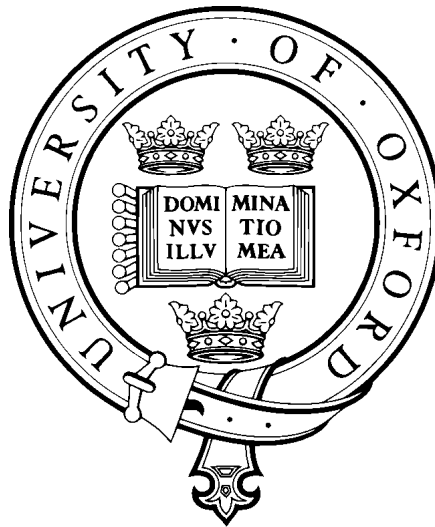


Physiological Patterns in Pregnancy



Lise Lørup
New College
University of Oxford

Supervised by
Professor Lionel Tarassenko
Trinity Term 2016

This thesis is submitted to the Department of Engineering Science,
University of Oxford, in fulfilment of the requirements for the degree of
Doctor of Philosophy

Declaration

I declare that this thesis is entirely my own work, and except where otherwise stated,
describes my own research.

This DPhil was supported by the Clarendon Fund, Scatcherd European Scholarship and
RCUK Digital Economy Programme grant number EP/G036861/1 (Oxford Centre for
Doctoral Training in Healthcare Innovation).

Lise Lørup
New College
University of Oxford

DPhil in Engineering Science
Trinity Term 2016

Abstract

Physiological Patterns in Pregnancy

Lise Lørup

New College

Thesis submitted for the degree of Doctor of Philosophy

University of Oxford, Trinity Term 2016

The Confidential Enquiries into Maternal Deaths in the UK have highlighted an urgent need for development of a national Modified Early Obstetric Warning System (MEOWS) to aid the more timely recognition and treatment of complications in pregnancy. Six vital signs are used to assess the clinical status of women presenting acutely during and immediately after pregnancy: heart rate, oxygen saturation, systolic and diastolic blood pressure, temperature and respiratory rate. In the case of MEOWS, these define the thresholds that determine if a woman requires further medical review. Dynamic changes in maternal vital sign physiology associated with pregnancy complicate the use of vital signs to detect physiological deterioration. These changes are currently poorly described in the literature.

The aim of this work is to establish the normal physiological response to pregnancy in low-risk pregnant women, in terms of expected changes in vital signs. Using this knowledge, we will propose the first evidence-based, gestation-specific MEOWS chart.

To achieve this, we have set up the Prediction of Physiological Patterns in Pregnancy (4P) study. Using a custom-designed digital health system, healthcare professionals and women self-monitoring at home collect large numbers of vital sign measurements electronically. In particular, as respiratory rate currently is the only vital sign measured manually in routine clinical care, we propose two methods for electronic measurement of respiratory rate: using the in-built accelerometer of a smartphone and using the photoplethysmogram signal output by a pulse oximeter. These methods are validated against the current reference standard derived from a 15-second nurse count of breaths.

This thesis uses a dataset of 2,645 sets of vital sign measurements, collected from 439 women during pregnancy, in the 4P study. Using the electronic methods for measurement of respiratory rate, we obtain respiratory rate estimates with a mean absolute error of 2.9 respirations per minute compared to the reference measurement. From these data, we design reference ranges for changes in all six vital signs conditional on gestational age. Using a centile-based approach, we propose gestation-specific thresholds of normality for each vital sign, which will form the basis of a new MEOWS score to alert for physiological deterioration. Defining normality is the first step in recognising abnormality.

Acknowledgements

First and foremost, I would like to thank my supervisor, Professor Lionel Tarassenko, for his invaluable support, guidance and encouragement throughout this research. I am also grateful to my clinical collaborators, Dr Lucy Mackillop and Dr Peter Watkinson, who have made this thesis possible. Thank you for managing the 4P project, for sharing your medical insights and for all of your support.

The 4P project was supported by an incredible team of research midwives. In particular I would like to thank Fenella Roseman, Jude Kemp, Fiona Kumar and Clare Edwards who were instrumental in the collection of data for the Oxford-arm of the study. I would also like to extend thanks to my engineering colleagues who contributed to all aspects of setting up and running the 4P project, including Rebecca Pullon, Dr Dario Salvi, Dr Andreas Triantafyllidis and Dr Oliver Gibson.

Thanks are due to the INTERBIO-21st project team, in particular to Professor Stephen Kennedy, who permitted access to the INTERBIO-21st patients and data, and to Stephen Ash and Rachel Craik, who assisted with the preparation of these datasets.

For the systematic review in Chapter 2, I would like to recognise the contribution of several people. Dr Susannah Fleming, Department of Primary Health Sciences, provided invaluable advice on every aspect of the systematic review methodology. Tatiana Petrinic, Oxford University Healthcare Libraries, assisted with developing the search strategy and carried out the literature search. Jacqueline Birks and Stephen Gerry, Centre of Statistics in Medicine, advised on the statistical methods for analysis of the systematic review data, as well as for analysis of the 4P data in Chapters 3 and 5. I would also like to thank Dr Eric Ohuma for being a great discussion partner and advisor in developing the statistical methods in Chapter 5.

I am grateful for the friendship and assistance of all my colleagues in the Biomedical Signal Processing and m-health group and in Time Series Terrace. Thank you for making the lab such an inspiring and enjoyable environment to be part of. I would also like to extend my gratitude to all the CDT staff and students with whom I shared the DPhil journey, and particularly to Jo Armitage, who kept us all in line and motivated.

Being part of the wonderful institution that is the University of Oxford has for me been about a lot more than purely the academic aspects. I would especially like to thank all of my friends in the Oxford University Golf Club and in New College Boat Club for the wonderful memories we have created outside of the lab during the past four years.

Finally, I would like to thank Milos and my family for their endless love and support.

Contents

1	Background	1
1.1	Maternal health and care in the UK	1
1.2	An introduction to early warning scores	7
1.3	Vital signs and their measurement	11
1.4	The need for an electronic measurement of respiratory rate	18
1.5	Aims and overview of this thesis	19
2	Systematic review of physiological trends in pregnancy	21
2.1	Systematic review methods	22
2.2	Methods for construction of centile charts	29
2.3	Results	35
2.4	Discussion	44
3	The 4P study and collection of vital sign data	48
3.1	Study protocol	49
3.2	System for collection of vital sign data	51
3.3	Description of the data used for this thesis	58
3.4	Validation of vital sign measurements	62
3.5	Discussion and future improvements	74
4	Electronic measurement of respiratory rate	76
4.1	Current measurement of respiratory rate	77
4.2	Review of literature	80
4.3	Methods for collection and analysis of respiratory waveforms	87
4.4	Respiratory rate estimates from the 4P data	104
4.5	Improving estimation accuracy using a signal quality index	118
4.6	Computation of a single electronic estimate	122
4.7	Limitations and future improvements of respiratory estimation methods .	124

5	Design of vital sign reference ranges	126
5.1	The use of reference ranges in obstetric care	126
5.2	Method identification	127
5.3	Application of centile estimation models to 4P data	134
5.4	Final centiles for all vital signs	147
5.5	Modelling individual patterns of change	151
5.6	Discussion	159
6	Design of a Modified Obstetric Early Warning Score	161
6.1	Design and evaluation of early warning systems	161
6.2	MEOWS reported in the literature	165
6.3	Centile-based early warning scores	174
6.4	Directions for future validation of C-MEOWS	185
7	Conclusions and future work	187
7.1	Contributions of this thesis	187
7.2	Directions for future work	191
A	Supporting material for systematic review	195
A.1	Addition to methods	195
A.2	Addition to results	199
B	Mathematical methods	202
B.1	Classical kernel regression	202
B.2	Weighted variable bandwidth kernel regression	204
B.3	General multilevel model	205
B.4	Multilevel model for aggregate data	215
B.5	Gaussian processes	217
C	Supporting material: Vital sign measurement	219
C.1	Measurement of blood pressure	219
C.2	Measurement of heart rate and oxygen saturation	220
D	Supporting material: Centile estimation	224
D.1	Centile estimation	224
D.2	Equations for final centiles	226
	Bibliography	228

Glossary

Terms in common use

AIC; Akaike's Information Criterion

AM; amplitude modulation

AR; auto-regressive (typically AR model)

AUROC; Area under the Receiver-Operator-Characteristic curve

BIC; Bayesian Information Criterion

BMI; body mass index

BP; blood pressure

CEWS; Centile-based Early Warning Score

CO; cardiac output

DBP; diastolic blood pressure

ECG; electrocardiogram

EWS; Early Warning Score

FFT; Fast Fourier Transform

FIR; finite impulse response

FM; frequency modulation

GP; General Practitioner

HDU; high-dependency unit

HF; high frequency

HR; heart rate

HRV; heart rate variability

ICNARC; Intensive Care National Audit and Research Centre

ICU; intensive care unit

IIR; infinite impulse response

LF; low frequency

MAE; mean absolute error

MEOWS (or MOEWS); Modified Obstetric Warning Score

MGRS; Multicentre Growth Reference Study group

ML; maximum likelihood

NEWS; National Early Warning Score

NHS; National Health Service

NICE; National Institute for Health and Care Excellence
 NPV; negative predictive value
 OUH; Oxford University Hospitals
 PCA; principal component analysis
 PPG; photoplethysmogram
 PPV; positive predictive value
 PSD; power spectral density
 Q-Q; quantile-quantile (typically Q-Q plot)
 RIAV; respiratory induced amplitude variation
 RIFV; respiratory induced frequency variation
 RIIV; respiratory induced intensity variation
 RIP; respiratory inductance plethysmography
 ROC; Receiver-Operator-Characteristic (typically ROC curve)
 RR; respiratory (respiration) rate
 RSA; respiratory sinus arrhythmia
 SBP; systolic blood pressure
 SD; standard deviation
 SpO₂; oxygen saturation
 SQI; signal quality index
 Temp; body temperature
 VLF; very low frequency
 WHO; World Health Organization

Terms defined by the author

A_x, A_y, A_z, A_P ; accelerometer signals
 G_x, G_y, G_z, G_P ; gyroscope signals
 4P; Prediction of Physiological Patterns in Pregnancy
 CEMD; Confidential Enquiry into Maternal Deaths in the UK
 C-MEOWS; Centile-based Modified Obstetric Warning Score
 HDP; hypertensive disorders of pregnancy
 IG; INTERGROWTH–21st project
 KR; kernel regression
 MLAD; multilevel model for aggregate data
 WVBKR; weighted variable bandwidth kernel regression

Chapter 1

Background

In 2007, the Confidential Enquiry into Maternal Deaths in the UK stated that *"There is an urgent need for the routine use of a national obstetric early warning chart, similar to those in use in other areas of clinical practice, which can be used for all obstetric women to help in the more timely recognition, treatment and referral of women who have, or are developing, a critical illness"* [1]. This chapter aims to introduce the reader to the motivation for this recommendation and the concept of an early warning score. We will also discuss the need for development of accurate reference ranges, or normograms, for normal values of vital signs during pregnancy, as this is an essential pre-requisite for the development of an early warning score. The chapter concludes with an overview of the aims and contents of this thesis.

