of 5alpha-reductase type 2 deficiency during early infancy. *Horm Res Paediatr*, *75*(5), 380-382. https://doi.org/10.1159/000324646

Peter, M., Janzen, N., Sander, S., Korsch, E., Riepe, F. G., & Sander, J. (2008). A case of 11beta-hydroxylase deficiency detected in a newborn screening program by second-tier LC-MS/MS. *Horm Res*, *69*(4), 253-256. https://doi.org/10.1159/000113027

Peters, C. J., Nugent, T., Perry, L. A., Davies, K., Morel, Y., Drake, W. M., Savage, M. O., & Johnston, L. B. (2007). Cosegregation of a novel homozygous CYP11B1 mutation with the phenotype of non-classical congenital adrenal hyperplasia in a consanguineous family. *Horm Res*, *67*(4), 189-193. https://doi.org/10.1159/000097244

Peterson, R. E., Imperato-McGinley, J., Gautier, T., & Shackleton, C. (1985a). Male pseudohermaphroditism due to multiple defects in steroid-biosynthetic microsomal mixed-function oxidases. A new variant of congenital adrenal hyperplasia. *N Engl J Med*, *313*(19), 1182-1191. https://doi.org/10.1056/NEJM198511073131903

Peterson, R. E., Imperato-McGinley, J., Gautier, T., & Shackleton, C. (1985b). Urinary steroid metabolites in subjects with male pseudohermaphroditism due to 5 alpha-reductase deficiency. *Clin Endocrinol (Oxf)*, *23*(1), 43-53. https://doi.org/10.1111/j.1365-2265.1985.tb00181.x

Peterson, R. E., Imperato-McGinley, J., Gautier, T., & Sturla, E. (1977). Male pseudohermaphroditism due to steroid 5-alpha-reductase deficiency. *Am J Med*, *62*(2), 170-191. https://doi.org/10.1016/0002-9343(77)90313-8

Pezzi, V., Mathis, J. M., Rainey, W. E., & Carr, B. R. (2003). Profiling transcript levels for steroidogenic enzymes in fetal tissues. *J Steroid Biochem Mol Biol*, *87*(2-3), 181-189. https://doi.org/10.1016/j.jsbmb.2003.07.006

Pretorius, E., Africander, D. J., Vlok, M., Perkins, M. S., Quanson, J., & Storbeck, K. H. (2016). 11-Ketotestosterone and 11-Ketodihydrotestosterone in Castration Resistant Prostate Cancer: Potent Androgens Which Can No Longer Be Ignored. *PLoS One*, *11*(7), e0159867. https://doi.org/10.1371/journal.pone.0159867

Pretorius, E., Arlt, W., & Storbeck, K. H. (2017). A new dawn for androgens: Novel lessons from 11-oxygenated C19 steroids. *Mol Cell Endocrinol*, *441*, 76-85. https://doi.org/10.1016/j.mce.2016.08.014

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine Learning in Medicine. *N Engl J Med*, *380*(14), 1347-1358. https://doi.org/10.1056/NEJMra1814259

Reeder, A. Y., & Joannou, G. E. (1996). 15 beta-hydroxysteroids (part I). Steroids of the human perinatal period: the synthesis of 3 beta,15 beta, 17 alpha-

trihydroxy-5-pregnen-20-one. *Steroids*, *61*(2), 74-81. https://doi.org/10.1016/0039-128x(95)00193-t

Rege, J., Nakamura, Y., Satoh, F., Morimoto, R., Kennedy, M. R., Layman, L. C., Honma, S., Sasano, H., & Rainey, W. E. (2013). Liquid chromatography-tandem mass spectrometry analysis of human adrenal vein 19-carbon steroids before and after ACTH stimulation. *J Clin Endocrinol Metab*, *98*(3), 1182-1188. https://doi.org/10.1210/jc.2012-2912

