

UNIVERSIDADE DE LISBOA  
FACULDADE DE CIÊNCIAS  
DEPARTAMENTO DE BIOLOGIA VEGETAL



## **Anthropometric Data Analytics: a Portuguese Case Study**

António Pedro Pereira Barata

**Mestrado em Bioinformática e Biologia Computacional**  
Especialização em Bioinformática

Dissertação orientada por:  
Francisco José Moreira Couto  
Lucília da Conceição Mourão de Carvalho Cáceres Monteiro



*"Sometimes science is more art than science, Morty.  
A lot of people don't get that."*

– Justin Roiland



## **Agradecimentos**

Todo o organismo pode ser interpretado como um somatório de instantes. Ao longo do tempo, em cada instante, há troca de informação entre o indivíduo e o ambiente. De modo a proporcionar a continuidade do ser, a memória será uma das características de maior relevância que advém da experienciação da vida. O processo de aprendizagem será a retenção, interpretação, e utilização de informação passada. A capacidade de formular hipóteses face novos problemas, utilizando conhecimento adquirido através de eventos passados, tem para mim uma conotação fascinante. Fascinante é, também, o facto destas propriedades algorítmicas da mente humana em pouco ou nada me servirem para o exercício de elaboração de tese uma de mestrado. Quaisquer noções empíricas previamente adquiridas servem apenas para providenciar algum tipo de aconchego contra a perceção tardia da realidade. É nesse momento de maximização da variável incompetência que a contribuição intelectual de outros tem maior validade e aplicação prática. Por esse fator, é necessário mencionar as entidades sem as quais nenhum deste trabalho seria possível.

Independentemente de ordem, a todos agradeço igualmente. Considero que a omissão de um qualquer interveniente tenha a mesma consequência: a ausência do produto final que é o presente trabalho. Ao Professor Doutor Francisco Couto, agradeço pela sua paciência aparentemente infinita enquanto meu orientador, pela sua disponibilidade constante, associado ao seu sentido crítico e construtivo. Agradeço em particular à Dr.<sup>a</sup> Sância Ramos, diretora do serviço de Anatomia Patológica do Centro Hospitalar de Lisboa Ocidental, e à Dr.<sup>a</sup> Lucília Carvalho, minha co-orientadora e assistente com grau de consultor do mesmo serviço e responsável pelo setor de Fetopatologia associado ao mesmo centro hospitalar, por me facultarem o seu tempo e amabilidade durante todo este longo e árduo percurso académico. Relativamente à minha família, tenho em elevada estima o facto de me terem sempre apoiado incondicionalmente; por me terem aturado e às minhas lamúrias, por partilharem comigo as suas experiências, opiniões, conselhos, e considerações. Aos meus amigos, por não só partilharem os seus ombros carinhosos como por me ajudarem a desanuviar nos momentos de maior ansiedade, agradeço imensamente.

## Resumo

Durante o período neonatal, para produzir corretamente um diagnóstico patológico e permitir assim uma reposta adequada, é imperativo realizar uma rigorosa estimacão acerca da idade gestacional do feto. Esta previsão é aplicada como ferramenta essencial para o aconselhamento parental de modo a providenciar um plano de cuidados perinatais apropriado. Durante uma autópsia fetal, a idade gestacional é uma variável a ter em consideracão, particularmente utilizada aquando de situaçoes de terminacão de gravidez medicamente assistida e/ou infanticídios. No nosso caso, foram colecionadas observaçoes representativas da populaçao Portuguesa da região Centro-Sul de Portugal através do procedimento de várias autópsias fetais, provenientes do Hospital de Egas Moniz (CHLO – Centro Hospitalar de Lisboa Ocidental). Desde há vários anos que o sector de fetopatologia deste hospital tem vindo a analisar e avaliar os casos de mortalidade fetal pertencentes à região Centro-Sul de Portugal. Cada caso de autópsia fetal produz um relatório representativo das medidas e pesos associados ao indivíduo em causa, entre outras informaçoes médicas relevantes; após a sua conclusao, cada relatório é arquivado num dossier (organizado cronologicamente). Este tipo de processamento e armazenamento de informacão não proporciona um acesso direto nem estruturado aos valores antropométricos específicos previamente registados, derivados de relatórios médicos elaborados durante um ou mais procedimentos de autópsia fetal. Cada relatório arquivado é, entao, tido em consideracão como independente de todos os outros casos, tornando trabalhoso e demorado qualquer abordagem ao estudo do seu conteúdo. Para enfrentar este desafio primário, foi necessário desenvolver uma base de dados, assim como toda a metodologia relacionada com a inserçao de dados na mesma.

Neste presente estudo, um banco de dados nada mais é senao um depósito seguro para informacão, servindo o propósito de acomodar estruturalmente dados. Foram registados 24 parâmetros fetais para cada caso individual, incluindo idade gestacional e mediçoes de distâncias e pesos de características antropométricas e órgãos, respetivamente. Obtidas de acordo com o protocolo em vigor, segue a exaustiva lista de mediçoes fetais registadas em cada autópsia: idade gestacional, comprimento total, comprimento craniocaudal, perímetro cefálico, perímetro torácico, perímetro abdominal, comprimento de pé, comprimento da mão, comprimento do dedo médio, distância intercomissural, comprimento do filtro, distância entre os cantos internos, distância entre os cantos externos, comprimento da fenda palpebral esquerda, comprimento da fenda palpebral direita, comprimento do pavilhão auricular esquerdo, comprimento do pavilhão auricular direito, peso corporal, peso dos rins, peso do timo, peso do baço, peso do fígado, peso dos pulmões, e peso das glândulas suprarrenais. Órgãos emparelhados (pulmões, por exemplo) são representados pelo seu peso combinado. Como unidades, são utilizadas semanas (idade gestacional), centímetros (comprimentos e distâncias), e gramas (pesos). Foi gerado código base para produzir programas capazes de criar e interagir com o construto. Após estipular a estrutura da base de dados, todos os processos de inserçao e consulta de informacão são geridos por algoritmos especificamente engendrados de modo a prevenir a adulteracão não propositada dos dados registados. A linguagem de programacão adotada foi Python, versão 2.7 devido às suas bibliotecas (notavelmente: SQLite3, NumPy, e SciPy) e por ser uma linguagem multiparadigmática.

A estrutura da base de dados é simples, apesar de relacional. É constituída por uma tabela em que linhas e colunas representam, respetivamente, os indivíduos e os valores dos seus parâmetros fetais registados durante a autópsia (incluindo uma chave primária). Assim, cada linha é representativa de um relatório de autópsia fetal, com a sua própria identidade, e medidas e pesos associados. Tal como a nossa base de dados, simples é também o mecanismo de inserçao de dados. Todos os relatórios escritos tiveram de ter a sua informacão transferida para o formato digital. Para esse efeito, foi desenvolvido um programa de apoio à inserçao de dados. Aquando da sua execuçao, surge uma interface compreensível que solicita iterativamente ao utilizador os valores registados de cada variável de um relatório de autópsia fetal. Assim que todos os campos estejam preenchidos, a informacão recolhida é automaticamente inserida na base de dados, simbolizando um indivíduo e os seus

respetivos atributos. Uma vez preenchida a base de dados com toda a informação necessária, é possível propor uma análise adequada. Na totalidade, recolhemos a informação referente a 450 fetos entre as 13 e as 42 semanas de idade (gestacional). Para o devido efeito, a manipulação de informação foi executada utilizando objetos abstratos baseados em tabelas de dispersão (Python) e SPSS.

Este trabalho procurou abordar a precisão de diferentes parâmetros fetais em termos de estimação da idade gestacional, fazendo uso de técnicas de regressão e análise em componentes principais (ACP). Na computação dos 2 modelos de regressão linear múltipla, foram utilizados algoritmos específicos de retenção de variáveis baseados na análise de variância (estatística- $F$ ). Enquanto ACP e regressões múltiplas foram processadas em SPSS, regressões polinomiais foram executadas em Python. Para cada uma das 23 variáveis (referente a todos os parâmetros fetais selecionados com a exceção de idade gestacional), foram calculadas regressões polinomiais de grau  $k$ ,  $k \in \{1, 2, 3, 4, 5\}$ , derivadas de cada conjunto de pares de pontos variável-idade. Para todas as regressões, múltiplas e polinomiais, os valores de  $R^2$  (coeficiente de determinação) foram registados com um valor- $p$  significativo contra a hipótese nula de que os coeficientes estimados de cada parâmetro são iguais zero. Os modelos de regressão foram comparados entre si, com base na proporção de variância da variável dependente (idade gestacional) previsível pela(s) variável(eis) independente(s), isto é, o erro associado a cada modelo (soma do quadrado dos resíduos). Tendo sido estabelecido um nível de significância de  $\alpha = 0.05$ , cada modelo de regressão linear múltipla foi comparado a cada um dos outros modelos de regressão (polinomial e linear múltipla); modelos polinomiais foram comparados a outros modelos derivados do mesmo tipo de regressão se e só se partilhassem o mesmo grau  $k$ .

Relativamente à ACP (com um índice de KMO de 0.972 e um valor de significância próximo de 0 para a homocedasticidade), a proporção de variância partilhada entre cada variável (comunalidade) apresentou maior valor para as variáveis comprimento total, comprimento craniocaudal, comprimento do pé. Associativamente, o único componente principal retido (com valor próprio maior ou igual a 1) apresenta valores de correlação maiores entre esses mesmos parâmetros originais (loadings) do que com qualquer outra variável. Podemos colocar a hipótese, então, de que essas variáveis sejam consideradas possíveis marcadores de desenvolvimento (preditores confiáveis de idade gestacional). De acordo com os algoritmos de seleção de variáveis (SPSS) utilizados para a computação de regressões lineares múltiplas, foram criados 2 modelos explicativos de idade gestacional. Estes modelos apresentaram valores de coeficiente de determinação semelhantes ( $R^2 \approx 0.953$ ), assim como valores de teste Durbin-Watson adequados. As variáveis retidas apresentadas pelos 2 algoritmos foram semelhantes entre si, exceto para as variáveis representativas de comprimentos total e craniocaudal, que se verificaram como sendo mutualmente exclusivas. Em ambos os modelos, as variáveis selecionadas foram, em ordem decrescente de pesos- $\beta$ : peso corporal ( $\beta \approx 0.393$ ), comprimento do pé ( $\beta \approx 0.347$ ), comprimento total ( $\beta \approx 0.266$ ), comprimento craniocaudal ( $\beta \approx 0.199$ ), pavilhão auricular esquerdo ( $\beta \approx 0.16$ ), peso dos pulmões, e peso das glândulas suprarrenais. Para as últimas duas variáveis mencionadas, o valor absoluto do peso- $\beta$  foi menor ou igual a 0.1. Através de comparações entre modelos polinomiais foi possível estabelecer um sistema de classificação para variáveis ou grupos de variáveis, indicativa da qualidade de cada variável (associada a um grau de polinómio) em estimar, de acordo com os nossos dados, a idade gestacional. O grupo de variáveis com maior valor para o coeficiente de determinação, para cada grau polinomial, conteve sempre as variáveis comprimento total, comprimento craniocaudal, e comprimento do pé. De entre todas as regressões, comprimentos total, craniocaudal, e do pé estão constantemente presentes nos grupos de melhores previsores de idade gestacional. Mediante o tipo de regressão aplicada, o peso corporal e o comprimento da mão são também variáveis pertencentes à categoria preditiva anterior.