1.1 Maternal health and care in the UK

Every year, approximately 650,000 deliveries take place in NHS hospitals across England. If we include episodes of miscarriage or ectopic pregnancy¹, the overall number of hospital admissions related to pregnancy is close to 700,000 admissions². This makes having a baby the single largest reason for admission to hospital [2].

¹An ectopic is a pregnancy in which the fetus is not viable because it develops outside the womb

²For the financial year 2013-14, the exact numbers were 646,904 deliveries, 38,869 miscarriages and 10,436 ectopic pregnancies

Pregnant women receive care from a range of health care professionals at multiple time points during the pregnancy and after the baby is born. The periods before and after childbirth are called the antenatal (or antepartum) and postnatal (or postpartum) periods, respectively. The period from the onset of labour to the end of labour is called the intrapartum period. Figure 1.1 outlines the maternity care pathway for women with low-risk pregnancies. Whereas women that are deemed to have a low risk of developing pregnancy complications are typically cared for by their GP and a midwifery service in the community, women with high-risk pregnancies should receive more frequent care from a multidisciplinary team of healthcare professionals [3]. In addition to a midwife, such teams may include an obstetrician, obstetric anaesthetist, obstetric or specialty physician, surgeon and other members of the allied health professions (e.g. physiotherapists and nutritionists) as appropriate [4].

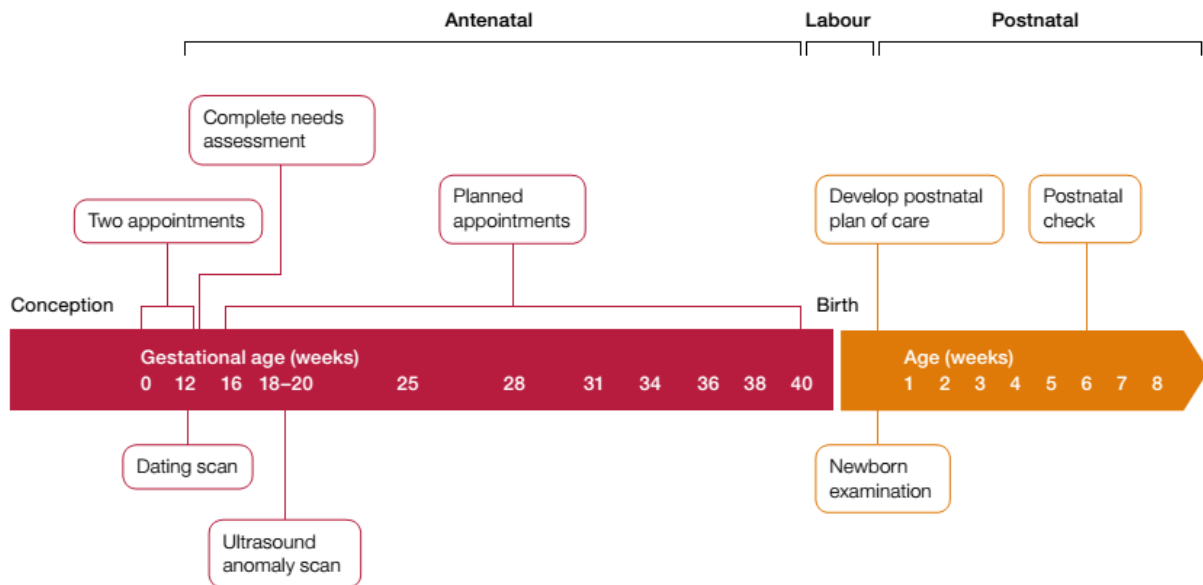


Figure 1.1: Simplified summary of the maternity care pathway for NHS England (only key indicative elements included). Reproduced from [3].

1.1.1 The increasing demand and cost of maternity care

The number of maternities is increasing; 110,000 more women gave birth in 2012 than in 2003. At the same time, a greater proportion of women experienced complex pregnancies that demand higher care needs [4]. This is predominantly a result of demographic

trends [5], that include older maternal age [2], higher rates of obesity [6] and a shift towards a more ethnically diverse maternity population. Moreover, as lifestyle and chronic diseases, such as diabetes and cardiac disease, are associated with a greater risk of maternal mortality and morbidity, the increasing prevalence of such diseases translates to more women being at risk of developing pregnancy complications [7].

The challenge of providing care to a more numerous and complex population of pregnant women puts a greater pressure on the capacity and financial resources of NHS maternity services. Currently, maternity care costs the NHS around £2.6 billion per year, or between £2,600 and £3,700 per pregnancy. This cost represents around 2.8 % of England's total health spending (estimates from the 2012-13 financial year) [3, 8].

Maternity services are unique in that the demand cannot be managed and activity (and hence cost) cannot be controlled through referrals [8]. At the same time, quality and safety considerations are of utmost importance. The care that pregnant women receive during and immediately after pregnancy may have a great impact on long-term health and wellbeing, both of the women and their babies, and on their future use of health services. Thus the need for constant improvement of maternity services continues to be a main point of emphasis for service providers, policy-makers and patients alike [4].

1.1.2 Prevalence and impact of pregnancy complications

Maternal deaths in the UK are, fortunately, rare and the incidence of maternal mortality is decreasing. The Confidential Enquiries into Maternal Deaths in the UK (CEMD) is the world's longest running audit programme for investigating the incidence and causes for maternal mortality. The most recent CEMD report stated that between the periods 2006 to 2008 and 2010 to 2012, the maternal mortality rate in the UK decreased from 11 to 10 maternal mortalities per 100,000 maternities³. This is equivalent to 357 maternal deaths that occurred either during or within 42 days of the end of their pregnancy, from 2009 to 2012 [4].

³The maternal mortality rate (MMRate) is defined as the number of maternal mortalities per 100,000 births (irrespective of neonatal outcome) at or beyond 24 weeks of gestation [4].

In comparison to the rest of the world, the number of maternal deaths in the UK is low. According to numbers from the Global Burden of Disease Study from 2013 [9], the maternal mortality ratio⁴ varies substantially by country; from 956.8 (95 % CI 685.1 to 1262.8) in South Sudan to 2.4 (1.6 to 3.6) in Iceland. Using an equivalent estimation method, the maternal mortality ratio for the UK was estimated to be 6.1 (5.1 to 6.9).

Maternal deaths are typically classified as arising from either direct or indirect causes [4]. Direct causes are complications that occur as a result of the pregnancy, delivery or management of the two. Below we have described the most common direct causes, which include hypertensive disorders of pregnancy (including pre-eclampsia), sepsis, obstetric haemorrhage and thrombosis/thromboembolism. Indirect causes are complications that arise from medical or mental health conditions exacerbated by pregnancy, such as cardiac disease, or causes unrelated to pregnancy, such as coincidental deaths or substandard care. The most common causes of maternal mortality and their incidence are summarised in Figure 1.2.

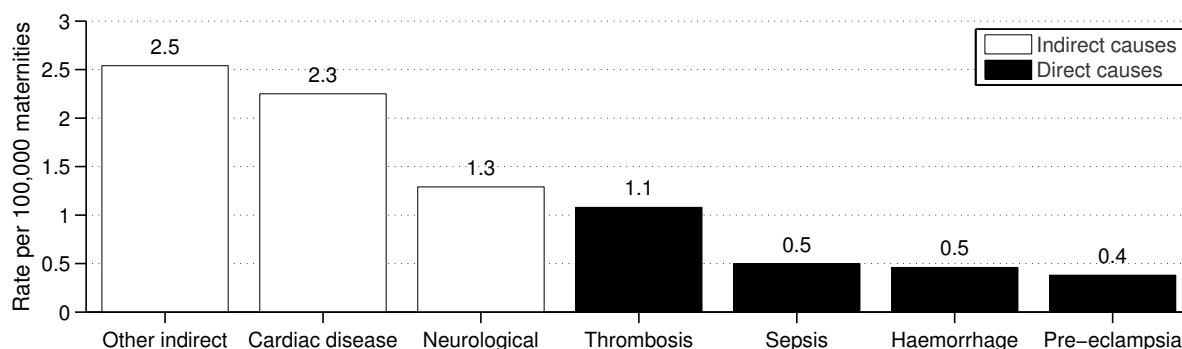


Figure 1.2: Causes for maternal mortality in the UK categorised as indirect (white) and direct (black) causes. Data are reproduced from the 2014 CEMD report [4]

Gestational hypertension

Gestational hypertension is high blood pressure that occurs for the first time during pregnancy. Specifically, the National Institute for Health and Care Excellence (NICE) defines gestational hypertension as *"new-onset hypertension that presents after 20 weeks*

⁴The maternal mortality ratio (MMRatio) is defined as the number of maternal mortalities per 100,000 live births. For the UK, the *MMRatio* is lower than the *MMRate* because only the number of deaths from national registries are counted.

of pregnancy" [10]. It complicates 6 to 7 % of pregnancies [11]. This is in contrast to essential or pre-existing hypertension, which affects 3 to 5 % of pregnancies but is not linked to the pregnant state. Gestational hypertension can occur in isolation, or in association with traces of protein in the urine (proteinuria) when it is known as pre-eclampsia.

Pre-eclampsia

Pre-eclampsia is a multisystem disorder that can affect almost all maternal organ systems, as well as the unborn baby. It is defined by the International Society for the Study of Hypertension in Pregnancy (ISSHP) as two consecutive high blood pressure measurements ($> 140/90$ mmHg) and proteinuria occurring after 20 weeks of gestation [10,12]. In severe cases, pre-eclampsia may lead to eclampsia when the mother develops seizures. Obstetric management of severe pre-eclampsia focuses on medical management of blood pressure (using anti-hypertensive medication) and prevention of seizures using magnesium sulfate; however, the ultimate treatment is delivery of the baby and placenta [13].

Sepsis

Sepsis is a life-threatening condition which in pregnancy typically arises from a genital tract (puerperal) infection. It is generally preceded by an increase in body temperature ($> 38^{\circ}\text{C}$), heart rate (>100 beats per minute) and respiratory rate (>20 respirations per minute) [14,15]. In many cases where sepsis results in morbidity and mortality, care has been found to be substandard [4]. Common themes are failure to recognise the sick woman and to provide timely referral and treatment, typically in the form of antibiotics and fluid management [16].

Haemorrhage

Postpartum haemorrhage, or a blood loss > 500 ml within the first 24 hours of delivery, is by far the most common complication of childbirth [2]. Severe haemorrhage of >2500 ml

occurs in 1 of 300 deliveries [17]. While most cases are successfully treated using either drugs or surgical intervention, severe morbidity and mortality from haemorrhage often result from a failure to detect early signs, such as tachycardia, and to initiate treatment at an early stage [17].

Thrombosis and thromboembolism

Thrombosis (local blood clots) and thromboembolism (obstruction of a blood vessel by a dislodged blood clot) are leading causes of maternal deaths. These conditions are strongly associated with obesity [18] and can occur at any point from early pregnancy to the early postnatal period. In the UK, the overall incidence of thrombosis in pregnancy is estimated to be 1 to 2 per 1,000 maternities. Most cases are treated by administration of a thromboprophylaxis drug [19].