Reisch, N., Hogler, W., Parajes, S., Rose, I. T., Dhir, V., Gotzinger, J., Arlt, W., & Krone, N. (2013). A diagnosis not to be missed: nonclassic steroid 11beta-hydroxylase deficiency presenting with premature adrenarche and hirsutism. *J Clin Endocrinol Metab*, *98*(10), E1620-1625. https://doi.org/10.1210/jc.2013-1306

Reisch, N., Idkowiak, J., Hughes, B. A., Ivison, H. E., Abdul-Rahman, O. A., Hendon, L. G., Olney, A. H., Nielsen, S., Harrison, R., Blair, E. M., Dhir, V., Krone, N., Shackleton, C. H., & Arlt, W. (2013). Prenatal diagnosis of congenital adrenal hyperplasia caused by P450 oxidoreductase deficiency. *J Clin Endocrinol Metab*, *98*(3), E528-536. https://doi.org/10.1210/jc.2012-3449

Reisch, N., Taylor, A. E., Nogueira, E. F., Asby, D. J., Dhir, V., Berry, A., Krone, N., Auchus, R. J., Shackleton, C. H. L., Hanley, N. A., & Arlt, W. (2019). Alternative pathway androgen biosynthesis and human fetal female virilization. *Proc Natl Acad Sci U S A*, *116*(44), 22294-22299. https://doi.org/10.1073/pnas.1906623116

Reiter, E. O., Fuldauer, V. G., & Root, A. W. (1977). Secretion of the adrenal androgen, dehydroepiandrosterone sulfate, during normal infancy, childhood, and adolescence, in sick infants, and in children with endocrinologic abnormalities. *J Pediatr*, *90*(5), 766-770. https://doi.org/10.1016/s0022-3476(77)81244-4

Remer, T., Boye, K. R., Hartmann, M. F., & Wudy, S. A. (2005). Urinary markers of adrenarche: reference values in healthy subjects, aged 3-18 years. *J Clin Endocrinol Metab*, *90*(4), 2015-2021. https://doi.org/10.1210/jc.2004-1571

Rheaume, E., Simard, J., Morel, Y., Mebarki, F., Zachmann, M., Forest, M. G., New, M. I., & Labrie, F. (1992). Congenital adrenal hyperplasia due to point mutations in the type II 3 beta-hydroxysteroid dehydrogenase gene. *Nat Genet*, *1*(4), 239-245. https://doi.org/10.1038/ng0792-239

Ribes, V., Otto, D. M., Dickmann, L., Schmidt, K., Schuhbaur, B., Henderson, C., Blomhoff, R., Wolf, C. R., Tickle, C., & Dolle, P. (2007). Rescue of cytochrome P450 oxidoreductase (Por) mouse mutants reveals functions in vasculogenesis, brain and limb patterning linked to retinoic acid homeostasis. *Dev Biol*, *303*(1), 66-81. https://doi.org/10.1016/j.ydbio.2006.10.032

Rich, B. H., Rosenfield, R. L., Lucky, A. W., Helke, J. C., & Otto, P. (1981). Adrenarche: changing adrenal response to adrenocorticotropin. *J Clin Endocrinol Metab*, *52*(6), 1129-1136. https://doi.org/10.1210/jcem-52-6-1129

Rogers, S. L., Hughes, B. A., Jones, C. A., Freedman, L., Smart, K., Taylor, N., Stewart, P. M., Shackleton, C. H., Krone, N. P., Blissett, J., & Tomlinson, J. W. (2014). Diminished 11beta-hydroxysteroid dehydrogenase type 2 activity is associated with decreased weight and weight gain across the first year of life. *J Clin Endocrinol Metab*, *99*(5), E821-831. https://doi.org/10.1210/jc.2013-3254

Rosler, A., Leiberman, E., & Cohen, T. (1992). High frequency of congenital adrenal hyperplasia (classic 11 beta-hydroxylase deficiency) among Jews from