### **Palavras-chave**

Fetopatologia; Feto; Previsão; Idade gestacional; Agrupamento.

## **Abstract**

Large amounts of information are systematically generated throughout the course of scientific research and progress. In our case, observations representing the Portuguese population within the central-southern region of Portugal were collected throughout various foetal autopsy procedures. Gestational age (GA) and measured distances and weights of numerous anthropometric features and organs, respectively, were recorded per singleton (24 variables in total). This work seeks to elaborate on the accuracy of different foetal parameters in terms of GA estimation, making use of principal component analysis (PCA) and regression techniques. We created a dataset of 450 fetuses, ranging from 13 to 42 weeks of age, to compute both PCA and regression models. Initial exploratory analysis shed light onto which variables are most explanatory in terms of foetal development, and are thus most likely suitable for predictive rolls. We produced clusters of models, based on coefficient of determination values ( $R^2$ ), by comparing the squared sum of residuals between models (significance level  $\alpha = 0.05$ ). Models comprised of linear combinations of different variables exhibited significantly higher values of  $R^2$  ( $p$ -value  $\leq 0.05$ ) when compared to single variable models. Multiple linear regression models, however, did not exhibit the same statistical significance when compared internally. Across all regression models (both polynomial and multiple linear), crown-heel length (CHL), crown-rump length (CRL), and foot length (FL) are constantly present within the cluster of best predictors of GA. Depending on the type of regression analysis applied, body weight (Body), hand length (HL) also fall onto the same category. Consistent with previously peer-reviewed work, variables such as CHL, CRL, and FL are found to be the most reliable sources of information for estimating developmental age. In cases where such measurements are impossible to obtain, other foetal features can be utilized (although less reliable) such as HL, HC, body weight, and ear length.

## **Keywords**

Foetopathology; Foetus; Prediction; Gestational age; Clustering.



**Table of Contents**

**List of Figures** ..... **X**

**List of Tables** ..... **X**

**List of Abbreviations**..... **X**

**1 Introduction**..... **1**

    1.1 Motivation..... 1

    1.2 Objectives..... 1

    1.3 Results and Contributions ..... 2

    1.4 Overview ..... 2

**2 Case Study: Background and Related Work**..... **3**

    2.1 Foetal Viability ..... 3

    2.2 Measurement Relevance ..... 4

**3 Methods**..... **5**

    3.1 Data Structure ..... 5

    3.2 Data Exploration and Model Comparison..... 8

**4 Results** ..... **9**

    4.1 Principal Component Analysis..... 9

    4.2 Regression Models ..... 11

    4.3 Comparison and Clustering..... 12

**5 Discussion**..... **18**

    5.1 Final Remarks ..... 20

**6 Conclusion**..... **21**

    6.1 Future Work ..... 21

**References** ..... **22**

## List of Figures

2.1 Prenatal development .....	4
3.1 Excerpt from createDB.py script file .....	6
3.2 Example of insertValues.py script file instancing.....	6
3.3 Information workflow .....	7
4.1 Scree plot.....	11
4.2 1 <sup>st</sup> degree polynomial regression goodness of fit clusters .....	13
4.3 2 <sup>nd</sup> degree polynomial regression goodness of fit clusters .....	14
4.4 3 <sup>rd</sup> degree polynomial regression goodness of fit clusters.....	15
4.5 4 <sup>th</sup> degree polynomial regression goodness of fit clusters.....	16
4.6 5 <sup>th</sup> degree polynomial regression goodness of fit clusters.....	17

## List of Tables

4.1 PCA communalities and loadings .....	9
4.2 Total variance explained .....	10
4.3 Multiple linear regression models .....	11
4.4 Polynomial regression models .....	12

## List of Abbreviations

CHL – Crown-heel length	ICD – Inner canthal distance
CRL – Crown-rump length	OCD – Outer canthal distance
HC – Head circumference	LPFW – Left palpebral fissure width
CC – Chest circumference	RPFW – Right palpebral fissure width
AC – Abdominal circumference	LEL – Left ear length
FL – Foot length	REL – Right ear length
HL – Hand length	PL – Philtrum length
MFL – Middle finger length	ID – Intercommissural distance

## Chapter 1

### Introduction

Performing rigorous estimations of GA is invaluable for correct diagnosis and optimum treatment of disease during the neonatal period. GA prediction is an essential tool for parental counselling and to plan for appropriate perinatal care. It is also a prime requisite for foetal autopsy, particularly in situations of criminal abortion, alleged infanticide, and medically-terminated pregnancies. Previous peer-reviewed studies have elaborated on the accuracy of different foetal parameters in GA prediction<sup>1</sup>, particularly head circumference (HC), HL, FL, CRL, and CHL<sup>2 - 5</sup>. Model analysis and hypothesis tests may help determine not only how different measurements and weights are linked to foetal developmental age, but also which variables might be classified and ordered in terms of their predictive capabilities. Regarding anthropometric data analytics, other published papers often approach the validity of different measured variables for conceptual age estimation<sup>6 - 10</sup>, and the quantitative standards of those measurements for foetal and neo-natal autopsy<sup>11</sup>. Regression analysis and model fitting are widely accepted and used in this field of work, hence being viewed as reliable tools for knowledge production<sup>12</sup>. Other relevant publications may also be found, discussing the relationship between different methods of analysis and discriminating regression properties, enabling model validation for subsequent selection<sup>13, 14</sup>. Currently, the application of analytical and statistical methods for the evaluation of information is accomplished with the use of data manipulative software<sup>15, 16</sup>. For these computer programs to be beneficial, however, all data must be made digitally available. Without a proper data frame, analysis of data becomes tedious and/or unfeasible.

### 1.1 Motivation

The underlying importance of having a well-established database is not only to be able to reliably keep information safely stored, but also to enable such data to be subjected to manipulation and analysis. The foetopathology and pathological anatomy departments of Hospital de Egas Moniz (HEM), part of Centro Hospitalar de Lisboa Ocidental (CHLO), have since long been creating, gathering, and evaluating singleton pre-natal and neo-natal clinical autopsy records derived from the population located in the central-southern region of Portugal. Medical professionals register organ weights, anthropometric distances, and other features for each individual episode so that a diagnosis may be conjectured to explain the most likely cause of death. To make any causality assessments, each measured variable must be associated with the value of the most probable gestation period for that measurement; to this effect, a reference table of expected anatomical details at various postmenstrual gestational periods is utilized. Having fully concluded an autopsy report, all information regarding it is archived. Thousands of files are stacked in dossiers, making it effortful to inquire such data. Without the aid of a more suitable storage-query system, it is not feasible to produce any kind of meaningful studies relating the contents of different autopsy reports. A database would have to be created. Moreover, an efficient way of inputting and manipulating information had to be devised.

### 1.2 Objectives

Our primary objective in this study is to devise a method for storing and manipulating information pertaining to the autopsy files collected by CHLO medical professionals. Additionally, we hope to apply different algorithmic approaches to our collected data to, not only test the adequacy of our database, but also produce meaningful knowledge by using different methodological approaches (such as PCA, polynomial regression, and multiple regression techniques) that may help establish which foetal parameters are most associated to foetal development (gestational age). Another goal is to make our scripts simple and user-friendly, specifically in terms of database creation, and data insertion and visualisation. Thus, data interaction can be easily applied without much background information.

### 1.3 Results and Contributions

With the application of exploratory analysis and distinct regression techniques, by means of SPSS and Python scripts, it is possible to check which variables are either most explicative regarding foetal development variance or serve best as the basis for GA estimation models, respectively. High values for communality ( $\geq 0.946$ ) and loading ( $\geq 0.972$ ) can be witnessed for CRL, CHL, and FL variables, which account for the shared variance with every other variable and produced component, respectively. Another example for GA estimation predictor assessment, is the evaluation of models with highest coefficient of determination (CHL, FL, CRL, body weight, and HL) and highest variable  $\beta$ -weights associated with multiple linear regression (CHL, FL, CRL, body weight, and ear length). Moreover, by comparing different models in terms of their associated error, regarding statistical significance, it is possible to produce clusters of variables which present the same prediction accuracy, despite exhibiting different coefficient of determination values and thus create a variable-based ranking system for GA estimation; for example, CRL, CHL, FL, and body weight are clustered as the least error-prone models for a 2<sup>nd</sup> degree polynomial regression ( $0.936 \leq R^2 \leq 0.942$ ).

Through the course of this work, software was developed to enable information gathering and manipulation, and derive the newly proposed statistically significant cluster and ranking system (to be applied to variable regression models). Our contributions in the field of database creation and handling are made available in a public repository<sup>17</sup>, as well as the actual database constructed and utilized in our work. Also, a research article related to our findings (concretely, GA estimation and variable adequacy) has also been accepted to the 11th International Conference on Practical Applications of Computational Biology & Bioinformatics<sup>18</sup> (PACBB), serving as a perfect peer-review process by which the validity of our endeavours can be testified.