While the maternal mortality rate in the UK is low, it is important to emphasise that for every maternal death, many more women suffer severe morbidities. Recent publications [4,20] quote a study of a cohort of 48,865 pregnant women from 2001, which found that for every woman who died, an estimated 118 women were subject to near-miss events or severe adverse maternal morbidity [21]. As severe morbidity is difficult to define and poorly documented in the literature, the exact incidence rate is difficult to estimate. What is clear is that morbidity from pregnancy complications may severely impact quality of life and can leave the women affected with lifelong disability. Pregnancy complications may also affect the baby and can result in stillbirth, low birth weight or pre-term delivery, all of which increase the risk of the baby developing developmental, behavioural or neurological problems later in life [1].

1.1.3 Improvements to the care of pregnant women

Studies estimate that as many as 30 to 40 % events of severe adverse maternal morbidity are preventable. A 'preventable' event implies that the event could have been either

completely avoided or had a less severe outcome had it been identified earlier [1]. Earlier recognition of abnormality could also prevent the progression from severe morbidity to maternal death, as well as perinatal mortality associated with maternal critical illness [22].

Preventable events may occur both because of system failure and substandard care. The recently published National Maternity Review [23] highlights meaningful differences in the standard of maternity care across the country. In particular, maternal deaths have frequently been associated with a lack of routine observations and a failure to recognise the significance of deteriorating vital signs. Vital signs are clinical measurements, such as heart rate, blood pressure, temperature, respiratory rate and oxygen saturation, which indicate the status of a patient's essential body functions. A more timely recognition, treatment and referral of women who have, or are developing, a critical illness is needed in order to improve the standard of care [1, 4, 14].

Based on this, the 2007 CEMD report recommended the introduction of routine vital sign monitoring for pregnant women with any signs of ill health during and immediately after pregnancy and child birth, coupled with the introduction of a national obstetric early warning score designed specifically for use in the obstetric population. The concept of an early warning score is introduced in the following section.

1.2 An introduction to early warning scores

Early warning scores (EWS), based on vital sign measurements, are a tool commonly used in other areas of critical care medicine for early identification of deteriorating physiology in any in-hospital patient. The concept of the EWS is based on the fact that critical illness is frequently preceded by abnormal vital sign values. By measuring, recording and assigning a score to each vital sign in turn, according to a pre-defined cut-off defining normality, the patient is assigned an overall score that reflects her physiological status. This score is subsequently compared to some pre-determined threshold to determine whether or not the patient should be referred to a higher level of care. For the acutely ill

patient, the vital signs recommended by NICE for scoring are heart rate, respiratory rate, systolic blood pressure, oxygen saturation, temperature and level of consciousness [24].

To refer patients who present with deteriorating physiology to a higher level of care within the hospital, a four-level system (where levels are denoted 0 to 3) is typically used to classify patient-care requirements. These levels are outlined in the list below (adapted from [20, 25]):

- **Level 0:** Normal ward care, including outpatients
- **Level 1:** Acute ward care, including patients at risk of deterioration, needing extra observations and/or recently returned from level 2 care
- **Level 2:** High-dependency unit (HDU), including patients requiring invasive monitoring and/or intervention, single organ failure and/or recently returned from level 3 care
- **Level 3:** Intensive care unit (ICU), including patients requiring advanced respiratory support or multi organ failure

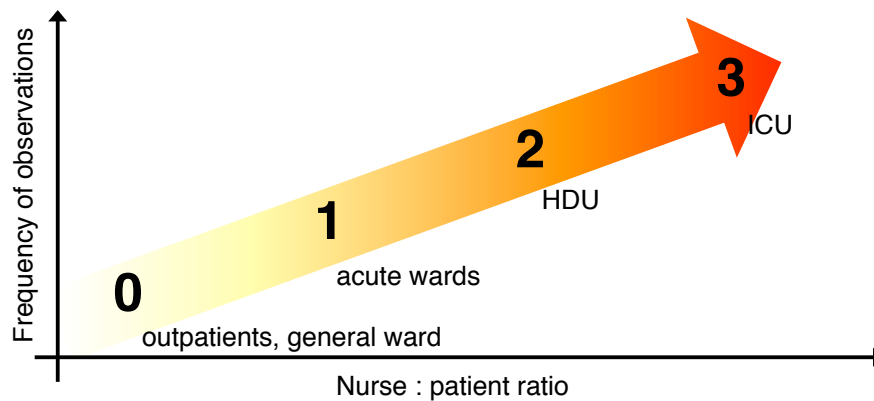


Figure 1.3: The four levels of referral for patient care within the hospital. HDU stands for high-dependency unit; ICU stands for intensive care unit. Adapted from [26]

Figure 1.3 illustrates how the frequency of monitoring and the nurse-to-patient ratio increases for each level of care. A graded response that discriminates between patients according to their abnormality score, and consequently the level of care required, is recommended in order to prioritise limited hospital resources and to ensure that the individual patient's care needs are met.

1.2.1 EWS for the general adult population

In 2012, the first National Early Warning Score (NEWS) was published. NEWS defines how the healthcare system should identify and respond to acute illness in people presenting to, or within hospitals. The NEWS scoring bands are based on data from a large vital sign dataset and were set to maximise the ability to discriminate in-hospital death within 24 hours based on a complete set of vital sign observations [27,28]. However, NEWS is not valid for use in pregnancy because the physiological adjustments of pregnancy affect the response of the pregnant woman to illness and injury [27]. Findings from several studies further support the fact that EWS designed to identify deterioration in the general population are not suitable for use in pregnancy [29,30]. While the 2007 CEMD report recommended the introduction of a national Modified Obstetric Early Warning Score (MEOWS) in the UK [1], no such nationally agreed score has yet been published.

1.2.2 Current use and development of an obstetric EWS

The use of obstetrical charts to record vital signs during pregnancy is not a new concept; as early as in 1970, Seiler proposed the charting of antenatal observations so that *"the condition of the patient would be seen at a glance."* [31]. Following the recommendation from the 2007 CEMD report, a survey of the UK's lead obstetric anaesthetists during 2012 to 2013 found that 100 % of the respondents had implemented a locally agreed version of MEOWS in their unit [32]; this represented an increase from 19 % in a similar survey from 2007 [33]. A high MEOWS uptake rate has also been reported across inpatient midwifery practices, while use in the community settings was reported to be infrequent [34].

In terms of vital signs included in current MEOWS, the surveys found that the vast majority of the scores include all of the vital signs recommended by NICE, as well as diastolic blood pressure. The inclusion of diastolic blood pressure is thought to be important for the diagnosis of conditions such as pre-eclampsia and haemorrhage [1,17]. However, the scores vary widely with respect to chart formats, number of types of

physiological parameters recorded (in addition to vital signs), trigger thresholds and escalation procedures. This is likely to be a consequence of the fact that these scores are not evidence-based but rather agreed by expert opinion and consensus.

To date in the UK, only a single evidence-based MEOWS has been proposed. The score, published in 2013 by Carle *et al.* [30], was designed to predict maternal mortality from physiological data collected during the first 24 hours of ICU-admission of maternity patients in the ICNARC dataset. Using statistically weighted variables with some adjustments to meet the minimum monitoring requirements set out by NICE [24], they designed a score that bears close resemblance to the NEWS score. Although their score accurately predicted mortality, it is not clear whether it would be applicable for identification of deterioration with less severe consequences [22, 30].

Moreover, the efficiency and ability of the MEOWS currently in use to discriminate women at risk of deterioration has not been thoroughly validated [32, 34]. The only prospective validation of the utility of a MEOWS has been described by Singh *et al.* [35]. While the analysis of their score showed that it was effective⁵ in detecting morbidity (based on a broad definition of morbidity presented in the paper), the authors concluded that it produced a high rate of false positive triggers that could be avoided with further refinement of trigger thresholds [22, 35].

In the UK, data collected by the Intensive Care National Audit and Research Centre (ICNARC) showed that in 2007, 260 of 100,000 women in the maternity population required level 3 (ICU) care [36]. This is comparable with the rates of admission quoted by Baskett *et al.* of 0.05% to 1.7% in developed countries during the early 2000's [37]. The number of women who require level 2 care is much less clear – a recent editorial estimated that as many as 5 % of deliveries may require level 2 (HDU) care [25].

⁵The exact metrics used to assess score effectiveness will be introduced later in Chapter 6.

1.2.3 The challenge of designing an obstetric EWS

The maternity population represents an extremely challenging target group for development of screening tools. Firstly, pregnancy is associated with an altered physiology that changes over the course of the pregnancy. Secondly, most pregnant women are of a relatively young age, which leads to an increased physiological reserve before decompensating, a lesser risk of medical comorbidities and lower rates of severe deterioration. Thirdly, the risk of physiological deterioration is to a large extent caused by pregnancy-specific medical disorders that are associated with unique symptoms. All of these factors complicate the ability of screening tools to distinguish accurately between sick and healthy pregnant women [38].

The biggest problem, however, is that there is little knowledge about how vital signs change in response to the physiological adaptations that occur over the course of pregnancy and in the postpartum period [22]. This may explain why the appropriate thresholds for identifying women at risk remain undecided, even when there appears to be a consensus on which variables to include in a MEOWS chart. An evidence-based approach, based on a large number of vital sign measurements, is required to identify when a woman is becoming physiologically abnormal for her stage of pregnancy. The first step to identifying abnormality is to recognise *normal* physiological behaviour. Since pregnancy is associated with physiological adaptations, the knowledge of normal ranges of vital signs for each stage of normal pregnancy is required.

1.3 Vital signs and their measurement

According to the 2014 CEMD report [4], *"All women with any symptoms or signs of ill health, including those who are postnatal, should have a full set of basic observations taken (temperature, pulse rate, respiratory rate and blood pressure), and the results documented and acted upon. Normality cannot be presumed without measurement."*

In the antenatal clinic, some vital signs, such as blood pressure, are typically measured

during routine out-patient appointments. On the hospital ward, vital sign measurements are often used to prioritise patients for referral to higher levels of care.

Below we present the six vital signs commonly used for physiological assessment and the various techniques that are commonly used to measure vital signs in the hospital and out-patient clinic. We will also discuss some of the problems associated with the current measurement techniques, in particular with the measurement of respiratory rate.

1.3.1 Description of vital signs

Heart rate: Heart rate (HR) is the measure of the number of contractions of the heart in one minute, reported in beats per minute (bpm). Resting HR is an important indicator of the patient's clinical condition and affected by several factors such as age, fitness level, activity level, body size and medications [27, 39]. The establishment of a normal value of HR is complicated by the difficulty of establishing standard conditions at rest. In non-pregnant adults, a normal resting HR is typically between 50 and 90 bpm [27].

During normal pregnancy, physiological cardiovascular adaptations occur in order to meet the increasing metabolic demands of the developing fetus. Cardiac output (CO)⁶ increases by 50% in the first trimester of pregnancy, together with an increase in blood volume. While it is unclear whether the observed increase in CO is primarily driven by an increase in HR or stroke volume, the fact that HR progressively increases during pregnancy is well established. At full term (gestational age 37 to 40 weeks), the resting HR is typically 10 to 15 bpm (5 to 20%) higher than pre-pregnancy [40, 41].