Morocco. *Am J Med Genet*, *42*(6), 827-834. https://doi.org/10.1002/ajmg.1320420617

Rosler, A., Leiberman, E., Sack, J., Landau, H., Benderly, A., Moses, S. W., & Cohen, T. (1982). Clinical variability of congenital adrenal hyperplasia due to 11 beta-hydroxylase deficiency. *Horm Res*, *16*(3), 133-141. https://doi.org/10.1159/000179494

Rousson, V., Ackermann, D., Ponte, B., Pruijm, M., Guessous, I., d'Uscio, C. H., Ehret, G., Escher, G., Pechere-Bertschi, A., Groessl, M., Martin, P. Y., Burnier, M., Dick, B., Bochud, M., Vogt, B., & Dhayat, N. A. (2021). Sex- and age-specific reference intervals for diagnostic ratios reflecting relative activity of steroidogenic enzymes and pathways in adults. *PLoS One*, *16*(7), e0253975. https://doi.org/10.1371/journal.pone.0253975

Saez, J. M., De Peretti, E., Morera, A. M., David, M., & Bertrand, J. (1971). Familial male pseudohermaphroditism with gynecomastia due to a testicular 17-ketosteroid reductase defect. I. Studies in vivo. *J Clin Endocrinol Metab*, *32*(5), 604-610. https://doi.org/10.1210/jcem-32-5-604

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, *3*(3), 210-229. https://doi.org/10.1147/rd.33.0210

Sato, A., & Yamada, K. (1995). Generalized learning vector quantization. *Advances in neural information processing systems*, *8*.

Savchuk, I., Morvan, M. L., Antignac, J. P., Gemzell-Danielsson, K., Le Bizec, B., Soder, O., & Svechnikov, K. (2017). Androgenic potential of human fetal adrenals at the end of the first trimester. *Endocr Connect*, *6*(6), 348-359. https://doi.org/10.1530/EC-17-0085

Schiffer, L., Arlt, W., & Storbeck, K. H. (2018). Intracrine androgen biosynthesis, metabolism and action revisited. *Mol Cell Endocrinol*, *465*, 4-26. https://doi.org/10.1016/j.mce.2017.08.016

Schiffer, L., Barnard, L., Baranowski, E. S., Gilligan, L. C., Taylor, A. E., Arlt, W., Shackleton, C. H. L., & Storbeck, K. H. (2019). Human steroid biosynthesis, metabolism and excretion are differentially reflected by serum and urine steroid metabolomes: A comprehensive review. *J Steroid Biochem Mol Biol*, *194*, 105439. https://doi.org/10.1016/j.jsbmb.2019.105439

Schneider, P. (2010). *Advanced methods for prototype-based classification*. Atto Producties Europe.

Schneider, P., Biehl, M., & Hammer, B. (2009). Adaptive relevance matrices in learning vector quantization. *Neural Comput*, *21*(12), 3532-3561. https://doi.org/10.1162/neco.2009.11-08-908

Schwartz, W. B. (1970). Medicine and the computer. The promise and problems of change. *N Engl J Med*, *283*(23), 1257-1264. https://doi.org/10.1056/NEJM197012032832305

Seeman, T. E., McEwen, B. S., Singer, B. H., Albert, M. S., & Rowe, J. W. (1997). Increase in urinary cortisol excretion and memory declines: MacArthur studies of successful aging. *J Clin Endocrinol Metab*, *82*(8), 2458-2465. https://doi.org/10.1210/jcem.82.8.4173

Seneviratne, M. G., Shah, N. H., & Chu, L. (2020). Bridging the implementation gap of machine learning in healthcare. *BMJ Innovations*, *6*(2), 45-47. https://doi.org/10.1136/bmjinnov-2019-000359

Shackleton, C., Marcos, J., Malunowicz, E. M., Szarras-Czapnik, M., Jira, P., Taylor, N. F., Murphy, N., Crushell, E., Gottschalk, M., Hauffa, B., Cragun, D. L.,