### 1.4 Overview

Based on foetal autopsy records, we created a dataset of 450 individuals, each comprised of 24 foetal parameters. PCA produced results indicating CHL, CRL, and FL variables as the most explanatory in terms of total data variance. By comparing regressions models, Body and HL parameters were also found to be significantly viable measurements for GA estimation, depending on the polynomial degree applied within each regression. We hope to reinforce the many advantages of data manipulation by computation over manual activity. With an adequately ample data set, it could be possible to establish, for example, certain specific pre-natal characteristics associated with a distinct disease, enabling pathology detection. Background information regarding this work is discussed in Section 2, which serves as context for the appreciation for our attempts and achievements. The following section describes the methodological approaches used (programming language used, noteworthy package applications, statistical approaches, etc.). Section 4 presents the results of applying said methods, which are mostly visual reference tables and figures. Discussion of obtained results and final remarks pertain to the 5th Section of this dissertation, where the properties of each approach are taken into consideration during result evaluation. Section 6 relates to the conclusions derived from our work, while attempting to foresee possible new outcomes, making use of our data.

## Chapter 2

### Case Study: Background and Related Work

Fetal and perinatal pathology is mainly a posthumous specialty concerned with the causes and mechanisms behind the reproductive loss in humans<sup>19</sup>. In its majority, pregnancy loss occurs in the first half of the gestational period<sup>20-24</sup>. Causes of death in this stage vary depending on the gestational age of an individual. For instance, chromosomal and genetic defects have their highest frequencies in earlier weeks – accounting for approximately 60% of all non-viable, and thus naturally aborted embryos – while infections and premature rupture of the membranes are most associated with mortality during later embryonic stages<sup>25-28</sup>.

For several years, the foetopathology department of Hospital de Egas Moniz, has been conducting the analysis and evaluation of foetal mortality cases pertaining to the central-southern region of Portugal. Each foetal autopsy produces a physical report file containing, amongst other relevant medical information, measurements and weights of the foetus. Whenever a foetopathology instance is concluded, the file is then archived within a dossier. This type of information processing and storage does not permit direct access to harboured values in more than a few cases at a time. Reports are regarded independently of each other, making any data study laborious and time-consuming. To address this challenge, we developed a database representing foetal autopsy records. Each report had to be manually inserted, due to discrepancies of cursive between files, excluding the use of optical character recognition (OCR) software.

### 2.1 Foetal Viability

Foetal viability is the ability or potential of the foetus to survive outside the uterus after birth while supported by modern medical technology<sup>29</sup>; an individual's viability is largely dependent upon its organ maturity and environmental conditions. There exists no well-defined set of developmental values – age, weight, or other measures – for which a human conceptus becomes automatically viable<sup>30</sup>. For instance, seldom does any infant weighing less than 500g persist outside the womb (although it has been reported). In accordance with the scientific community in this field, 20 to 35 percent of babies born at 23 weeks of gestation survive, while 50 to 70 percent of babies born at 24 to 25 weeks and over 90 percent born at 26 to 27 weeks survive<sup>31-33</sup>. Between weeks 23<sup>rd</sup> and 24<sup>th</sup> of gestation, an average individual's chance for survival is augmented 3 to 4% per day. From 24 to 26 weeks of development, the increment in viability per day is reduced to 2 to 3%. The following gestational periods exhibit a decrease in rate of viability augmentation, due to the already present high chance of survival.

The GA at which the expectation that a foetus has as much chance of surviving as not surviving post-partum is a medical concept known as the limit of viability. With the development and support of neonatal intensive care units (NICU) – a special department of a hospital or health care facility catering to ill or premature new-born babies – the limit of viability has been declining since half a century ago, although stagnant for the past 12 years<sup>34</sup>. The 50 percent survivability threshold is currently around the GA of 24 weeks<sup>35</sup>.

From all possible factors that affect an infant's chance of survival, the most influential are age, weight, gender, and race. Foetal viability is also influenced by several types of health problems: breathing problems, congenital abnormalities or malformations, and infections threaten the survival of the neonate<sup>36</sup>. Other factors may influence the foetus' ability to withstand birth by altering the rate of organ maturation or oxygen supply. Progeny whose maternal entity is conditioned by diabetes mellitus, as an example, have a higher mortality rate (comparatively to non-diabetic mothers).

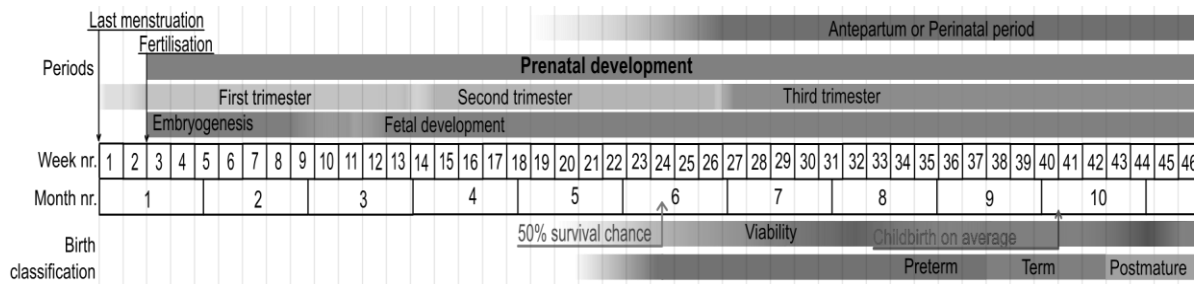


Figure 2.1: **Prenatal development.** Stages in prenatal development, presenting viability and point of 50% chance of survival (limit of viability) at bottom. Weeks and months are numbered by gestation. Source: Häggström M. Medical gallery of Mikael Häggström 2014. WikiJournal of Medicine. 2014, 1 (2).

## 2.2 Measurement Relevance

To analyse diagnoses and evaluate the recurrence risk for disadvantageous pregnancy conclusions, medical and pathological professionals gather and disclose various anatomical details during neonatal autopsies. While questing for meaningful answers, numerous types of information play an important part; family and personal health history of the parents, obstetric events, biometrics, radiography, histological examination, and laboratory studies, for example, are some of the paramount details required to produce any knowledge – and consequently, wisdom – in this field of work. For roughly one third of cases, a precise cause of death may not be accurately determined despite all comprehensive attempts performed<sup>37</sup>. Consequently, and due to lack of adequate explanation to family members for their affliction, socially-impaired mourning behaviours may rise<sup>38</sup>. Specialists aim to unearth distinct syndromic diagnoses as families are best supplied by having unambiguous acumen into future liabilities. Without normative tables, presumptively important findings such as hypoplasia – the underdevelopment or incomplete development of a tissue or organ – and hypertrophy – the increase in the volume of an organ or tissue due to the enlargement of its component cells – for example, may not be accurately denoted while clinically assessing an individual during any biometric procedure.

In 2006, the conjoint effort of John Archie, Julianne Collins, and Robert Lebel produced quantitative standards for foetal and neonatal autopsies. The data used to create such a construct was available at the time, sourcing from other information repositories which had been assembled by other researchers. Singleton foetal measurements and their associated gestational periods derived from specific circumstances: data was collected from different geographical origins<sup>39 – 53</sup> with varying gathering conditions; for example, normal term infants, electively aborted foetuses, and stillborn data were utilized, acquired from contrasting regression analysis models – linear and polynomial.

Portuguese professionals make use of this meta-analytical informational design daily to produce viable output from their gathered observations, providing overwhelming importance to the contribution of John Archie and his team. CHLO medical specialists select specific measurements and weights from the list of all variables studied throughout the foetal developmental process, while assessing any foetopathological event. Preceding any diagnosis, professionals must associate each measurement to a specific gestational period in weeks, following a unified table of lengths, distances, and weights. Discrepancies between age values from different variables within the same individual provide insight into determining a probable cause of death and/or factors most linked to fatality.

## Chapter 3

### Methods

For this study, a database is no more than a safe-deposit space for data. It serves the purpose of being able to structurally accommodate data, rendering it as information. Concretely, Python code is required to create and interact with said construct. In this fashion, once structure is defined, all processes of data insertion and query must fall onto the responsibility of specific code scripts; this provides a practical barrier against uncared practices towards data. For example, while it possible to easily visualize data by means of a spreadsheet, the information itself is kept separate from the observation-enabling file, thus not being directly possible to alter or delete any given values or structure within the informational scheme itself.

24 quantitative variables were selected to represent each foetal autopsy case. Retrieved according to autopsy protocol, the extensive list of recorded foetal parameters follows: GA, CHL, CRL, HC, chest circumference (CC), abdominal circumference (AC), FL, HL, middle finger length (MFL), intercommissural distance (ID), philtrum length (PL), inner canthal distance (ICD), outer canthal distance (OCD), left palpebral fissure width (LPFW), right palpebral fissure width (RPFW), left ear length (LEL), right ear length (REL), body, kidneys, thymus, spleen, liver, lungs, and adrenals. Paired organs are represented by their combined weight. Units comprise of week (GA), centimetre (distances and lengths), and gram (organ and body weights).

Given the format of each autopsy report file in this work, a database was constructed and algorithms to store, retrieve, and manipulate information were devised. Python 2.7 was applied as the programming language for these tasks mainly due to its extensive libraries and packages, notably SQLite3 (providing SQL interface compliant with the DB-API 2.0 specification described in PEP – Python Enhancement Proposal – 249), NumPy, and SciPy modules<sup>54–56</sup>, while also prioritizing code readability. Another Python-promoting key factor is that it facilitates script development by being multi-paradigmatic, fully supporting aspect-oriented, object-oriented, structured, imperative, functional, contract, and logic styles of programming<sup>57–63</sup>. IBM's SPSS software<sup>64</sup> was also utilized due to its inbuilt statistical applications, concretely PCA and variable selection algorithms for multiple linear regression.

### 3.1 Data Structure

The actual database structure utilized to store and retrieve information is a simplistic one. Despite being modelled as a relational database, no more than a single table was created due to the underlying nature of selected data. Within this specified database there exists a table where the first column corresponds to an identifier for every individual (primary key), and each other column represents a certain variable of interest. Hence, each row denotes a singleton foetal autopsy report, with its own identity, and associated measures and weights.