Oxygen saturation: Oxygen saturation (SpO_2) is a measure of the oxygen content in the blood, defined as the ratio of haemoglobin bound to oxygen (oxyhaemoglobin) to the total haemoglobin (oxyhaemoglobin and deoxyhaemoglobin). It is a function of the partial pressure of oxygen, as described by the sigmoid-shaped oxygen dissociation curve in Figure 1.4a [42]. In healthy adults, normal levels of peripheral SpO_2 are between 95

⁶The product of heart rate and stroke volume, defined as the volume of blood pumped by the left ventricle of the heart in one beat

and 100%. A low measurement of SpO_2 is a late sign of an airway, breathing or circulation problem. Hypoxia is diagnosed at a value of SpO_2 below 90% and at values below 80%, organ failure may incur [43].

Little is known about how SpO_2 changes in pregnancy. It is hypothesised that in order to meet an increase in the demand of oxygen of tissues such as the placenta and fetus, blood volume increases and the affinity of haemoglobin to oxygen decreases. This manifests as a right-shift of the oxyhaemoglobin dissociation curve, as illustrated in Figure 1.4b. A right-shift of the curve implies that a greater partial pressure of oxygen is required to maintain the same level of SpO_2 . The extent of this right-shift has been observed to be directly related to the duration of pregnancy [44, 45].

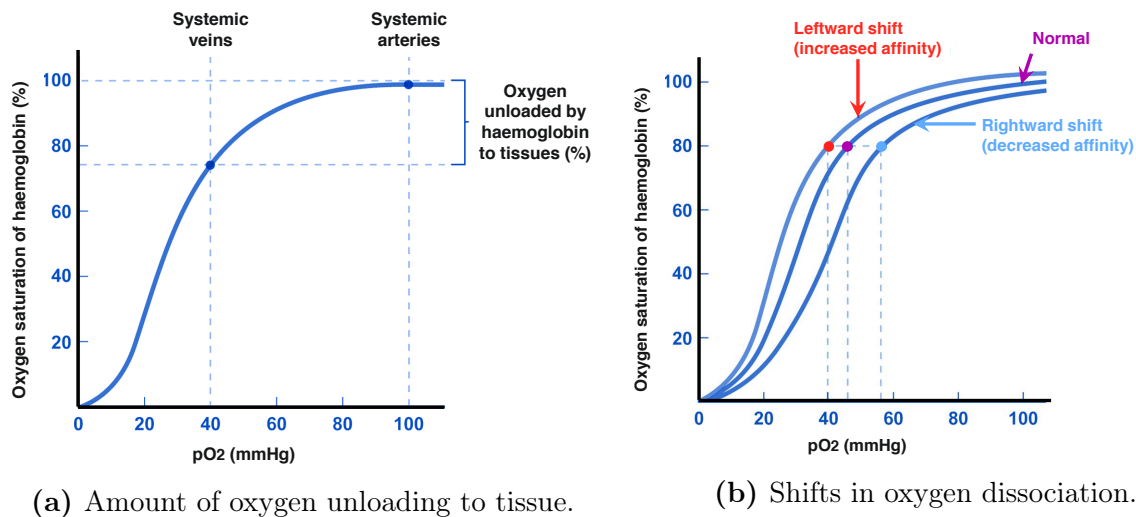


Figure 1.4: The dissociation of oxygen from haemoglobin is described by a sigmoid curve that depends on the partial pressure of oxygen.

Blood pressure: Blood pressure (BP), reported in millimeters of mercury (mmHg), is a measure of the pressure exerted by blood against the arterial wall when the heart is contracting (systole) and relaxing (diastole). In obstetric care, a measurement of BP is the most commonly used screening method for gestational hypertensive disorders [40]. According to the NICE guidelines for hypertension in pregnancy, a blood pressure higher than 140/90 mmHg (systolic BP/diastolic BP) is diagnosed as mild hypertension.

Treatment is initiated at 150/100 mm Hg (or 140/90 mmHg with organ failure) and severe hypertension is diagnosed at 160/110 mmHg [10].

In non-pregnant women, BP values depend on age, BMI and factors such as the circumstances under which the measurement is taken and the time of day. Pregnant women are affected by many of these factors but typically exhibit a narrower range of BP. In normal pregnancy, a gradual lowering of maternal BP until mid-pregnancy (22 to 24 weeks) is expected. This is thought to be a result of the effect of progesterone (a hormone) on the smooth muscle of the vasculature and increases in endothelial production of vasodilating factors. This is expected to be followed by a subsequent rise in BP from mid-pregnancy until delivery which is a consequence of the rise in blood volume. This pattern has been observed for both systolic BP (SBP) and diastolic BP (DBP); however, the mid-pregnancy dip is typically more pronounced for DBP [41, 46, 47].

Temperature: Body temperature (Temp), reported in Celsius degrees ($^{\circ}\text{C}$), is a measure of the body's temperature. The body's ability to maintain its core temperature is known as thermoregulation. In healthy individuals, normal Temp is generally considered to be 37.0°C . It may vary by 0.5 to 1.0°C according to circadian rhythm and menstrual cycle or by as much as 2 to 3°C in response to heavy exercise. An abnormal Temp can be a reflection of an abnormal physiological process, a drug, or an environmental influence [48].

The existing literature on normal values of Temp during pregnancy is limited. In general, as maternal hormonal and metabolic levels are progressively changing, small but transient changes in Temp may be expected. A small increase in temperature of 0.5°C with a subsequent slight decrease in late pregnancy has been described, which could be related to decreases in progesterone and physical activity during this time [49].

Respiratory rate: Respiratory rate (RR), or breathing rate, is defined as the number of inspirations or expirations in one minute, measured in respirations (or breaths) per minute (rpm). A change in RR is a sensitive indicator of critical illness which often accompanies, and may precede, changes in HR and SpO_2 levels [50, 51].

During pregnancy, the additional energy requirements of the fetus and placenta increases the woman's metabolic need. To ensure adequate oxygenation of the increased blood volume, maternal respiratory efficacy must therefore be improved to ensure that the maternal blood is adequately oxygenated. The pregnant woman increases her ventilation by breathing more deeply rather than more frequently. Consequently, tidal volume⁷ is expected to increase by 30 to 40% from the 2nd trimester to term, while RR remains unchanged at 14 to 17 rpm [40,45].

1.3.2 Methods for measurement of vital signs

Heart rate: The gold standard for measurement of HR is a manual measurement by palpation of blood vessels in the wrist or neck [52]. During triage, an electronic HR measurement can be readily obtained from a photoplethysmogram (using a pulse oximeter) or blood pressure signal (using an automated blood pressure device). On the ward, HR is commonly measured from a continuous signal such as the electrocardiogram (ECG).

Oxygen saturation: The gold standard for measurement of oxygen partial pressure is an analysis of arterial blood gas from a blood sample. However, as this method is impractical and invasive, a measurement of SpO₂ is normally used instead, which is measured non-invasively using a pulse oximeter [43].

The pulse oximeter measures SpO₂ using the absorption of light by tissue. Most devices contain a light source and a light detector. The light source emits light at two different wavelengths that are transmitted through (or reflected from) the finger, earlobe or toe, depending on the sensor used. Typically, wavelengths of 660 and 940 nm are chosen, corresponding to red and infrared light. At low wavelengths, haemoglobin that is fully saturated with oxygen has a different absorption coefficient to de-oxygenated blood. However, at wavelengths greater than 805 nm (i.e. infra-red light), oxygenated

⁷Tidal volume is defined as the amount of air inspired per respiration

and de-oxygenated blood have similar absorption coefficients. The proportion of oxy- and deoxyhaemoglobin in the blood is thus derived from the absorption measured at the two wavelengths at the site of measurement [53, 54].

The signal recorded by the pulse oximeter at each wavelength is called the photoplethysmogram (PPG). Most commercial devices output the infra-red signal, which is less sensitive to changes in haemoglobin saturation and therefore more stable over time compared with the red signal. The PPG is a cardiac-synchronous pulsatile signal which depends cardiac activity and changes in blood volume. Heart rate can be calculated from the time between consecutive PPG peaks. Occasionally, a secondary, smaller peak may be observed, called the dicrotic notch. The dicrotic notch reflects the elasticity of the arteries and pulse velocity. It disappears with age as arteries get stiffer and pulse velocity decreases [54]. An example of a good-quality PPG signal, recorded from a 4P woman at 16 weeks of gestation, is shown in Figure 1.5.

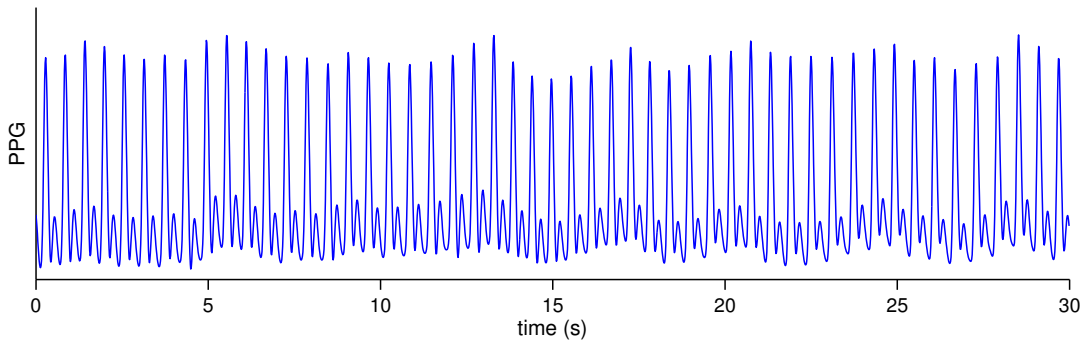


Figure 1.5: Example of a good-quality 30-second PPG signal with a prominent dichrotic notch

Blood pressure: In standard clinical practice, BP is generally measured with an auscultatory (manual) or oscillometric (automated) BP device using an inflatable brachial (arm) cuff. In pregnancy, the auscultatory method has traditionally been the gold standard, where BP is read off a mercury or aneroid sphygmomanometer by an observer who uses a stethoscope to listen for the Korotkoff (heart) sounds. Oscillometric monitors are now becoming more widespread in use. Other measurement techniques, such as those based on ultrasound or finger cuffs are less widely used in the clinic [55]. For measurements

performed by a patient at home, only validated automated oscillometric arm cuff devices are recommended [56]. Any automated monitor used to measure maternal BP must be validated for use in pregnancy [57]. This is particularly important for women in their third trimester and for women with high BP, for whom automated monitors not suitable for use in pregnancy tend to underestimate BP. This is a result of the way that SBP and DBP are *estimated* from the mean arterial pressure using the oscillometric principle, rather than *measured*. The oscillometric principle is explained in Appendix C.1.1.