Hopkin, R. J., Adachi, M., & Arlt, W. (2004). Biochemical diagnosis of Antley-Bixler syndrome by steroid analysis. *Am J Med Genet A*, *128A*(3), 223-231. https://doi.org/10.1002/ajmg.a.30104

Shackleton, C., Pozo, O. J., & Marcos, J. (2018). GC/MS in Recent Years Has Defined the Normal and Clinically Disordered Steroidome: Will It Soon Be Surpassed by LC/Tandem MS in This Role? *Journal of the Endocrine Society*, *2*(8), 974-996. https://doi.org/10.1210/js.2018-00135

Shackleton, C. H. (1986). Profiling steroid hormones and urinary steroids. *J Chromatogr*, *379*, 91-156. https://doi.org/10.1016/s0378-4347(00)80683-0

Shackleton, C. H. (2008). Genetic Disorders of Steroid Metabolism Diagnosed by Mass Spectrometry. *Laboratory Guide to the Methods in Biochemical Genetics*, 549-605. https://doi.org/10.1007/978-3-540-76698-8_26

Shackleton, C. H., Honour, J. W., & Taylor, N. F. (1979). Metabolism of fetal and neonatal adrenal steroids. *J Steroid Biochem*, *11*(1B), 523-529. https://doi.org/10.1016/0022-4731(79)90077-3

Shackleton, C. H., Irias, J., McDonald, C., & Imperato-McGinley, J. (1986). Late-onset 21-hydroxylase deficiency: reliable diagnosis by steroid analysis of random urine collections. *Steroids*, *48*(3-4), 239-250. https://doi.org/10.1016/0039-128x(86)90007-3

Shackleton, C. H., Taylor, N., & Honour, J. (1980). *An atlas of gas chromatographic profiles of neutral urinary steroids in health and disease*. Packard-Becker.

Shackleton, C. H., & Taylor, N. F. (1975). Identification of the androstenetriolones and androstenetetrols present in the urine of infants. *J Steroid Biochem*, *6*(10), 1393-1399. https://doi.org/10.1016/0022-4731(75)90075-8

Shackleton, C. H., & Whitney, J. O. (1980). Use of Sep-pak cartridges for urinary steroid extraction: evaluation of the method for use prior to gas chromatographic analysis. *Clin Chim Acta*, *107*(3), 231-243. https://doi.org/10.1016/0009-8981(80)90451-9

Shackleton, C. H. L., & Marcos, J. (2005). GC/MS steroid profiling: diagnosis of disorders affecting steroid synthesis and metabolism. *The Encyclopedia of Mass Spectrometry*, *8*, 789-813.

Shaw, G., Fenelon, J., Sichlau, M., Auchus, R. J., Wilson, J. D., & Renfree, M. B. (2006). Role of the alternate pathway of dihydrotestosterone formation in virilization of the Wolffian ducts of the tammar wallaby, Macropus eugenii. *Endocrinology*, *147*(5), 2368-2373. https://doi.org/10.1210/en.2005-1251

Shepard, T. H., & Clausen, S. W. (1951). Case of adrenogenital syndrome with hypertension treated with cortisone. *Pediatrics*, *8*(6), 805-811. https://www.ncbi.nlm.nih.gov/pubmed/14911251

Sherbet, D. P., Tiosano, D., Kwist, K. M., Hochberg, Z., & Auchus, R. J. (2003). CYP17 mutation E305G causes isolated 17,20-lyase deficiency by selectively altering substrate binding. *J Biol Chem*, *278*(49), 48563-48569. https://doi.org/10.1074/jbc.M307586200

Shikita, M., & Hall, P. F. (1973). Cytochrome P-450 from bovine adrenocortical mitochondria: an enzyme for the side chain cleavage of cholesterol. II. Subunit structure. *J Biol Chem*, *248*(16), 5605-5609. https://www.ncbi.nlm.nih.gov/pubmed/4723902