To create a .db extension file using Python (createDB.py), it is firstly necessary to create a Connection object (herein referred to as conn) representing the database. Once conn is established, it is imperative to conceive a Cursor object (derived from conn's cursor() method) and make use of its execute() method to perform SQL commands. Therefore, a command variable must be initialized with a string attributed to it, representing the SQL-syntax statement for database creation. As a DB-API (database-application programming interface) requirement, after connecting to the database and thus ensuing a new transaction, it is necessary to confirm any alterations made: the commit() method, belonging to the conn object, applies such confirmations. During a database creation procedure, the referenced method is not strictly necessary, but rather demonstrates good coding practice as it is required when inserting, deleting, or updating values within the database. Lastly, to terminate the previously established connection, the close() method (belonging to conn) is evoked.

```

import sqlite3 as sql
conn = sql.connect(`dbName.db`)
c = conn.cursor()
command = `CREATE TABLE tableName(idd VARCHAR(255) PRIMARY KEY, ..., GA FLOAT);`
c.execute(command)
conn.commit()
conn.close()

```

Figure 3.1: **Excerpt from createDB.py script file.** Snippet of Python code used to produce the database. All variables are declared as FLOAT, except for idd (type VARCHAR) which is the primary key. Exemplary, only 2 variables are depicted. When run, outputs a dbName.db file consisting of a database with the specified table properties.

A simple database structure should be accompanied by a straightforward data insertion mechanism. All scribed reports had to be translated into the form of digital information. To achieve this, a Python script was formulated (insertValues.py) to aid the exhausting task of allocating all data heretofore gathered into the specified database. When run, a user-accessible interface emerges, iteratively requesting the recorded values of each variable within an autopsy log. While running the insertion script, the names displayed for each variable are derived from the variable names given when creating the database. After all fields are filled, the gathered information is automatically inserted into the database, symbolizing a unique individual and its corresponding features. A SQLite3 module approach is used, in resemblance to the previously mentioned database creation script, to execute SQL-syntax commands for data insertion.

```

idd:      A 05 00199
GA:       34
Body:     1385
CHL:      39
CRL:      26.5
HC:       29.5
CC:       24
AC:       21.5
FL:       5.5
HL:       4.6
MFL:      2
ICD:      2
OCD:      4.5
LPCFW:    1.2
RPFW:     1.3
LEL:      2.3
REL:      2.5
PL:       0.6
ID:       2.1
Lungs:    24.32
Thymus:   3.73
Liver:    55.54
Spleen:   3.51
Adrenals: 1.46
Kidneys:  14.05

```

Figure 3.2: **Example of insertValues.py script file instancing.** Practical application of the devised algorithm for value insertion. All numerical values are considered as floating point numbers when inserted into the database. In this case, the 199<sup>th</sup> autopsy report from the year 2005 is displayed. Variable names prompted by the script are shown to the left, while user input values are shown to the right. Noticeably, idd was selected as the primary key variable name instead of id; this decision was made because ID was already utilized as the variable for processing intercommissural distance.

Once our database has been populated with all necessary information, it is possible to elaborate on that information so that proper analysis can ensue. For this purpose, a third Python script (analysis.py) was created not only to retrieve information from the database, but also produce meaningful output. This output comes in two different forms. One type of output is merely a data frame containing data



(for example, a .csv extension file). This output is then manually imported onto IBM’s SPSS software through the Import Data option, for the application of exploratory and multiple regression analyses. The second type of output consists of every other result enunciated within this thesis (polynomial regression and clustering, for example), described along this work, including SPSS output analysis.

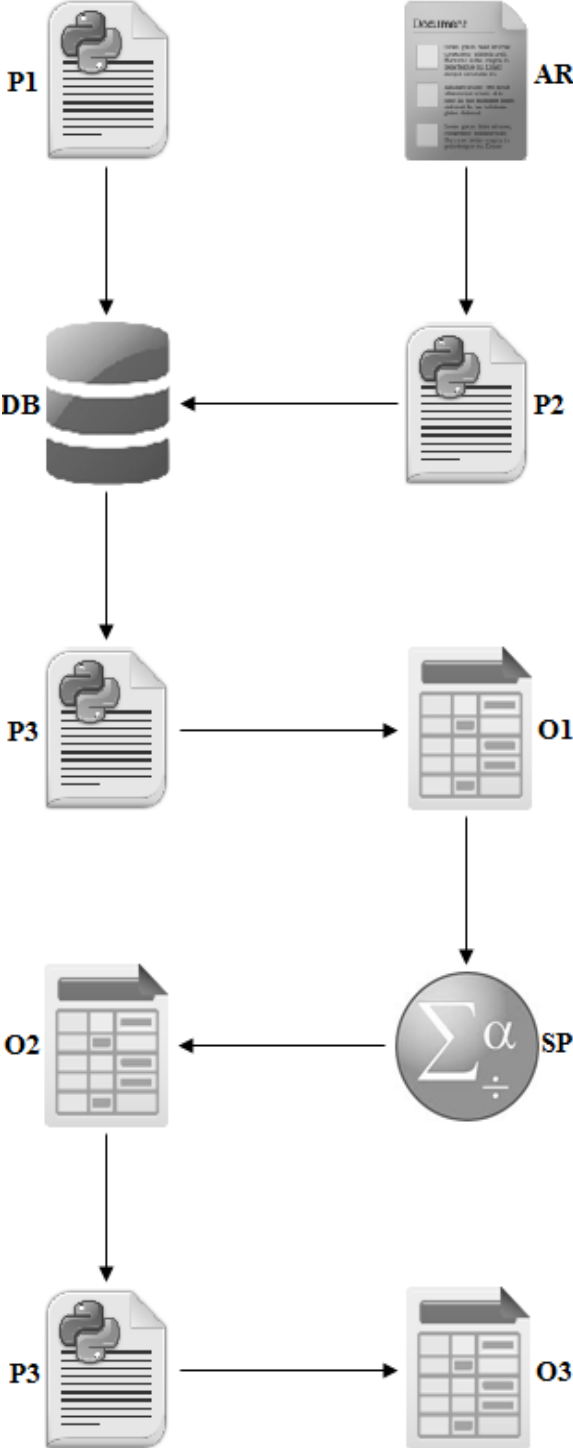


Figure 3.3: **Information workflow.** Practical illustration of all tasks and items required within this work. Arrows indicate the direction of information flow. P1 (createDB.py) creates the specified database (DB); P2 (insertValues.py) inserts all information retrieved from an AR (Autopsy Report) into DB; P3 (analysis.py), after retrieving data from the database, outputs a .csv file (O1) containing all DB information; O1 is manually passed onto IBM’s SPSS software (SP), which outputs its analysis results (O2); O2 is manually incorporated onto P3, which outputs the end results (O3) shown in this dissertation.

### 3.2 Data Exploration and Model Comparison

SPSS was used to conduct the initial PCA, which would provide foresight onto possible outcomes of successive regression models. Computed extraction communalities, loadings, explained variance per component, and adequacy parameters were consequently inspected. Computation of multiple linear regression models was performed through the same IBM software. GA was selected as the dependent variable, while the remaining 23 features were used as predictors. All available regression algorithms for variable selection (Enter, Stepwise, Remove, Backward, and Forward) were utilized and their outputs taken into consideration. Models were selected based on statistically significant coefficient values ( $\alpha = 0.05$ ), as well as Durbin-Watson and  $R^2$  values. Standardized  $\beta$ -weights were also a point of interest for later model comparison.

In total, 5 different  $k^{\text{th}}$  degree polynomial regression functions were fit onto each of the 23 variables, for  $k \in \{1, 2, 3, 4, 5\}$ . Each variable dataset consisted of pairs of variable-age points, where each pair represents the GA and recorded variable value of a singleton foetus. The NumPy module `polyfit()` function was used to output each single variable model.  $R^2$  and estimated parameter values were recorded for all regressions presenting a significant  $p$ -value for the null hypothesis that the estimated coefficients are equal to zero.

Regression models were compared based on each model's proportion of variance in the dependent variable predictable by the independent variable. The F-statistic was selected and computed using the squared sum of residuals (SSR) and degrees of freedom of the models being compared<sup>65</sup>. A significance level of  $\alpha = 0.05$  was established. Each multiple linear regression model was compared to all other multiple and polynomial models, while polynomial models were compared to other polynomial models if and only if both models pertained to the same polynomial degree. When comparing 2 models with the same degree of freedom, the F-statistic was computed as

$$F = \frac{SSR1}{SSR2}$$

where SSR1 and SSR2 indicate the squared sum of residuals for each model being compared. The upper critical value of the F distribution was then deduced with both numerator and denominator values equal to the degrees of freedom of either model. In contrast, to compare models exhibiting different degrees of freedom (polynomial versus multiple regressions), the F-statistic was computed as

$$F = \frac{(SSR1 - SSR2)/(df1 - df2)}{(SSR2/df2)}$$

where df1 and df2 represent the degrees of freedom of the first and second models, respectively. The first model must be the one with fewer parameters between the 2 models being compared. The upper critical value of the F distribution (which is directly related to the computed  $p$ -value) is deduced for a numerator value of the difference between df1 and df2, and a denominator value of df2.

The SciPy module `stats.f.cdf()` function was used to compute all  $p$ -values associated with the previously computed F-statistics. For models with the same degrees of freedom between them,

$$p\text{-value} = 1 - \text{stats.f.cdf}(F, df1, df2)$$

while for models with different degrees of freedom between them,

$$p\text{-value} = 1 - \text{stats.f.cdf}(F, df1 - df2, df2)$$

## Chapter 4

### Results

#### 4.1 Principal Component Analysis

For our dataset, the Kaiser-Meyer-Olkin (KMO) index for sampling adequacy had a value of 0.973 while the  $p$ -value corresponding to the  $\chi^2$ -statistic associated with Bartlett's test of homoscedasticity was below  $5 \times 10^{-4}$ . PCA produced only one significant component (eigenvalue  $\geq 1$ ) explaining 79.624% of total data variance. Communality and loading values for all variables are shown below, as well as total variance explained across components and scree plot (component versus eigenvalue).

Table 4.1: **PCA communalities and loadings.** PCA-generated communality and loading values per variable within our dataset. Darker shades represent lower values.