Temperature: The most common methods for the measurement of Temp are listed in Table 1.1. In the outpatient clinic, tympanic thermometers are generally preferred because these are non-invasive and quick and easy to use relative to the other methods. Tympanic thermometers measure the infrared heat generated by the eardrum and its surrounding tissue and convert this to a Temp reading [58]. Tympanic Temp measurements have been observed to be somewhat lower and more variable than measurements obtained from thermistors (measuring core body temperature) or oral thermometers [59].

Table 1.1: Accuracy of methods used for measuring temperature (reproduced from O’Grady *et al.* [48])

Methods used for measuring temperature
Most accurate
Pulmonary artery thermistor
Urinary bladder catheter thermistor
Esophageal probe
Rectal probe
Other acceptable methods in order of accuracy
Oral probe
Infrared ear (tympanic) thermometry
Other methods less desirable
Temporal artery thermometer
Axillary thermometer
Chemical dot

Respiratory rate: To measure RR, the World Health Organization recommends a count of breaths for one minute by observation of chest wall movements or auscultation [60].

This differs from standard clinical practice on the ward, where breaths are generally counted for only 15 seconds and the count is multiplied by four to obtain the respiratory rate in respirations (breaths) per minute. Electronic measurements for respiratory rate exists, but due to their high complexity and cost, these are not suitable for use in the antenatal clinic [61].

1.4 The need for an electronic measurement of respiratory rate

Of the six vital signs, RR remains the only vital sign that currently is not measured electronically during routine clinical assessment [61,62]. It is also the only vital sign that is not entirely controlled by the autonomic nervous system, thus it can to some extent be affected and manipulated by conscious control [63].

Clinical staff have frequently been observed to count breaths for 15 seconds only. To obtain the RR in breaths or respirations per minute (rpm), this count is subsequently multiplied by four [64]. This practice causes greater quantisation of the measurement and amplifies any potential counting error. In a comparison of RR measured by nurses in the Emergency Department (counting chest wall movements for 15 seconds) to the gold standard (counting chest wall movements for 60 seconds), Lovett *et al.* [65] found a low correlation between the two methods. More recently, Bianchi *et al.* [61] found that respiratory rates estimated from a 15-second nurse count resulted in estimates with lower variability than the 1-minute reference criterion, as well as a lower sensitivity for measuring abnormally slow (< 12 rpm) or fast (> 20 rpm) respiratory rates. In particular, they found that 77% of tachypnoeic (> 20 rpm) cases were missed when chest wall movements were only counted for 15 seconds.

That a manual measurement is less objective than an electronic measure is unavoidable. Moreover, nurses have frequently been assumed to *estimate* RR rather than formally counting the number of breaths [64]. Even when RR is measured for the full 60 seconds,

poor interobserver agreement has been observed. For example, Edmonds *et al.* [62] found independent, consecutive measures of RR to differ by more than 35%; a greater inaccuracy than that reported for measurement of any other vital sign.

Finally, RR is the only vital sign that is not entirely controlled by the autonomic nervous system and therefore to some extent affected by conscious control. This may further complicate the measurement. In a study by Hogan *et al.* [63], patients were observed to change their respiratory pattern in response to a nurse counting their breaths. In spite of these difficulties, changes in RR have been shown to be a sensitive indicator of critical illness which often accompanies, and may precede, changes in heart rate and oxygen saturation levels [50,51].

1.5 Aims and overview of this thesis

The overall aims of the investigations presented in thesis are *(i)* to establish what the expected changes in vital signs (HR, SpO₂, SBP, DBP, Temp and RR) associated with normal pregnancy are; and, *(ii)* to use this knowledge to design the first evidence-based, gestation-specific MEOWS.

1.5.1 Overview of thesis chapters

This thesis is split into seven chapters. In this chapter, we described the need for a gestation-specific MEOWS, to be based on the knowledge of the normal ranges of vital signs over the course of pregnancy and the immediate postpartum period.

In Chapter 2, a systematic review is presented which summarises the current evidence base of studies that measure changes in vital signs across pregnancy. Due to the limited evidence available for all six vital signs, the systematic review only included studies of measurements of SBP, DBP and HR during the antenatal period. We present a quality assessment used to qualitatively assess the current evidence-base and centile charts that summarise the evidence in a quantitative form. This systematic review was conducted in

collaboration with colleagues from the Department of Engineering Science, Department of Primary Care Health Sciences, Centre of Statistics in Medicine, Nuffield Department of Obstetrics & Gynaecology and Nuffield Department of Clinical Neurosciences.

In Chapter 3, the Prediction of Physiology Patterns in Pregnancy (4P) study is introduced. This study was set up in order to collect vital sign measurements from 1,000 women during pregnancy and the immediate postpartum period. Since the establishment of accurate reference ranges of vital signs demands measurements of the highest possible quality and accuracy, the discussion focuses on how we designed a custom-made system to ensure that the vital sign data were reliably collected. We also discuss how the measurements were validated.

In Chapter 4, methods for the electronic estimation of respiratory rate are presented. Using data from the 4P study, we develop and deploy signal processing techniques for estimation of respiratory rate from two different signals: the accelerometer signal collected using the in-built accelerometer of a smartphone and the photoplethysmogram collected using a pulse oximeter. The electronic estimates are compared to reference measurements (a manual count of breaths by a midwife). A discussion of the benefits and limitations of the different approaches is also presented.

In Chapter 5, vital sign data from the 4P study are used to design gestation-specific reference ranges of normal values of vital signs throughout the antenatal period. A range of methods for the construction of these charts are compared, including multi-level models, Gaussian Processes, kernel regression and empirical centile estimation.

In Chapter 6, we describe how the gestation-specific reference ranges were used to design the first evidence-based MEOWS score. We also include some examples of how the score could be used to identify women who suffer from complications during their pregnancy.

Finally, in Chapter 7, the main findings of this thesis are summarised and future work is discussed.

Chapter 2

Systematic review of physiological trends in pregnancy

Assessment of vital signs is a key part of the clinical assessment of pregnant women. Our understanding of the normal thresholds of these vital signs, such as the thresholds recommended by NICE for diagnosing hypertension in pregnancy [10], underpins their use. Current MEOWS thresholds use the perceived normal ranges of vital signs, which appear to be based on clinical consensus [4, 22]. Many modern clinical guidelines do not reference their sources for normal ranges [10, 17, 66]. Others refer to a core textbook, which references data from small individual studies published between 1970 and the mid-1990s [40]. The evidence underpinning current guidance is therefore weak.

The dynamic changes in maternal vital sign physiology associated with pregnancy complicate the use of vital signs to detect physiological deterioration [14]. While we know that vital signs change over the course of pregnancy, we have not been able to identify a clinical guideline or MEOWS which accounts for these expected changes. Accurate reference ranges, which take into account the expected changes in vital signs during pregnancy, are therefore required.

Vital signs are commonly recorded at a particular stage of pregnancy in different types of clinical studies. We postulated that large quantities of data may already be available to

inform vital sign thresholds. To investigate this we have conducted a systematic review, which aims to synthesise the current evidence base of studies that describe changes in vital signs during normal pregnancy.

2.1 Systematic review methods

This systematic review of observational studies was conducted and reported following the PRISMA guidelines [67, 68]. It has been registered on PROSPERO (registration number CRD42014009673). The review protocol, developed in collaboration with clinical colleagues and medical statisticians, has been published [69].

2.1.1 Search strategy and selection criteria

We searched electronic databases (MEDLINE, EMBASE, CINAHL from inception until April 2015) and reference lists to identify studies that measured vital signs in healthy (low-risk) pregnant women. Both MeSH¹ and free-text terms were used, without any language restrictions. The complete search strategy is included in Appendix A.1.1.

Studies with objective measurements of BP and HR made during the antepartum period by a healthcare professional were included, according to the inclusion and exclusion criteria described in Table 2.1. We included both cross-sectional studies (where each woman is measured once) and longitudinal studies (where each woman is measured at multiple occasions). Initially, we set out to include measurements of all six vital signs commonly used for clinical assessment (SBP, DBP, HR, SpO₂, RR and/or Temp), made during any of the three stages of pregnancy (antepartum, intrapartum and/or postpartum). However, we identified only a limited number of studies which described changes in the other vital signs and/or stages of pregnancies. Consequently, the scope of the review was narrowed to include only BP and HR data measured during the antepartum period. An obstetric physician (Dr Lucy Mackillop) and an intensive care

¹Medical Subject Headings (MeSH terms) are the National Library of Medicine’s controlled vocabulary used to index databases of research articles (<https://www.nlm.nih.gov/mesh/>)

specialist (Dr Peter Watkinson) assessed eligibility of studies for inclusion, and disagreements were resolved by consensus.

2.1.2 Data extraction and summary measures

For each study, we extracted information about the study (year of data collection and publication, study setting, country of study, data collection schedule), measurements (method, subject position, details of measurement device) and participants (age, weight, BMI², ethnicity, parity³, singleton/multiple pregnancies⁴, pregnancy dating method, reason for measurements). For each *period of gestational age* (gestational age band) reported in a paper, we extracted information about the associated sample size, sample characteristics and reported summary statistics for BP and HR. The period of gestational age is defined as the range of gestational ages associated with each reported measurement occasion. The following summary statistics were extracted:

- mean and/or median value
- standard deviation (*SD*) and/or standard error of the mean (*SE*)
- centiles, percentiles, quartiles etc.
- confidence intervals

Measurement data from multiple groups (e.g. for different ethnic groups or for subgroups defined based on subsequent assignment of a medical diagnosis) were classified independently. Where multiple measurements were presented for the same group at the same measurement occasion, a single data point was selected to avoid over-representation, using the pre-specified rules summarised in Appendix A.1.3. For studies that did not report relevant summary statistics (mean $\pm SD$ or mean $\pm SE$ of vital signs and gestational age), these were estimated by assuming a normal distribution, as described below. We did not formally assess skewness as most papers did not report the required summary statistics.

²Body Mass Index (BMI) is body weight divided by height squared (kg/m^2)

³Parity is the number of previous births (irrespective of neonatal outcome).