Shimozawa, K., Saisho, S., Yata, J., & Kambegawa, A. (1988). Age-related changes in serum 17-hydroxypregnenolone and 17-hydroxypregnenolone sulfate concentrations in human infancy and childhood. *Endocrinol Jpn*, *35*(2), 189-195. https://doi.org/10.1507/endocrj1954.35.189

Sidey-Gibbons, J. A. M., & Sidey-Gibbons, C. J. (2019). Machine learning in medicine: a practical introduction. *BMC Med Res Methodol*, *19*(1), 64. https://doi.org/10.1186/s12874-019-0681-4

Siest, G., Henny, J., Grasbeck, R., Wilding, P., Petitclerc, C., Queralto, J. M., & Hyltoft Petersen, P. (2013). The theory of reference values: an unfinished symphony. *Clin Chem Lab Med*, *51*(1), 47-64. https://doi.org/10.1515/cclm-2012-0682

Simpson, E. R., & MacDonald, P. C. (1981). Endocrine physiology of the placenta. *Annu Rev Physiol*, *43*, 163-188. https://doi.org/10.1146/annurev.ph.43.030181.001115

Speiser, P. W., Azziz, R., Baskin, L. S., Ghizzoni, L., Hensle, T. W., Merke, D. P., Meyer-Bahlburg, H. F., Miller, W. L., Montori, V. M., Oberfield, S. E., Ritzen, M., White, P. C., & Endocrine, S. (2010). Congenital adrenal hyperplasia due to steroid 21-hydroxylase deficiency: an Endocrine Society clinical practice guideline. *J Clin Endocrinol Metab*, *95*(9), 4133-4160. https://doi.org/10.1210/jc.2009-2631

Storbeck, K. H., Bloem, L. M., Africander, D., Schloms, L., Swart, P., & Swart, A. C. (2013). 11beta-Hydroxydihydrotestosterone and 11-ketodihydrotestosterone, novel C19 steroids with androgenic activity: a putative role in castration resistant prostate cancer? *Mol Cell Endocrinol*, *377*(1-2), 135-146. https://doi.org/10.1016/j.mce.2013.07.006

Storbeck, K. H., Gilligan, L., Jenkinson, C., Baranowski, E. S., Quanson, J. L., Arlt, W., & Taylor, A. E. (2018). The utility of ultra-high performance supercritical fluid chromatography-tandem mass spectrometry (UHPSFC-MS/MS) for clinically relevant steroid analysis. *J Chromatogr B Analyt Technol Biomed Life Sci*, *1085*, 36-41. https://doi.org/10.1016/j.jchromb.2018.03.033

Storbeck, K. H., Schiffer, L., Baranowski, E. S., Chortis, V., Prete, A., Barnard, L., Gilligan, L. C., Taylor, A. E., Idkowiak, J., Arlt, W., & Shackleton, C. H. L. (2019). Steroid Metabolome Analysis in Disorders of Adrenal Steroid Biosynthesis and Metabolism. *Endocr Rev*, *40*(6), 1605-1625. https://doi.org/10.1210/er.2018-00262

Sulcova, J., Hill, M., Hampl, R., & Starka, L. (1997). Age and sex related differences in serum levels of unconjugated dehydroepiandrosterone and its sulphate in normal subjects. *J Endocrinol*, *154*(1), 57-62. https://doi.org/10.1677/joe.0.1540057

Sun, M., Mueller, J. W., Gilligan, L. C., Taylor, A. E., Shaheen, F., Noczynska, A., T'Sjoen, G., Denvir, L., Shenoy, S., Fulton, P., Cheetham, T. D., Gleeson, H., Rahman, M., Krone, N. P., Taylor, N. F., Shackleton, C. H. L., Arlt, W., & Idkowiak, J. (2021). The broad phenotypic spectrum of 17alpha-hydroxylase/17,20-lyase (CYP17A1) deficiency: a case series. *Eur J Endocrinol*, *185*(5), 729-741. https://doi.org/10.1530/EJE-21-0152