	Communality	Loading
<b>CRL</b>	0.963	0.981
<b>CHL</b>	0.956	0.978
<b>FL</b>	0.946	0.972
<b>GA</b>	0.937	0.968
<b>HC</b>	0.931	0.965
<b>Body</b>	0.925	0.962
<b>REL</b>	0.924	0.961
<b>LEL</b>	0.918	0.958
<b>AC</b>	0.908	0.953
<b>OCD</b>	0.897	0.947
<b>MFL</b>	0.872	0.934
<b>Liver</b>	0.847	0.921
<b>Kidneys</b>	0.804	0.897
<b>Lungs</b>	0.800	0.894
<b>RPFW</b>	0.800	0.894
<b>LPFW</b>	0.781	0.884
<b>ICD</b>	0.743	0.862
<b>Spleen</b>	0.695	0.834
<b>Adrenals</b>	0.694	0.833
<b>Thymus</b>	0.679	0.824
<b>PL</b>	0.651	0.807
<b>CC</b>	0.572	0.756
<b>HL</b>	0.460	0.678
<b>ID</b>	0.406	0.637

Table 4.2: **Total variance explained.** PCA-generated eigenvalue per component produced and associated percentage of total explained variance. Darker shades represent non-retained components.

<b>Component</b>	<b>Eigenvalue</b>	<b>% Total Explained Variance</b>
<b>1</b>	19.11	79.624
<b>2</b>	0.921	3.839
<b>3</b>	0.585	2.437
<b>4</b>	0.558	2.325
<b>5</b>	0.46	1.916
<b>6</b>	0.394	1.641
<b>7</b>	0.36	1.5
<b>8</b>	0.296	1.234
<b>9</b>	0.251	1.047
<b>10</b>	0.24	0.998
<b>11</b>	0.153	0.638
<b>12</b>	0.141	0.59
<b>13</b>	0.11	0.458
<b>14</b>	0.094	0.393
<b>15</b>	0.065	0.273
<b>16</b>	0.06	0.249
<b>17</b>	0.048	0.201
<b>18</b>	0.045	0.186
<b>19</b>	0.032	0.135
<b>20</b>	0.029	0.119
<b>21</b>	0.019	0.078
<b>22</b>	0.012	0.05
<b>23</b>	0.011	0.045
<b>24</b>	0.006	0.023

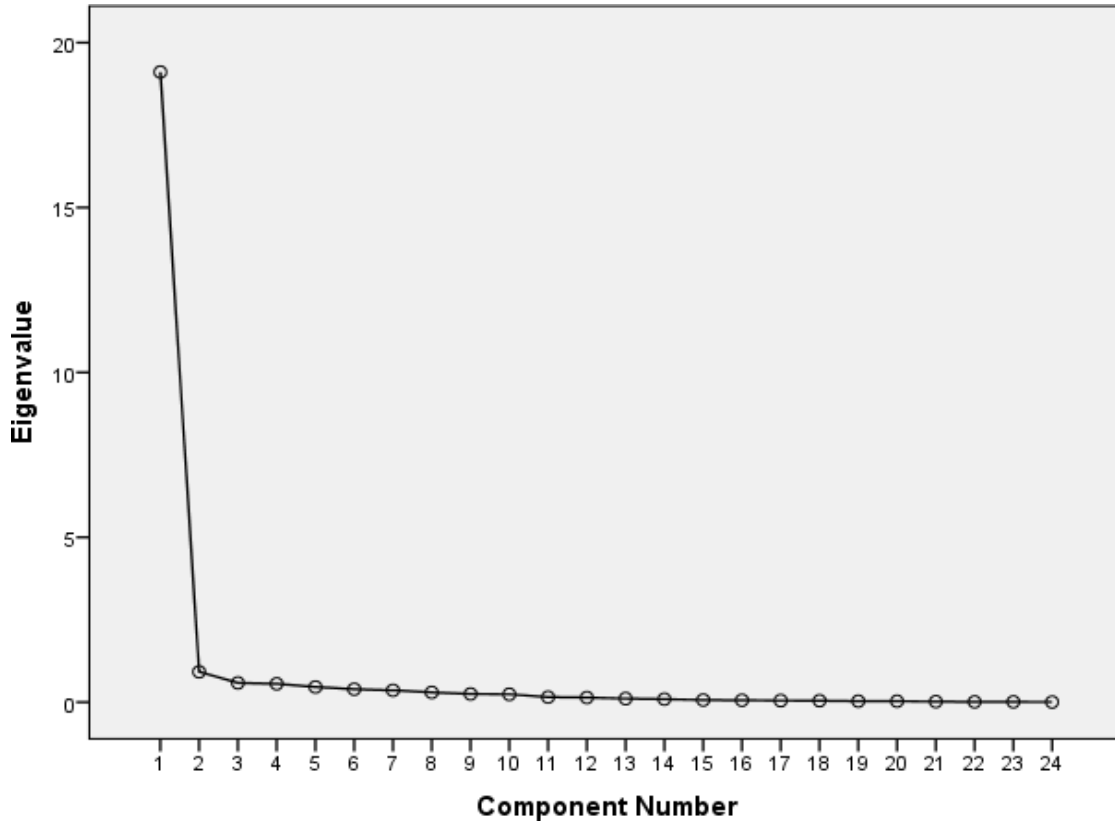


Figure 4.1: **Scree plot.** Eigenvalues of associated components versus the number of the component.

## 4.2 Regression Models

Across all variable selection methods for multiple regression, outputs presenting models with non-significant variable coefficients were excluded (Enter and Remove). The Backward selection algorithm was discarded for presenting the same output as the Forward approach, while yielding a Durbin-Watson statistic further away from 2. Stepwise and Forward algorithms produced models with Durbin-Watson values of 1.961 and 1.958, respectively, and similar coefficients of determination values ( $R^2 \approx 0.953$ ). Both regressions share 5 retained variables, one exclusive variable each. Only statistically significant variable coefficients are present in either model ( $p$ -value  $\leq 0.05$ ).

Table 4.3: **Multiple linear regression models.** Standardized  $\beta$ -weights for each variable selected associated with each variable selection algorithm method for regression. Darker shades represent lower values.

	Body	FL	CHL	CRL	REL	Lungs	Adrenals
Stepwise	0.402	0.310	0.266	-	0.157	-0.070	-0.087
Forward	0.384	0.384	-	0.199	0.163	-0.069	-0.083

In terms of polynomial regression, a collection of 115 single variable-based models for GA estimation were generated, comprised of 5 different degree polynomial regressions for each of the 23 independent variables. Models were retained after checking the statistical significance of each model's estimated parameters ( $p$ -value  $\leq 0.05$ ). Every  $k^{\text{th}}$  degree polynomial regression model follows the form

$$f(x) = \beta_k \cdot x^k + \beta_{k-1} \cdot x^{k-1} + \dots + \beta_0 \cdot x^0$$

where  $\beta_k, \beta_{k-1}, \dots, \beta_0$  are the computed weights associated with variable  $x$ , for any polynomial degree  $k$ .

Table 4.4: **Polynomial regression models.**  $R^2$  values computed for all polynomial regressions. Polynomial degrees are represented by numbers 1 through 5, for each variable-derived model. Darker shades represent lower values.

	1	2	3	4	5
<b>CHL</b>	0.931	0.942	0.943	0.943	0.944
<b>FL</b>	0.927	0.940	0.942	0.945	0.945
<b>Body</b>	0.868	0.937	0.942	0.942	0.942
<b>CRL</b>	0.931	0.936	0.938	0.940	0.940
<b>HL</b>	0.410	0.917	0.930	0.934	0.936
<b>HC</b>	0.896	0.911	0.914	0.916	0.917
<b>REL</b>	0.893	0.902	0.904	0.907	0.907
<b>LEL</b>	0.885	0.891	0.895	0.896	0.896
<b>Kidneys</b>	0.734	0.876	0.877	0.881	0.881
<b>CC</b>	0.503	0.871	0.883	0.898	0.899
<b>MFL</b>	0.849	0.864	0.917	0.917	0.920
<b>AC</b>	0.840	0.840	0.852	0.853	0.857
<b>Liver</b>	0.759	0.840	0.842	0.843	0.843
<b>OCD</b>	0.834	0.835	0.854	0.857	0.860
<b>Lungs</b>	0.720	0.808	0.813	0.814	0.816
<b>Spleen</b>	0.623	0.791	0.833	0.847	0.849
<b>RPFW</b>	0.730	0.759	0.800	0.803	0.809
<b>Thymus</b>	0.608	0.756	0.816	0.820	0.820
<b>LPFW</b>	0.711	0.738	0.777	0.779	0.784
<b>ICD</b>	0.710	0.726	0.742	0.750	0.751
<b>ID</b>	0.363	0.715	0.722	0.777	0.787
<b>Adrenals</b>	0.589	0.681	0.689	0.691	0.692
<b>PL</b>	0.595	0.598	0.606	0.606	0.608

### 4.3 Comparison and Clustering

In terms of multiple linear regression, both previously selected models exhibited no statistically significant difference between them. In contrast, when either model was compared to any of the 115 polynomial regression models, a recurring  $p$ -value  $\leq 0.05$  was systematically observed. By clustering models presenting no significant difference between other variable models, and creating different variable clusters based on statistical evidence for divergence, a goodness of fit hierarchy was established. CHL, CRL, and FL were the only single parameter-based regressions to be present in the top tier throughout all polynomial degrees. The hierarchical dissimilarities were most evident between 1st degree polynomial regressions and the remaining polynomial degree models. Notably, body weight was placed alongside the best GA estimators for any polynomial degree  $\geq 2$ , as HL for any degree  $\geq 3$ .

0.931	CRL			
0.931	CHL			
0.927	FL			
0.896				HC
0.893				REL
0.885			LEL	<b>LEL</b>
0.868		Body	<b>Body</b>	Body
0.849	MFL	<b>MFL</b>	MFL	
0.840	<b>AC</b>	AC		
0.834	<b>OCD</b>	OCD		
0.759		Liver	<b>Liver</b>	
0.734	Kidneys	<b>Kidneys</b>	Kidneys	
0.730	RPWF	<b>RPWF</b>	RPWF	
0.720	Lungs	<b>Lungs</b>	Lungs	
0.711	<b>LPFW</b>	LPFW		
0.710	<b>ICD</b>	ICD		
0.623	Spleen			
0.608	Thymus			
0.595	PL			
0.589	Adrenals			
0.503	CC			
0.410	HL			
0.363	ID			

Figure 4.2: **1<sup>st</sup> degree polynomial regression goodness of fit clusters.** Numerical values represent the coefficients of determination of each aligned variable. Darker shades represent lower  $R^2$  values. Clusters are represented by boxes. Parameters in bold indicate cluster centre(s): variable models used as subject for model comparison across other higher  $R^2$  valued variable models. For example, while AC and OCD models (as a cluster centre) are statistically indistinguishable from MFL and one another, both have a significantly worse fit when compared to any other given model with a higher  $R^2$  value; MFL (as a cluster centre) is statistically identical to Body, and both AC and OCD models, and significantly different from every other model.