⁴Whether the woman carries one or more babies

Table 2.1: Inclusion and exclusion criteria for the systematic review

Inclusion criteria
1. Cross-sectional, case-control, or longitudinal study 2. Minimum of 50 participants 3. Age 14 years or older 4. Objective measurement[†] of heart rate, respiratory rate, blood pressure, oxygen saturations or temperature 5. Measurements taken during the antenatal period, up to the start of the intrapartum period* 6. Raw data or average measure reported and possible to extract within minimum accuracy
Exclusion criteria
1. Measurements from women with illnesses likely to affect the cardiac or respiratory systems[†] 2. Measurements from women at high risk of developing complications[†] 3. Measurements from women who at the time of recruitment were recruited for a specific study group 4. Measurements from women known to be taking medication which could affect the measurements 5. Measurements from women where the gestational age at the point of measurement was not reported in terms of weeks or days of gestation 6. Measurements from women where the reported gestational age band exceeded 16 weeks 7. Measurements from self-monitoring or other measurements not taken by a healthcare professional 8. Measurements from a group of women for whom fewer than 10% of the pregnancies were singleton pregnancies 9. For women undergoing fertility procedures, any measurements taken prior to a positive pregnancy test 10. Any of the following measurements (without valid baseline): – Measurements taken using ambulatory technologies – Measurements taken using invasive technologies – Measurements taken during anaesthesia – Measurements taken during sleep – Measurements taken during exercise – Measurements taken at heights greater than 1000m above sea level
[†] Defined in Appendices A.1.2 and A.1.3
* Defined as progressive cervical dilatation with regular contractions

Vital sign measurements: Where the median was reported as the only measure of central tendency, we considered reported median values to be equivalent to the mean (assuming a normal distribution). To represent the spread of data, both SD and a number of representative centiles were modelled. Where SE was reported as the only measure of spread, this was converted to a standard deviation using Equation 2.1, where SD is standard deviation, SE is standard error and n is sample size.

$$SD = SE\sqrt{n} \quad (2.1)$$

Where centiles of vital signs were reported as the only measure of spread, Equation 2.2 was used to convert to SD . Here, X is the value of the variable at the P^{th} centile and Z is the value of the P^{th} centile in a unit normal distribution (i.e. a distribution with $\mu = 0$ and $SD = 1$). As an example, for the $P = 95^{th}$ centile, $Z = 1.96$.

$$SD = \frac{X - \mu}{Z} \quad (2.2)$$

Gestational age: Where the gestational age associated with each data point was reported as a mean gestational age $\text{mean}(t)$ (where t is in weeks), we used this for the analysis. The period of gestational age was then estimated as $[\text{mean}(t) - SD(t), \text{mean}(t) + SD(t)]$ where $\text{mean}(t)$ and $SD(t)$ is the reported mean and standard deviation. Where the period was reported as a full range $[\text{min}(t), \text{max}(t)]$, the mid-point of the range was taken as the mean. Where reported as a single week t , the range was set to $[t, t + 1]$ for consistency, as this generally implies a period of $t^{+0} \text{ days}$ to $t^{+6} \text{ days}$.

Sample size: Where the sample size associated with each measurement occasion was reported, we used this for the analysis. For studies where sample size was only reported at baseline, we imputed this to be the sample size associated with each measurement occasion, assuming no drop-outs or lost-to-follow-up.

Sample characteristics: Where characteristics of women (such as parity, weight/BMI, age) were reported at multiple measurement occasions (e.g. pre-pregnancy or at recruitment), we used the value reported closest to the time of recruitment.

Study year: If the year of data collection for the study was reported, we used the last year of data collection as the study year. If the year of data collection was not reported, this was taken to be the year of publication of the study.

2.1.3 Assessment of bias

To assess sources of bias and heterogeneity, a list of methodological quality criteria was developed in advance of the review. These were adapted from previously published quality criteria [70, 71] and covered two domains: study design and reporting methods. Because the study objectives, and hence the statistical methods, differed greatly between studies we did not score statistical methods as a separate domain. Each criterion was scored as either *high* or *low* risk of bias. The quality criteria are summarised in Table 2.2.

Two reviewers (Lise Loerup and Rebecca Pullon) extracted the study data and performed the quality assessment of each study. The data were independently abstracted onto a spreadsheet (Microsoft Office Excel) which automatically highlighted disagreements. Disagreements were resolved by recourse to the original papers or in consultation with a medical statistician (Ms Jacqueline Birks), intensive care specialist (Dr Peter Watkinson) or obstetric physician (Dr Lucy Mackillop), as required. The overall quality score was defined as the percentage of *low risk of bias* marks over the total number of criteria for each study.

2.1.4 Data synthesis

To examine trends in vital sign values with gestational age, we used the mean, standard deviation and representative (1^{st} , 10^{th} , 25^{th} , 75^{th} , 90^{th} and 99^{th}) centiles of BP and HR from each study. The methods for quantitative data synthesis are described in the following section. As part of this analysis, we conducted a subgroup analysis to investigate the effect of study-level covariates including method of measurement, parity and study year. The analysis was done in MATLAB (v12b, MathWorks Inc) and R (v1.3.3).

Table 2.2: Quality assessment criteria for the systematic review

Assessment criteria	Low risk of bias	High risk of bias
Study design		
1. Study design	Clearly described as either cross-sectional or longitudinal (accept follow-up, serial etc)	Not reported OR mixture of cross-sectional and longitudinal data
2. Sample selection	Population-based study where there are attempts to identify and clearly define populations from a specific geographical area OR women selected either consecutively or at random	Not population based OR convenience sampling OR arbitrary recruitment
3. Number of assessments (cross-sectional only)	Each woman was measured and included only once	Some women were measured and included more than once
4. Gestational age interval of assessments (longitudinal only)	Interval of measures prospectively pre-specified and justified	Interval of measures not prospectively pre-specified and justified
5. Number of assessments (longitudinal only)	Clear documentation of the intended number of serial measurements	No clear documentation of the intended number of serial measurements
6. Inclusion and exclusion criteria	Women at high risk of pregnancy complications were not included AND all women were included in the final analysis	The study population included both low-risk and high-risk pregnancies OR women with abnormal outcome were excluded OR criteria not well explained
7. Sample size calculation	A priori determination of sample size and justification	Lack of a priori sample size determination and justification
8. Data collection time point	Prospective study	Retrospective study
9. Data collection purpose	Vital sign data collected (and analysed) specifically in order to investigate values or trends of vital sign in pregnancy	Data collected for another purpose (e.g. for validation of measurement equipment)
10. Pregnancy dating method	Clearly described AND a sonogram before 14 weeks	Sonogram after 14 weeks OR sonogram at unspecified gestational age OR sonogram not primary dating method
11. Collection of data on gestational age	The gestational age was calculated precisely to the day	Truncation of gestational age to the number of <i>completed weeks</i>

Continued on next page ...

... Table 2.2 Continued from previous page

Assessment criteria	Low risk of bias	High risk of bias
12. Number of measurements per assessment	More than one assessment per woman per appointment	Single measure OR not specified
13. Suitability of equipment used	Any BP monitor used approved for pregnancy OR calibrated sphygmomanometer	No evidence of calibration of sphygmomanometer OR BP monitor not approved for pregnancy
Reporting methods		
14. Characteristics of study population	Presented in a table or clearly described AND includes minimum dataset of age AND weight AND height (or BMI) and parity	Not presented in a table OR not clearly described OR does not contain minimum dataset
15. Number of women approached and enrolled (prospective studies only)	Both described	Not described OR only number enrolled described
16. Details of equipment used	Manufacturer and model for each device stated OR mercury sphygmomanometer (in case of BP measurement)	Missing information about some or all devices
17. Qualification of observers	Measurement taken by healthcare professional OR trained personnel	Taken by anyone else OR unreported
18. Number of observers	Reported	Unreported
19. Description of measurement technique	The study described sufficient and unambiguous details of the measurement techniques; including position of subject appropriate cuff size (BP only).	Description insufficient OR ambiguous
20. Description of raw data	(Presented in a table OR clearly described) AND (Presented per week of gestation OR for a clearly defined period of gestation)	Not presented in a table OR unclear OR incomplete

2.2 Methods for construction of centile charts

Data extracted for the systematic review were in the form of summary statistics, i.e. data combined from measurements of several women measured within a period of gestational age (time range). In statistics, such data are referred to as *aggregate data*. Aggregate data have several characteristics that must be accounted for in the choice of analysis method:

1. Aggregate data points from different studies may represent different sample sizes
2. Studies may report aggregate data across periods of gestational age (time ranges) of different length
3. The timing and total number of data points may vary across studies according to the studies' data collection schemes
4. While measurements at multiple measurement occasions from longitudinal studies are correlated (repeated measurements), multiple measurements from cross-sectional studies are independent
5. Studies generally do not report the correlation between measurements and the effect of missing data at the patient level.

Methods for analysis of longitudinal aggregate data are not well established; a recent review of meta-analysis methods by Jones *et al.* [72] found that the most frequently used approach is to ignore the correlation between repeated observations. Based on this, the review recommended that *"It is preferable to obtain IPD [individual patient data] from all studies to correctly account for the correlation between repeated observations"*. For studies included in our review, IPD data were generally not available.

From the literature, we have identified two methods that have been specifically developed for analysis of aggregate data; weighted variable bandwidth kernel regression proposed by Fleming [73] and a multilevel model adapted for the analysis of longitudinal aggregate data proposed by Ishak *et al.* [74]. The exact specifications of these models are described in Appendices B.2 and B.4, respectively. Here, we will discuss the merits and limitations of each method. Table 2.3 summarises the key similarities and differences between the two methods. For the purpose of this thesis, both methods will be applied to our data set and compared.

Table 2.3: Comparison of different methods for estimation of centiles from aggregate data

Model	Multilevel model	Kernel regression
Distribution assumption	Normal distribution	No distribution assumed
Smoothing method	Polynomial specification (conventional or fractional) or splines fitted to individual vital sign trajectories	Non-parametric specification using smoothing parameter h
Yields equation for prediction	Yes	No
Independence assumption	Dependencies accounted for by random terms	Data assumed independent
Additional considerations when adapted for aggregate data		
Model name	Multilevel aggregate data model (MLAD)	Weighted variable bandwidth kernel regression (WVBKR)
Accounts for different sample sizes (point 1)	Indirectly, data weighted by variance of the mean (SE^2)	Yes, data weighted by sample size (n)
Accounts for different time ranges (point 2)	No	Yes, includes variable bandwidth
Accounts for different data collection schemes (point 3)	Yes, data weighted by variance of the mean (SE^2)	No
Accounts for correlated repeated measurements (point 4)	Yes, through specification of covariance structures $\mathbf{D}$ and $\mathbf{R}$	No

2.2.1 Weighted variable bandwidth kernel regression

Weighted variable bandwidth kernel regression (WVBKR), an adaptation of the classical kernel regression described in Appendix B.1, was first proposed in order to design centile charts of changes in HR and RR data with age in children [73, 75]. The method uses a non-parametric specification of time, which implies that the shape of the curve is not constrained by the shape of a pre-defined function. Instead, the smoothness and shape of the curve is adjusted using the smoothing parameter h_c , also referred to as the common bandwidth multiplier, and p , the order of the local polynomial kernel estimator. Moreover, it does not assume a distribution of the measured response, hence the mean and centiles are estimated separately.

WVBKR accounts for the fact that the aggregate data points represent different

sample sizes and periods of gestational age (points 1 and 2) by assigning a weight w_j and variable bandwidth h_j to each data point j . A key limitation is that it treats all data points as independent. As a result, studies with more data points are assigned a greater weight in the computation of the overall trend and any correlation between longitudinal data points from the same study are not accounted for (points 3 and 4).

2.2.2 Multilevel model for aggregate data

The multilevel model is also referred to as a hierarchical linear model [76] because it accounts for the hierarchical structure of data. In the case of aggregate data, the hierarchical structure is in the form of repeated measurement occasions (level 1, subscript j) nested within studies (level 2, subscript i). The model is defined in terms of *fixed* effects, which define the shape of the underlying population mean, and *random* effects, which represent individual variation that cannot be explained by characteristics of the population. Correlations between data points are specified by the within-studies (level 1) and between-studies (level 2) covariance matrices. These are denoted $\mathbf{R}_i$ and $\mathbf{D}$, respectively. The shape of the curve is typically defined as a parametric function of time. Additional study-level covariates may be added to explain some of the variation between studies. The model is based on strong assumptions of normality: at level 1, the data points are assumed to be normally distributed about a mean within-study trajectory; and, at level 2, the mean study trajectories are assumed to be normally distributed about an overall population mean. In this thesis, we refer to the pattern of vital sign values as they evolve over time as the vital sign *trajectory*. A comprehensive description of the general multilevel model and the associated covariance matrices is given in Appendix B.3.