Suzuki, T., Sasano, H., Takeyama, J., Kaneko, C., Freije, W. A., Carr, B. R., & Rainey, W. E. (2000). Developmental changes in steroidogenic enzymes in human postnatal adrenal cortex: immunohistochemical studies. *Clin Endocrinol (Oxf)*, *53*(6), 739-747. https://doi.org/10.1046/j.1365-2265.2000.01144.x

Takeyama, J., Suzuki, T., Hirasawa, G., Muramatsu, Y., Nagura, H., Iinuma, K., Nakamura, J., Kimura, K. I., Yoshihama, M., Harada, N., Andersson, S., & Sasano, H. (2000). 17beta-hydroxysteroid dehydrogenase type 1 and 2

expression in the human fetus. *J Clin Endocrinol Metab*, *85*(1), 410-416. https://doi.org/10.1210/jcem.85.1.6323

Taylor, N. F., Curnow, D. H., & Shackleton, C. H. (1978). Analysis of glucocorticoid metabolites in the neonatal period: catabolism of cortisone acetate by an infant with 21-hydroxylase deficiency. *Clin Chim Acta*, *85*(3), 219-229. https://doi.org/10.1016/0009-8981(78)90299-1

Tezuka, Y., Atsumi, N., Blinder, A. R., Rege, J., Giordano, T. J., Rainey, W. E., & Turcu, A. F. (2021). The Age-Dependent Changes of the Human Adrenal Cortical Zones Are Not Congruent. *J Clin Endocrinol Metab*, *106*(5), 1389-1397. https://doi.org/10.1210/clinem/dgab007

Therrell, B. L., Jr., Berenbaum, S. A., Manter-Kapanke, V., Simmank, J., Korman, K., Prentice, L., Gonzalez, J., & Gunn, S. (1998). Results of screening 1.9 million Texas newborns for 21-hydroxylase-deficient congenital adrenal hyperplasia. *Pediatrics*, *101*(4 Pt 1), 583-590. https://doi.org/10.1542/peds.101.4.583

Thomas, J. L., Frieden, C., Nash, W. E., & Strickler, R. C. (1995). An NADH-induced conformational change that mediates the sequential 3 beta-hydroxysteroid dehydrogenase/isomerase activities is supported by affinity labeling and the time-dependent activation of isomerase. *J Biol Chem*, *270*(36), 21003-21008. https://doi.org/10.1074/jbc.270.36.21003

Thomas, J. L., Myers, R. P., & Strickler, R. C. (1989). Human placental 3 beta-hydroxy-5-ene-steroid dehydrogenase and steroid 5----4-ene-isomerase: purification from mitochondria and kinetic profiles, biophysical characterization of the purified mitochondrial and microsomal enzymes. *J Steroid Biochem*, *33*(2), 209-217. https://doi.org/10.1016/0022-4731(89)90296-3

Tillin, T., Hughes, A. D., Whincup, P., Mayet, J., Sattar, N., McKeigue, P. M., Chaturvedi, N., & Group, S. S. (2014). Ethnicity and prediction of cardiovascular disease: performance of QRISK2 and Framingham scores in a U.K. tri-ethnic prospective cohort study (SABRE--Southall And Brent REvisited). *Heart*, *100*(1), 60-67. https://doi.org/10.1136/heartjnl-2013-304474

Timmermans, S., Souffriau, J., & Libert, C. (2019). A General Introduction to Glucocorticoid Biology. *Front Immunol*, *10*, 1545. https://doi.org/10.3389/fimmu.2019.01545

Tomalik-Scharte, D., Maiter, D., Kirchheiner, J., Ivison, H. E., Fuhr, U., & Arlt, W. (2010). Impaired hepatic drug and steroid metabolism in congenital adrenal hyperplasia due to P450 oxidoreductase deficiency. *Eur J Endocrinol*, *163*(6), 919-924. https://doi.org/10.1530/EJE-10-0764