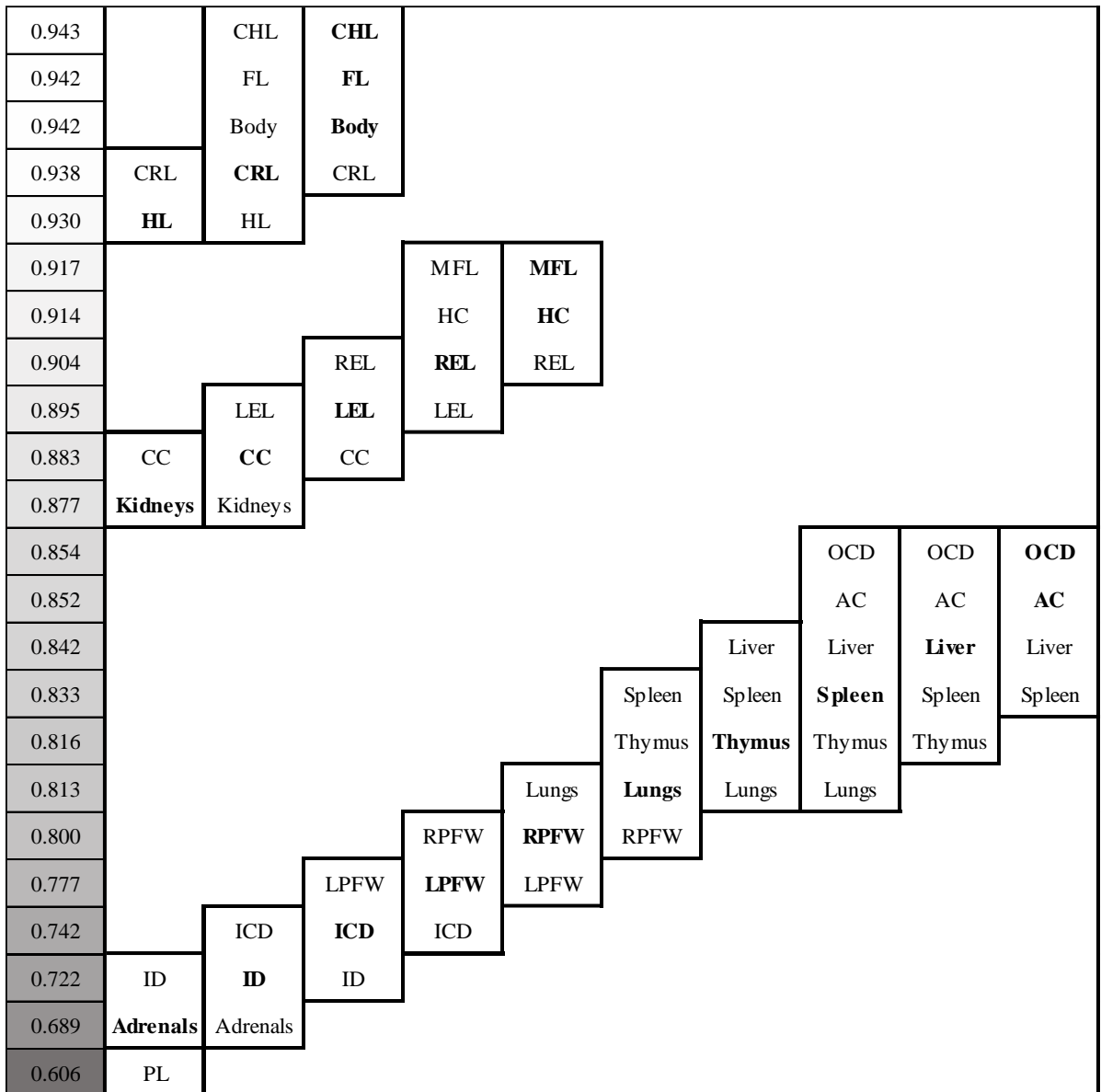


Figure 4.4: 3<sup>rd</sup> degree polynomial regression goodness of fit clusters. Numerical values represent the coefficients of determination of each aligned variable. Darker shades represent lower  $R^2$  values. Clusters are represented by boxes. Parameters in bold indicate cluster centre(s).

0.945		FL	<b>FL</b>		
0.943	CHL	<b>CHL</b>	CHL		
0.942	Body	<b>Body</b>	Body		
0.940	CRL	<b>CRL</b>	CRL		
0.934	<b>HL</b>	HL			
0.917			MFL	<b>MFL</b>	
0.916			HC	<b>HC</b>	
0.907		REL	<b>REL</b>	REL	
0.898	CC	<b>CC</b>	CC		
0.896	LEL	<b>LEL</b>	LEL		
0.881	<b>Kidneys</b>	Kidneys			
0.857				OCD	<b>OCD</b>
0.853				AC	<b>AC</b>
0.847				Spleen	<b>Spleen</b>
0.843				Liver	<b>Liver</b>
0.820			Thymus	Thymus	<b>Thymus</b>
0.814			Lungs	<b>Lungs</b>	Lungs
0.803		RPCL	<b>RPCL</b>	RPCL	RPCL
0.779	LPCL	<b>LPCL</b>	LPCL		
0.777	ID	<b>ID</b>	ID		
0.750	<b>ICD</b>	ICD			
0.691	Adrenals				
0.606	PL				

Figure 4.5: 4<sup>th</sup> degree polynomial regression goodness of fit clusters. Numerical values represent the coefficients of determination of each aligned variable. Darker shades represent lower  $R^2$  values. Clusters are represented by boxes. Parameters in bold indicate cluster centre(s).

0.945	FL								
0.944	CHL								
0.942	Body								
0.940	CRL								
0.936	HL								
0.920				MFL	MFL				
0.917				HC	HC				
0.907		REL	REL	REL	REL				
0.899		CC	CC	CC					
0.896	LEL	LEL	LEL	LEL					
0.881	Kidneys	Kidneys							
0.860						OCD	OCD		
0.857						AC	AC		
0.849						Spleen	Spleen		
0.843						Liver	Liver		
0.820				Thymus	Thymus	Thymus	Thymus		
0.816			Lungs	Lungs	Lungs	Lungs			
0.809		RPCL	RPCL	RPCL	RPCL	RPCL			
0.787	ID	ID	ID	ID	ID				
0.784	LPCL	LPCL	LPCL	LPCL					
0.751	ICD	ICD	ICD						
0.692	Adrenals								
0.608	PL								

Figure 4.6: **5<sup>th</sup> degree polynomial regression goodness of fit clusters.** Numerical values represent the coefficients of determination of each aligned variable. Darker shades represent lower  $R^2$  values. Clusters are represented by boxes. Parameters in bold indicate cluster centre(s). Comparatively to the previous table, HL is now indistinguishable from any of the top 5 predictors.

## Chapter 5

### Discussion

The adequacy of exploratory analysis by PCA, applied to our dataset, can be determined by inspecting the results from Bartlett's test of sphericity – statistical test for the overall significance of all correlations within the correlation matrix – and the Kaiser-Meyer-Olkin test for sampling adequacy (KMO index). Without statistical significance of correlations, the remaining outputs of PCA (components, communalities, and loadings, for instance) would be statistically invalid. The KMO index (values ranging from 0 to 1), once correlation significance has been inferred, indicates how efficiently our original variables can be factorized, given that the correlation between any 2 variables can be influenced by any other given variable present within the dataset. A sphericity test significance value lower than  $5 \times 10^{-4}$ , and a KMO index of 0.973 indicate that our dataset is viable for a PCA approach. Having a total of 24 variables and a dataset comprised of 450 individuals helped stipulate which component retention criterion to be used<sup>66</sup>. For this reason, a single component was selected with a corresponding eigenvalue of 19.1. This component exhibited a percentage of total variance explained of 79.62, which is adequate<sup>67</sup>. The total amount of variance shared between each variable and every other parameter within our analysis (communality) presented higher values for variables such as CRL, CHL, and FL (order from highest to lowest). On a similar note, for our retained principal component (which can be described as a developmental marker), loading values have the same variable-value order (as in the communalities table), which translates into the correlation between the original variables and that component. For variables yielding high loading values (CRL, CHL, FL, for example), one can assume those variables might be considered as potentially good developmental markers (or rather, reliable GA predictors). Such claims, however, can only be induced by different methods, such as regression analysis.

An important step in regression model validation is testing the hypothesis that the squared sum of residuals in a model is significantly different than the SSR of a constant-valued model. Every regression model (both multiple and polynomial) presented, in accordance with the associated test statistic, statistically significant  $R^2$  values ( $p$ -value  $\leq 0.05$ ). Given the high correlation values between several variables (as foetal development acts positively on all measurements and weights), and to filter possible GA estimation candidates, variable selection algorithms were used to produce multiple linear regressions. For a significance level  $\alpha = 0.05$ , the least possible number of features presenting significantly distinguishable effects were selected by each of the 2 algorithms (Stepwise and Forward). Because these algorithms are based on variable iteration, autocorrelation is factor to be taken into consideration. Durbin-Watson (DW) test values (where the null hypothesis assumes that model errors are serially uncorrelated against the alternative that they follow a first order autoregressive process) were inspected for model validation. With an optimal value of 2, both output models presented reliable Durbin-Watson values ( $2 - \text{abs}(\text{DW}) \leq 0.042$ ). In SPSS, the Stepwise algorithm incorporates both F-in and F-out (F-statistic critical values for considering variables as having significantly distinguishable effects or not) parameters used in Backward (starting from a full set of variables and iteratively removing each variable) and Forward (beginning from a single variable and iteratively inserting each variable) variable selection methods, respectively. Hence, Stepwise and Backward algorithms produced multiple linear regression models with identical retained variables and their associated  $\beta$ -weights, and coefficient of determination values ( $R^2 \approx 0.953$ ), varying only in DW values. Because Backward/Stepwise and Forward approaches have different computational starting points and direction (empty versus full set of variables), variables CHL and CRL (which have statistically indistinguishable effects) were retained, respectively. This is understandable when taking into consideration the correlation value between the 2 variables (0.992) and the nature of the measured parameters.