Ishak *et al.* [74] have proposed a modification of the multilevel model which we will refer to as the multilevel aggregate data model (MLAD). This differs from the general multilevel model, as the entries of the within-study (level 1) covariance matrix $\mathbf{R}_i$ are populated using the reported values from each study, as opposed to estimated from the data. While Ishak *et al.* set the diagonal elements of $\mathbf{R}_i$ to the reported variance (SD_{ij}^2)

of the measurements at each measurement occasion, we propose using the variance of the mean (SE_{ij}^2) to explicitly account for differences in sample size between studies (point 1). The SE_{ij}^2 for study i at measurement occasion j is given by Equation 2.3, where SD_{ij}^2 is the reported variance and w_{ij} is the associated sample size.

$$SE_{ij}^2 = \frac{SD_{ij}^2}{w_{ij}} \quad (2.3)$$

By fixing the diagonal elements of $\mathbf{R}_i$, studies that report a lower variance and a higher sample size will contribute more towards the estimation of the mean. This weighting also accounts for differences in the data collection schemes of the included studies (point 3). Each study only contributes once to the estimation of the mean, but with a weighting that is dependent on the number and spacing of reported measurement occasions. Because the within-study (level 1) variances are set using study-specific information (as opposed to estimated from the data), it is not possible to perform inference to estimate centiles for the measured responses⁵. As for the general multilevel model, dependencies between longitudinal data points (point 4) are modelled using the off-diagonal elements of the within-study (level 1) and/or between-study (level 2) covariance matrices. This is described in more detail in the following section and in Appendix B.4.

2.2.3 Selection of model parameters

Because WVBKR and MLAD account for very different characteristics of the data, we have implemented and compared both models. Note that while both models yield an estimate for the pooled mean, only WVBKR may be used to estimate pooled centiles from aggregate data. The model parameters were selected empirically by testing a range of values and selecting values that produced good fits across the full range of gestational ages. Additional considerations for each method are described below.

⁵Note that this is not the case for the general multilevel model where the covariance matrices are estimated entirely from the data

Parameter estimation for weighted variable bandwidth kernel regression

For SBP, DBP and HR, we chose a polynomial kernel of order $p = 1$ which reflected the fact that the data did not tend to a horizontal line at either boundary (for which $p = 0$ would have been the preferred choice). The value of the smoothing parameter h_c must be set to account for a trade-off between optimising for a close fit to the data, using a small value of h_c , and optimising for a smooth curve, using a large value of h_c . For SBP and DBP, a value of $h_c = 3$ was found to produce a good fit to the data. For HR, we chose a value of $h_c = 4$. Given that the data available for measurements of HR were limited, it is expected that a higher value of h_c would be needed to obtain a smooth curve. Centiles were estimated separately using the standard deviations and centiles, either using the values reported in the papers or obtained with appropriate conversions, as described in Section 2.1.2.

Parameter estimation for the multilevel aggregate data model

For MLAD, the main modelling choices are the number and types of fixed and random effects. The number of fixed effects depends on the choice of polynomial parameterisation of gestational age (time) and of study level covariates. The polynomial function of time was chosen based on the shape of the curves reported in the individual studies. The number of random effects was estimated based on inspection of a number of criteria for assessment of model fit, including the likelihood, Akaike's Information Criterion and the Bayesian Information Criterion (explained in Appendix B.3.6).

For all three vital signs, the final model was in the form of a quadratic polynomial with a random intercept and slope as described by Equation 2.4, where y_{ij} are the vital sign values at gestational ages x_{ij} , β and b_i are the fixed and random effect coefficients, respectively, and e_{ij} are the occasion-specific residuals for each study.

$$y_{ij} = \beta_0 + b_{0i} + \beta_1 x_{ij} + b_{1i} x_{ij} + \beta_2 x_{ij}^2 + e_{ij} \quad (2.4)$$

To investigate the effect of method of measurement and study year on the reported values of BP, we included these as study-level covariates in the MLAD model as shown in Equation 2.5. Here, the method of measurement is defined in the form of a dichotomous variable m_{1i} where $m_{1i} = 1$ signified an automated measurement (using an automated BP monitor) and $m_{1i} = 0$ signified a manual measurement (using mercury sphygmomanometry). The study year, $year_i$, is defined as a continuous variable $m_{2i} = year_i - year_{min}$, where $year_{min}$ is the earliest reported study year.⁶

$$y_{ij} = \beta_{00} + \beta_{01}m_{1i} + \beta_{02}m_{2i} + b_{0i} + \beta_{10}x_{ij} + b_{1i}x_{ij} + \beta_{20}x_{ij}^2 + e_{ij} \quad (2.5)$$

To investigate the effect of parity on the reported values of BP, we defined a second model with parity as the only study-level covariate as shown in Equation 2.6. Here, parity was defined as a continuous variable m_{3i} equal to the reported proportion of multiparous women. Thus $m_{3i} = 0$ for a study group with nulliparous⁷ women only and $m_{3i} = 1$ for a study group with multiparous⁸ women only.

$$y_{ij} = \beta_{00} + \beta_{03}m_{3i} + b_{0i} + \beta_{10}x_{ij} + b_{1i}x_{ij} + \beta_{20}x_{ij}^2 + e_{ij} \quad (2.6)$$

For all models, the level 1 covariance matrix $\mathbf{R}_i$ was set as a diagonal matrix using the study-reported values of SE_{ij}^2 (Equation 2.3) for each data point. Due to the limited number of data points reported for each study, as well as the wide variation between studies in gestational age associated with observations, we were not able to estimate the level 1 time-dependent correlations between observations. Consequently, the off-diagonal elements of $\mathbf{R}_i$ were set to zero. Correlations between observations were modelled using the level 2 covariance matrix $\mathbf{D}$. This matrix took an unstructured (not parameterised) form, in which all matrix entries were estimated from the data. The model parameters were estimated using maximum likelihood, as detailed in Appendix B.3.4.

⁶The details of how study level covariates are included in the Equation are given in Appendix B.3

⁷A woman is nulliparous if she has not previously been pregnant beyond 24 weeks.

⁸A woman is multiparous if she has previously been pregnant beyond 24 weeks.

As the level 1 covariance matrix is fixed rather than estimated from the data, it is not possible to estimate centiles of the underlying observations using the MLAD model. Instead, we report the spread in the study-specific trajectories across gestational age, computed as a standard deviations of the estimated within-study mean trajectories using the estimated level 2 covariance matrix. This standard deviation was computed using Equation 2.7 where D_{11} , D_{12} and D_{22} are the entries of the two-by-two level 2 covariance matrix $\mathbf{D}$ (matrix $\mathbf{D}$ is symmetric so that $D_{12} = D_{21}$).

$$\text{SD}(b_{0i} + b_{1i}x_{ij}) = \sqrt{\text{Var}(b_{0i} + b_{1i}x_{ij})} = \sqrt{D_{11} + 2D_{12}x_{ij} + D_{22}x_{ij}^2} \quad (2.7)$$

2.3 Results

Figure 2.1 illustrates the process of identification and selection of studies. The reasons for rejection of studies are summarised in Appendix A.2.1. In total, from 1,702 publications we identified 36 studies that were suitable for inclusion. Of these, 32 studies contained measurements of BP and nine studies contained measurements of HR (five of the studies contained measurements of both BP and HR). The studies were undertaken in 18 different countries, of which 16 were high or very high-income countries and two were middle or low-income countries. Only one study, by Demidov *et al.* [77], was reported in a non-English language (Russian). The details of the included studies are summarised in Appendix A.2.2.

2.3.1 Changes in blood pressure

Of the studies that reported measurements of BP, 24 studies were longitudinal and 8 were cross-sectional. In total, 139,086 measurements of BP from 36,691 women were reported at 301 measurement occasions. The longitudinal studies reported measurements from 39 independent groups of women, with summary measurements reported at between 2 and 35 (median 4) measurement occasions per group. All of the cross-sectional studies of BP,

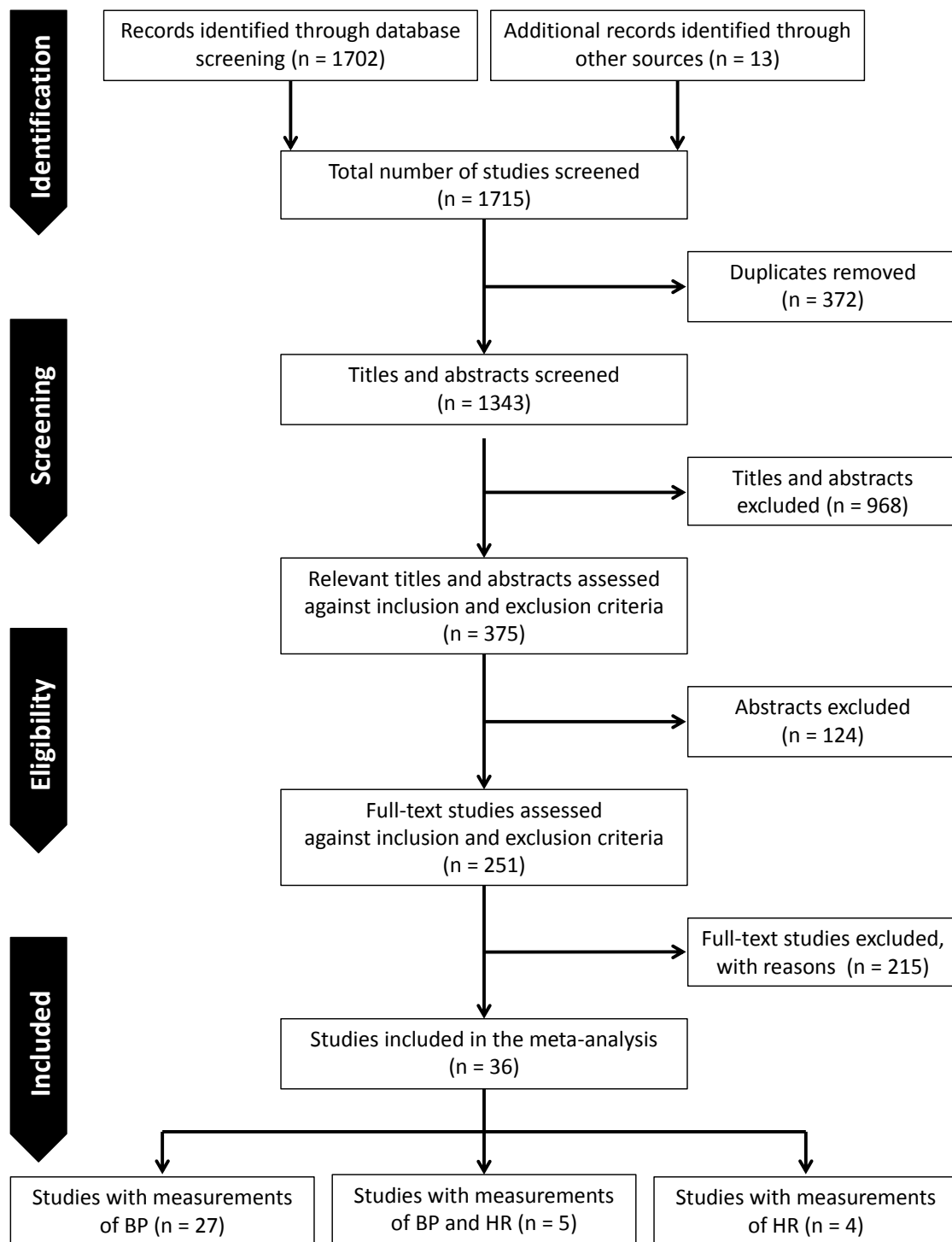


Figure 2.1: PRISMA flow diagram of the study selection process.