Travers, S., Martinerie, L., Boileau, P., Xue, Q. Y., Lombes, M., & Pussard, E. (2018). Comparative profiling of adrenal steroids in maternal and umbilical cord blood. *J Steroid Biochem Mol Biol*, *178*, 127-134. https://doi.org/10.1016/j.jsbmb.2017.11.012

Tuckey, R. C. (2005). Progesterone synthesis by the human placenta. *Placenta*, *26*(4), 273-281. https://doi.org/10.1016/j.placenta.2004.06.012

Turcu, A. F., Nanba, A. T., Chomic, R., Upadhyay, S. K., Giordano, T. J., Shields, J. J., Merke, D. P., Rainey, W. E., & Auchus, R. J. (2016). Adrenal-derived 11-oxygenated 19-carbon steroids are the dominant androgens in classic 21-hydroxylase deficiency. *Eur J Endocrinol*, *174*(5), 601-609. https://doi.org/10.1530/EJE-15-1181

TURING, A. M. (1950). I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, *LIX*(236), 433-460. https://doi.org/10.1093/mind/LIX.236.433

Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform Decis Mak*, *19*(1), 281. https://doi.org/10.1186/s12911-019-1004-8

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, *9*(11).

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLoS Med*, *15*(11), e1002689. https://doi.org/10.1371/journal.pmed.1002689

Vogeser, M., & Stone, J. A. (2020). A suggested standard for validation of LC-MS/MS based analytical series in diagnostic laboratories. *Clin Mass Spectrom*, *16*, 25-32. https://doi.org/10.1016/j.clinms.2020.02.002

Watson, D. S., Krutzinna, J., Bruce, I. N., Griffiths, C. E., McInnes, I. B., Barnes, M. R., & Floridi, L. (2019). Clinical applications of machine learning algorithms: beyond the black box. *BMJ*, *364*, l886. https://doi.org/10.1136/bmj.l886

White, P. C., Curnow, K. M., & Pascoe, L. (1994). Disorders of steroid 11 beta-hydroxylase isozymes. *Endocr Rev*, *15*(4), 421-438. https://doi.org/10.1210/edrv-15-4-421

White, P. C., New, M. I., & Dupont, B. (1986). Structure of human steroid 21-hydroxylase genes. *Proc Natl Acad Sci U S A*, *83*(14), 5111-5115. https://doi.org/10.1073/pnas.83.14.5111

WIckham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag. http://ggplot2.org

Wilkes, E. H., Rumsby, G., & Woodward, G. M. (2018). Using Machine Learning to Aid the Interpretation of Urine Steroid Profiles. *Clin Chem*, *64*(11), 1586-1595. https://doi.org/10.1373/clinchem.2018.292201

Wilkins, L., Crigler, J. F., Jr., Silverman, S. H., Gardner, L. I., & Migeon, C. J. (1952). Further studies on the treatment of congenital adrenal hyperplasia with cortisone. III. The control of hypertension with cortisone, with a discussion of variations in the type of congenital adrenal hyperplasia and report of a case with probable defect of carbohydrate-regulating hormones. *J Clin Endocrinol Metab*, *12*(8), 1015-1030. https://doi.org/10.1210/jcem-12-8-1015

Wilson, J. D., Auchus, R. J., Leihy, M. W., Guryev, O. L., Estabrook, R. W., Osborn, S. M., Shaw, G., & Renfree, M. B. (2003). 5alpha-androstane-3alpha,17beta-diol is formed in tammar wallaby pouch young testes by a pathway involving 5alpha-pregnane-3alpha,17alpha-diol-20-one as a key intermediate. *Endocrinology*, *144*(2), 575-580. https://doi.org/10.1210/en.2002-220721

Wisniewski, A. B., Batista, R. L., Costa, E. M. F., Finlayson, C., Sircili, M. H. P., Denes, F. T., Domenice, S., & Mendonca, B. B. (2019). Management of 46,XY Differences/Disorders of Sex Development (DSD) Throughout Life. *Endocr Rev*, *40*(6), 1547-1572. https://doi.org/10.1210/er.2019-00049

Wright, E. M., & Royston, P. (1997). A comparison of statistical methods for age-related reference intervals. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *160*(1), 47-69.