When comparing both models in terms of  $R^2$ , there was no statistical evidence supporting the hypothesis that the squared sum of residuals of models varied significantly. Excluding that variable pair, both models retained the same remaining significant variables: Body, FL, CHL/CRL, REL, Lungs, and Adrenals (in descending standardized  $\beta$ -weight order for both models, albeit having different  $\beta$  values for the same variable). Standardized  $\beta$ -weight values indicate which previously validated variables (presenting statistical significance) contribute the most within a multiple linear regression model. Lungs and Adrenals, although selected by each algorithm, have small contribution values ( $\text{abs}(\beta) \leq 0.1$ ), for example. A ranking system based on weights can be interpreted, denoting body weight ( $\beta \approx 0.393$ ), FL ( $\beta \approx 0.347$ ), and CHL ( $\beta \approx 0.266$ ), CRL ( $\beta \approx 0.199$ ), and REL ( $\beta \approx 0.16$ ) as major contributors for GA estimation, following a linear combination approach. This mustn't mean, however, that a certain variable is better than another variable, individually, at GA estimation. To compare variables individually, single variable-GA pairs are used to compute polynomial regression models.

$R^2$  values increased, for each of the 23 variable-derived models, along all ordered  $k^{\text{th}}$  degree polynomial regression models, for  $k \in \{1, 2, 3, 4, 5\}$ . Polynomial models with higher values of  $k$  are more likely to be subjected to overfitting; should  $k$  tend to an infinitely large value, then the training error would approach 0 ( $R^2$  would approach 1). Due to the nature of our dataset, cross validation (which would account for cases of overfitting) could not be executed, as the division of data would produce datasets with missing representative GA values. For this reason, variable models were only compared to other variable models for the same polynomial degree  $k$ . The concept of overfitting was also taken into consideration while assessing our results. Inspecting the results of residual comparison testing for a significance level  $\alpha = 0.05$ , several clusters and meta-clusters (groups of clusters branching outward in figures 4.2 through 4.6) are distinguishable. Cluster hierarchy for models where  $k = 1$ , simple linear regression, presented a significantly different variable order when compared to all other  $k^{\text{th}}$  degree models. The cluster of variables with the highest coefficient of determination values – CRL, CHL, and FL with  $0.927 \leq R^2 \leq 0.931$  – exhibits significantly fewer error comparatively to the meta-cluster comprised of variables OCD, AC, MFL, Body, LEL, REL, and HC, for example. This top tier cluster does not, however, discern which of the 3 foetal parameters is statistically superior ( $p$ -value  $\geq 0.05$ ) to serve as the best possible GA estimator. Clusters, thus, indicate the hierarchy by which foetal parameters are selected as developmental predictors. For  $k \geq 2$ , body weight is incorporated into the cluster or meta-cluster of variables with highest  $R^2$ .

Other significant changes between  $k$  equal to 1 and  $k \geq 2$  can be witnessed with variables models based on HL and CC. When  $k = 2$ , both models are placed within the meta-cluster of 2<sup>nd</sup> hierarchical position. The latter model, for any  $k \geq 2$ , always stands within the meta-cluster of 2<sup>nd</sup> hierarchical position; however, the model derived from HL measurements, throughout  $3 \leq k \leq 5$ , is positioned within the meta-cluster (for  $k = 3$ , and  $k = 4$ ) or cluster ( $k = 5$ ) of variables with higher estimation capabilities. Absolute  $R^2$  values for every variable regression model, excluding Body, HL, CC, and ID kept their relative position across different polynomial degrees. The regression model based on ID, despite rising in hierarchy throughout ordered values of  $k$ , was never witnessed within any cluster or meta-cluster of rank 2 or superior. The changes in hierarchy configuration are understandable when comparing linear ( $k = 1$ ) to non-linear ( $k \geq 2$ ) regression models. Body weight, as a predictive variable for GA estimation, fits a quadratic function better than a linear one, for example. Such is the nature of that variable, and overfitting can be excluded. However, the same might not be said regarding HL and its hierarchical position variations across  $k$ . It is only when  $k = 5$  that this model is present within the unique cluster of most appropriate GA estimator variables (highest  $R^2$  values). The process of overfitting by  $k$  increment may be at play, for this variable.

### **5.1 Final Remarks**

In our case of 450 foetal autopsy cases, findings suggest that across all variables, CHL, CRL, and FL are the most appropriate candidate foetal parameters for GA estimation. Within all approaches (PCA and regression techniques), certain specific variables showed a tendency to present values indicative of superior estimation capabilities (either by correlation or by SSR, for instance). CRL, CHL, and FL are the only variables possessing this property. Other variables can also be considered as proper developmental markers, depending on the technique utilized. For any degree of polynomial regression, these variables were always displayed within the significantly highest  $R^2$  cluster. The same variables were also selected by multiple linear regression, exhibiting positive standardized  $\beta$ -weights  $\geq 0.199$  (ascendingly ordered CRL, CHL, and FL), and presented the highest PCA communality and loading values. Body weight, HC, HL, and ear length are also noteworthy candidate variables for either presenting high PCA communality and loading values, or having significantly meaningful  $\beta$  and/or  $R^2$  values.

## Chapter 6

### Conclusion

Accurately estimating foetal GA is essential for pregnancy management. As a further matter, GA estimation during autopsy procedures is key in assessing legal and criminal abortion cases. During these events, the estimation of GA depends on the foetal parameters used. Measurements of various foetal anthropometric features are frequently used for this purpose.

The primary goal for this thesis of devising a simple method for storing and manipulating information (regarding the foetal autopsy report files pertaining to Hospital de Egas Moniz) was achieved. This was made possible by developing a Python application which enabled the creation of an information system, integrating a computer-assisted data insertion tool (createDB.py and insertValues.py files, respectively). Moreover, by applying different algorithmic approaches to our collection of structured data (such as the previously discussed PCA and polynomial/multiple regression techniques), we produced statistically meaningful results directly enabling a better understanding of the real-world problem of GA assessment or estimation. We also established a novel approach to determine measurement adequacy, through the course of this work, by associating our computed regression models to statistical hypothesis test for divergence in variance ( $F$ -statistic on squared sum of residuals, conclusively). This new solid and well-founded variable clustering approach is one of our many contributions formulated during this thesis, which we hope will assist foetopathologists everywhere during their medical procedures.

Consistent with previously published work, CHL, CRL, and FL are found to be the most reliable sources of information for estimating foetal developmental age. Particularly in cases of 1<sup>st</sup> degree polynomial regression models, clustering algorithms based on  $R^2$  values placed exactly those 3 variables in the top tier cluster of best GA estimators. These same variables were also witnessed within the cluster of best development predictors for any other  $k^{\text{th}}$  degree polynomial regression model, albeit being accompanied by other variables (such as body weight for  $k \geq 2$  and HL for  $k \geq 3$ ); CHL, CRL, and FL were also retained in multiple linear regression models (with high  $\beta$ -weight values, second only to Body and closely followed by REL), derived from variable selection algorithms based on the statistical distinguishability of variable effects. In cases where these 3 preferable measurements are impossible to obtain, other foetal features can be utilized (albeit less reliable, as our findings suggest) such as HL, HC, body weight, and ear length.

### 6.1 Future Work

After having validated the usability and adequacy of our methodology, it is feasible to assume that progressive endeavours related to our data and methods can ensue; specifically, in the field of biomedical and health sciences. By making use of open linked data – a previously validated method<sup>68</sup> of publishing structured data so that it can be interlinked and become more useful through semantic queries –, it is possible to cross validate, counter-examine, and derive additional knowledge (to name a few practical applications) from our own findings deduced from this work. In this manner, it is possible to provide continuity to our studies not only in temporal terms but also in knowledge-gathering and, consequently, wisdom acquirement.

As our database evolves, and different foetal parameters are recorded, different studies can emerge. By analysing features such as cause of death and family background, in association with measurements and weights, machine learning algorithms (such as neural networks, for instance) can be executed to create a pathological prediction tool. Having a chronological set of the same parameters along a pregnancy events may also help determine certain developmental particularities associated to illness and pregnancy abnormalities. These approaches would be useful for early diagnosis of disease, aiding professionals and family members in taking the appropriate set of actions accordingly.