Wudy, S. A., Hartmann, M., & Homoki, J. (2000). Hormonal diagnosis of 21-hydroxylase deficiency in plasma and urine of neonates using benchtop gas chromatography-mass spectrometry. *J Endocrinol*, *165*(3), 679-683. https://doi.org/10.1677/joe.0.1650679

Wudy, S. A., Schuler, G., Sanchez-Guijo, A., & Hartmann, M. F. (2018). The art of measuring steroids: Principles and practice of current hormonal steroid

analysis. *J Steroid Biochem Mol Biol*, *179*, 88-103. https://doi.org/10.1016/j.jsbmb.2017.09.003

Yamamoto, Y., Hosogaya, S., Osawa, S., Ichihara, K., Onuma, T., Saito, A., Banba, K., Araki, H., Nagamine, Y., Shinohara, K., Okada, G., Matsumoto, H., Oguri, T., Gonaikawa, S., Iwagami, M., & Japanese Association of Medical, T. (2013). Nationwide multicenter study aimed at the establishment of common reference intervals for standardized clinical laboratory tests in Japan. *Clin Chem Lab Med*, *51*(8), 1663-1672. https://doi.org/10.1515/cclm-2012-0413

Yanase, T., Simpson, E. R., & Waterman, M. R. (1991). 17 alpha-hydroxylase/17,20-lyase deficiency: from clinical investigation to molecular definition. *Endocr Rev*, *12*(1), 91-108. https://doi.org/10.1210/edrv-12-1-91

Yiallouris, A., Tsioutis, C., Agapidaki, E., Zafeiri, M., Agouridis, A. P., Ntourakis, D., & Johnson, E. O. (2019). Adrenal Aging and Its Implications on Stress Responsiveness in Humans. *Front Endocrinol (Lausanne)*, *10*, 54. https://doi.org/10.3389/fendo.2019.00054

Ying, X. (2019). An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series*, *1168*, 022022. https://doi.org/10.1088/1742-6596/1168/2/022022

Yong, A. B., Pitt, J. J., Montalto, J., Davies, H. E., Warne, G. L., & Connelly, J. F. (1988). Diagnosis of 21-hydroxylase deficiency in newborn infants by GC-MS of urinary steroids. *Aust Paediatr J*, *24*(5), 280-285. https://doi.org/10.1111/j.1440-1754.1988.tb01363.x

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32-35.

Zalas, D., Reinehr, T., Niedziela, M., Borzikowsky, C., Flader, M., Simic-Schleicher, G., Akkurt, H. I., Heger, S., Hornig, N., Holterhus, P. M., & Kulle, A. E. (2018). Multiples of Median-Transformed, Normalized Reference Ranges of Steroid Profiling Data Independent of Age, Sex, and Units. *Horm Res Paediatr*, *89*(4), 255-264. https://doi.org/10.1159/000488028

Zhang, H., Hamdane, D., Im, S. C., & Waskell, L. (2008). Cytochrome b5 inhibits electron transfer from NADPH-cytochrome P450 reductase to ferric cytochrome P450 2B4. *J Biol Chem*, *283*(9), 5217-5225. https://doi.org/10.1074/jbc.M709094200

Zuber, M. X., Simpson, E. R., & Waterman, M. R. (1986). Expression of bovine 17 alpha-hydroxylase cytochrome P-450 cDNA in nonsteroidogenic (COS 1) cells. *Science*, *234*(4781), 1258-1261. https://doi.org/10.1126/science.3535074