## References

1. Hern WM. Correlation of fetal age and measurements between 10 and 26 weeks of gestation. *Obstet Gynecol.* 1984, 63 (1): 26 – 32.
2. Gandhi D, Masand R, Purohit A. A simple method for assessment of gestational age in neonates using head circumference. *Pediatrics.* 2014, 3 (5): 211 – 213.
3. Kumar GP, Kumar UK. Estimation of gestational age from hand and foot length. *Med Sci Law.* 1994, 34 (1): 48 – 50.
4. Mercer BM, Sklar S, Shariatmadar A, Gillieson MS, D'Alton ME. Fetal foot length as a predictor of gestational age. *Am J Obstet Gynecol.* 1987, 156 (2): 350 – 355.
5. Patil SS, Wasnik RN, Deokar RB. Estimation of gestational age using crown heel length and crown rump length in India. *International J. of Healthcare & Biomedical Research.* 2013, 2 (1): 12 – 20.
6. Selbing A, Fjällbrant B. Accuracy of conceptual age estimation from fetal crown-rump length. *J Clin Ultrasound.* 1984, 12 (6): 343 – 346.
7. Scheuer JL, MacLaughlin-Black S. Age estimation from the pars basilaris of the fetal juvenile occipital bone. *Int J Osteoarchaeol.* 1994, 4 (4): 377 – 380.
8. Scheuer JL, Musgrave JH, Evans SP. The estimation of late fetal and perinatal age from limb bone length by linear and logarithmic regression. 1980, 7 (3): 257 – 265.
9. Chikkannaiah P, Gosavi M. Accuracy of fetal measurements in estimation of gestational age. In *J Pathol Oncol.* 2016, 3 (1): 11 – 13.
10. Gupta DP, Saxena DK, Gupta HP, Zeeshan Zaidi, Gupta RP. Fetal femur length in assessment of gestational age in thirds trimester in women of northern India (Lucknow, UP) and a comparative study with Western and other Asian countries. In *J Clin Prac.* 2013, 24 (4): 372 – 375.
11. Archie JG, Collins JS, Lebel RR. Quantitative standards for fetal and neonatal autopsy. *Am J Clin Pathol.* 2006, 126 (2): 256 – 265.
12. Sherwood RJ, Meindl RS, Robinson HB, May RL. Fetal age: methods of estimation and effects of pathology. *Am J Phys Anthropol.* 2000, 113 (3): 305 – 315.
13. Andrews DT, Chen L, Wentzell PD, Hamilton DC. Comments on the relationship between principal components analysis and weighted linear regression for bivariate data sets. *Chemometrics and Intelligent Laboratory Systems.* 1996, 34 (2): 231 – 244.
14. Nadaraya EA. On estimating regression. *Theory of Probability & Its Applications.* 1964, 9 (1): 141 – 142.
15. R Core Team. R: a language and environment for statistical computing, version 3.3.2. Vienna, Austria: R Foundation for Statistical Computing. 2016.
16. Eaton JW, Bateman D, Hauberg S. GNU Octave version 3.0.1 manual: a high-level interactive language for numerical computations. CreateSpace Independent Publishing Platform. 2009.
17. Barata AP. Anthropometric data analytics: a portuguese case study. 2017. <https://github.com/BarataAP/Anthropometric-Data-Analytics-Portugal.git/>.
18. Barata AP, Couto FM, Carvalho LC. Anthropometric data analytics: a portuguese case study. 11<sup>th</sup> International Conference on Practical Applications of Computational Biology & Informatics. 2017. <http://www.pacbb.net/>.
19. Wigglesworth JS, Singer DB. Textbook of fetal and perinatal pathology. Blackwell Scientific Publications. 1991.
20. Edmonds DK, Lindsay KS, Miller JF, Williamson E, Wood PJ. Early embryonic mortality in women. *Fertil Steril.* 1982, 38 (4): 447 – 453.



21. Opitz JM. The Farber lecture. Prenatal and perinatal death: the future of developmental pathology. *Pediatr Pathol*. 1987, 7 (4): 363 – 394.
22. Stein Z. Early fetal loss. *Birth Defects Orig Artic Ser*. 1981, 17 (1): 95 – 111.
23. Warburton D, Fraser FC. Spontaneous abortion risk in man: data from reproductive histories collected in a medical genetics unit. *Am J Hum Genet*. 1964, 16: 1 – 25.
24. Wilcox AJ, Weinberg CR, Wehmann RE, Armstrong EG, Canfield RE, Nisula BC. Measuring early pregnancy loss: laboratory and field methods. *Fertil Steril*. 1985, 44 (3): 366 – 374.
25. Bauld R, Sutherland GR, Bain AD. Chromosomal studies in investigations of stillbirths and neonatal deaths. *Arch Dis Child*. 1974, 49 (10): 782 – 788.
26. Boué A, Boué J, Gropp A. Cytogenetics of pregnancy wastage. *Adv Hum Genet*. 1985, 14: 1 – 57.
27. Boué J, Boué A, Lazar P. Retrospective and prospective epidemiological studies of 1500 karyotyped spontaneous human abortions. 1975. *Birth Defects Res A Clin Mol Teratol*. 2013, 97 (7): 471 – 486.
28. Gilbert EF, Opitz JM. Developmental and other pathologic changes in syndromes caused by chromosome abnormalities. *Perspect Pediatr Pathol*. 1982, 7: 1 – 63.
29. Moore KL, Persaud TVN, Torchia MG. *The developing human: clinically oriented embryology*. Elsevier Health Sciences. 2015.
30. Breborowicz GH. Limits of fetal viability and its enhancement. *Early Pregnancy*. 2001, 5 (1): 49 – 50.
31. Tyson JE, Parikh NA, Langer J, Green C, Higgins RD. Intensive care for extreme prematurity – moving beyond gestational age. *N Engl J Med*. 2008, 358 (16): 1672 – 1681.
32. Luke B, Brown MB. The changing risk of infant mortality by gestation, plurality, and race: 1989-1991 versus 1990-2001. *Pediatrics*. 2006, 118 (6): 2488 – 2497.
33. American College of Obstetricians and Gynecologists. ACOG practice bulletin: clinical management guidelines for obstetrician-gynecologists: number 38, September 2002. Perinatal care at the threshold viability. *Obstet Gynecol*. 2002, 100 (3): 617 – 624.
34. Walsh F. Prem baby survival rates revealed. *BBC News*. 11 April 2008.
35. Kaempf JW, Tomlinson M, Arduza C, Anderson S, Campbell B, Ferguson LA, Zabari M, Stewart VT. Medical staff guidelines for periviability pregnancy counseling and medical treatment of extremely premature infants. *Pediatrics*. 2006, 117 (1): 22 – 29.
36. Morgan MA, Goldenberg RL, Schulkin J. Obstetrician-gynecologists' practices regarding preterm birth at the limit of viability. *J Matern Fetal Neonatal Med*. 2008, 21 (2): 115 – 121.
37. Incerpi MH, Miller DA, Samadi R, Settlege RH, Goodwin TM. Stillbirth evaluation: what tests are needed? *Am J Obstet Gynecol*. 1998, 178 (6): 1121 – 1125.
38. Pine VR. *Unrecognized and unsanctioned grief: the nature and counseling of unacknowledged loss*. Charles C Thomas Pub Limited. 1990.
39. Cussen L, Scurry J, Mitropoulos G, McTigue C, Gross J. Mean organ weights of an Australian population of fetuses and infants. *J Paediatr Child Health*. 1990, 26 (2): 101 – 103.
40. Chambers HM, Knowles S, Staples A, Tamblyn M, Hann EA. Anthropometric measurements in the second trimester fetus. *Early Hum Dev*. 1993, 33 (1): 45 – 59.
41. Larroche JC. *Developmental pathology of the neonate*. Excerpta Medica. 1977.
42. Guihard-Costa AM, Menez F, Delezoide AL. Standards for dysmorphological diagnosis in human fetuses. *Pediatr Dev Pathol*. 2003, 6 (5): 427 – 434.
43. Kulkarni ML, Rajendran NK, Sangam DK. Inner canthal, outer canthal and inter pupillary distance in newborns. *Indian Pediatr*. 1992, 29 (6): 759 – 763.
44. Kulkarni ML, Rajendran NK. Internipple distance in the newborns. *Indian Pediatr*. 1992, 29 (5): 619 – 620.

45. Madhulika, Kabra SK, Barar V, Purohit A, Saxena S, Sharma U, Bansal RK. Upper and lower limb standards in newborn. *Indian Pediatr.* 1989, 26 (7): 667 – 670.
46. Merlob P, Sivan Y, Reisner SH. Lower limb standards in newborns. *Am J Dis Child.* 1984, 138 (2): 140 – 142.
47. Gruenwald P, Minh HN. Evaluation of body and organ weights in perinatal pathology. I. Normal standards derived from autopsies. *Am J Clin Pathol.* 1969, 34: 247 – 253.
48. Hansen K, Sung CJ, Huang C, Pinar H, Singer DB, Oyer CE. Reference values for second trimester fetal and neonatal organ weights and measurements. *Pediatr Dev Pathol.* 2003, 6 (2): 160 – 167.
49. FitzSimmons J, Chinn A, Shepard TH. Normal length of the human fetal Gastrointestinal tract. *Pediatr Pathol.* 1988, 8 (6): 633 – 641.
50. Gruenwald P, Minh HN. Evaluation of body and organ weights in perinatal pathology. II. Weight of body and placenta of surviving and of autopsied infants. *Am J Obstet Gynecol.* 1961, 82: 312 – 319.
51. Munsick RA. Similarities of negro and caucasian fetal extremity lengths in the interval from 9 to 20 weeks of pregnancy. *Am J Obstet Gynecol.* 1987, 156 (1): 183 – 185.
52. Potter EL, Craig JM. *Pathology of the fetus and the infant.* Year Book. 1975.
53. Bhat GJ, Mukelabai K, Shastri GN, Tamina C. Anthropometric parameters of zambian infants at birth. *J Trop Pediatr.* 1989, 35 (3): 100 – 104.
54. Oliphant TE. *Python for scientific computing.* Computing in Science & Engineering. 2007, 9 (3): 10 – 20.
55. Millman KJ, Aivazis M. *Python for scientists and engineers.* Computing in Science & Engineering. 2011, 13 (2): 9 – 12.
56. Walt S, Colbert SC, Varoquaux G. The NumPy array: a structure for efficient numerical computation. *Computing in Science & Engineering.* 2011, 13 (2): 22 – 30.
57. Abadia M, Cardelli L. *A theory of objects.* Springer Science & Business Media. 1998.
58. Dahl RO, Dijkstra E, Hoare CAR. *Structured programming.* CreateSpace Independent Publishing Platform. 2012.
59. Pratt TW. *Programming languages: design and implementation.* Prentice-Hall. 1975.
60. Turner DA. Total functional programming. *J Univers Comput Sci.* 2004, 10 (7): 751 – 768.
61. Steimann F. The paradoxical success of aspect-oriented programming. *Proceedings of the 21st Annual ACM SIGPLAN Conference on Object-Oriented Programming Systems, Languages, and Applications.* 2006, 41 (10): 481 – 497.
62. Mithcell R, McKim J. *Design by contract, by example.* Addison Wesley. 2002.
63. Baral C, Gelfond M. *Logic programming and knowledge representation.* The Journal of Logic Programming. 1994, 19 – 20: 73 – 148.
64. IBM Corp. *IBM SPSS Statistics for Windows, version 24.0.* Armonk, NY: IBM Corp. 2016.
65. Judd CM, McClelland GH, Ryan CS. *Data analysis: a model comparison approach.* Routledge. 2011.
66. Jambu M. *Exploratory and multivariate data analysis.* Elsevier. 1991
67. Tabachnick BG, Fidell LS. *Using multivariate statistics.* Allyn & Bacon. 2001.
68. Barros M, Couto FM. *Knowledge representation and Management: a linked data perspective.* IMIA Yearbook. 2016, 178 – 183.