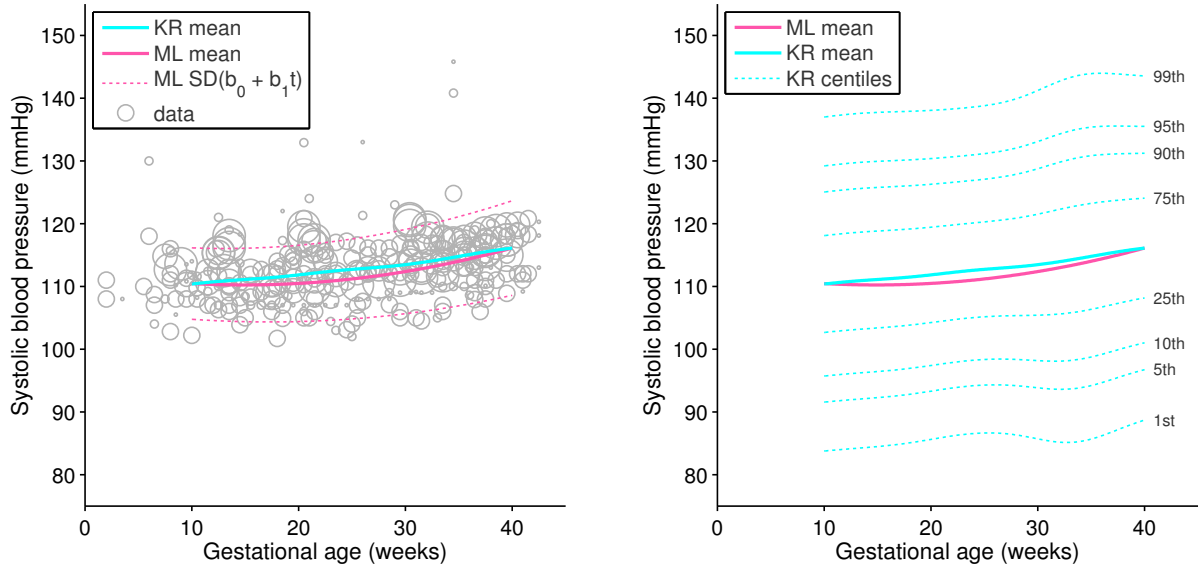
apart from one study by Ochsenbein-Kolble *et al.* [78], reported summary measurements at a single measurement occasion.

Figures 2.2a and 2.2b show for SBP and DBP, respectively, the extracted data and vital sign trajectories computed using MLAD and WVBKR. Both SBP and DBP increased from 10 to 40 weeks of gestation (mean SBP increased from 100 to 117 mmHg; mean DBP increased from 65 to 70 mmHg). However, while the trajectories for mean SBP and DBP computed using WVBKR increased monotonically, the trajectories estimated using MLAD initially decreased and reached a minimum at between 18 and 20 weeks of gestation before increasing to a maximum value at 40 weeks of gestation.

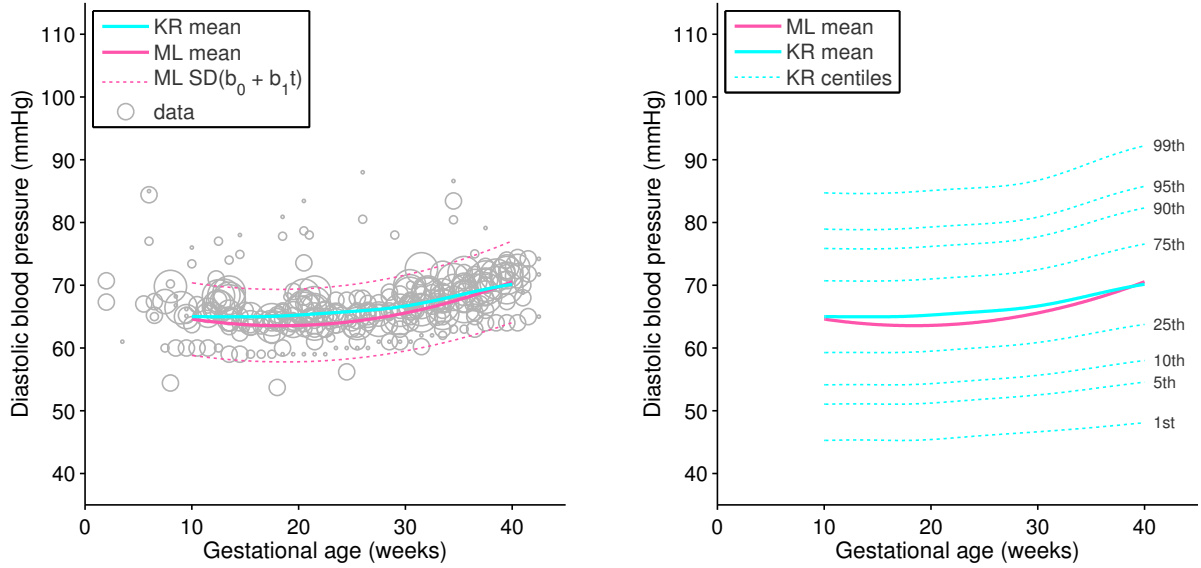
We found significant heterogeneity between studies in terms of the reported measurement values and sample sizes. This is likely to have caused the difference in shape of trajectories estimated from the two models. The difference could (*i*) be a consequence of the different weightings used in the estimation; or (*ii*) result from the relatively high value of h_c that was required to obtain smooth curves, as a large value of h_c could potentially have caused an "over-smoothing" of the underlying trend.

Across the same time range, the standard deviations of mean within-studies trajectories estimated using MLAD, given by Equation 2.7, increased slightly. This implied that the difference between studies in reported values of BP increased with gestational age. The separation between the centiles estimated by WVBKR indicated a greater variation in the values of SBP compared to DBP.

Table 2.4 summarises the results from the subgroup analysis with respect to method of measurement, study year and parity. Methods of measurement were reported as mercury sphygmomanometry (manual; 16 studies, 123 measurement occasions), automated BP monitor (automated; 9 studies, 87 measurement occasions) or not specified (7 studies, 91 measurement occasions). Studies using a manual measurement reported higher mean BP compared with studies using an automated measurement (difference of 5.6 and 8.6 mmHg for SBP and DBP, respectively). We also found a positive correlation between BP and study year (increase by 0.46 and 0.55 mmHg/year for SBP and DBP, respectively).



(a) Results from the meta-analysis of SBP



(b) Results from the meta-analysis of DBP

Figure 2.2: Results from the meta-analysis of BP. **[left]** The extracted BP data (grey circles, sizes reflect sample size) with mean fitted using MLAD (solid pink line), standard deviation of the within-study means (dotted pink lines) and mean fitted using WVBKR (solid light blue line). **[right]** Mean fitted using WVBKR (solid light blue line), associated centiles (dotted blue lines) and mean fitted using MLAD (solid pink line).

Table 2.4: Comparison of different methods for estimation of centiles from aggregate data

Covariate description (unit)	Coefficient	SBP	DBP
Method of measurement (mmHg)	β_{01}	−5.6	−8.6
Study year (mmHg/year)	β_{02}	0.46	0.55
Parity (mmHg)	β_{03}	−4.6	−4.7

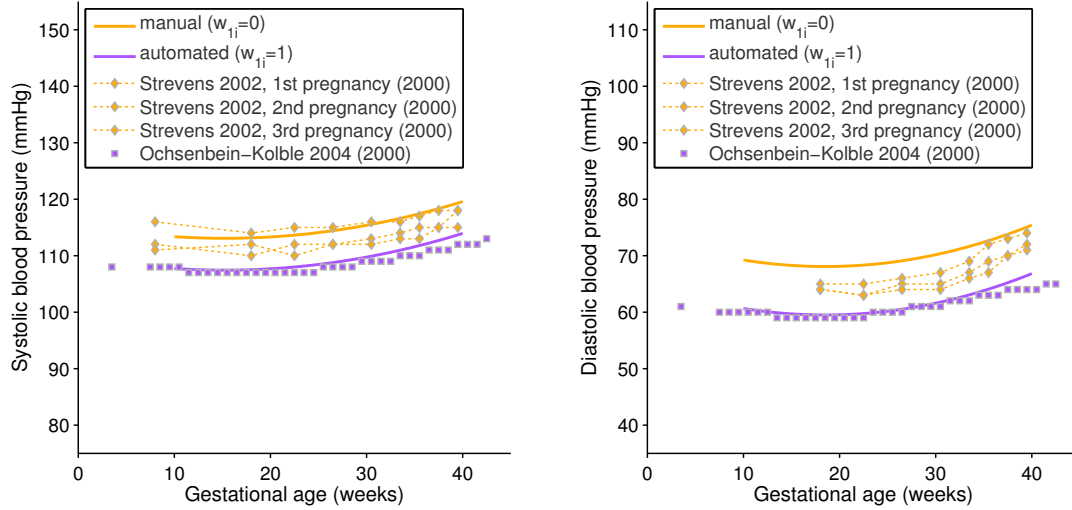


Figure 2.3: Examples of subgroup trajectories for SBP [left] and DBP [right] for studies with data collected in $\text{year}_i = 2000$ using a manual (yellow) or automated (purple) method of measurement. The dotted lines and data points show the summary measurements reported by the studies. The solid lines show the MLAD estimated subgroup trajectories.

Figure 2.3 illustrates the estimated SBP and DBP trajectories for manual and automated measurements where the study year $\text{year}_i = 2000$. The reported observations of BP for two studies, for which data were collected during this year, are also shown for comparison. The difference in BP according to method of measurement can potentially be attributed to findings from the quality assessment, which showed that most of the studies that used an automated method of measurement did not use a monitor that was approved for use in pregnancy (this will be discussed in more detail in Section 2.3.3). For the subgroup analysis with respect to parity, we found that multiparous women reported lower values of BP compared to nulliparous women (decrease by 4.6 and 4.7 mmHg for SBP and DBP, respectively). This finding is in line with findings from previous studies [79–81], but is potentially confounded by an increase in maternal age and BMI typically associated with subsequent pregnancies.

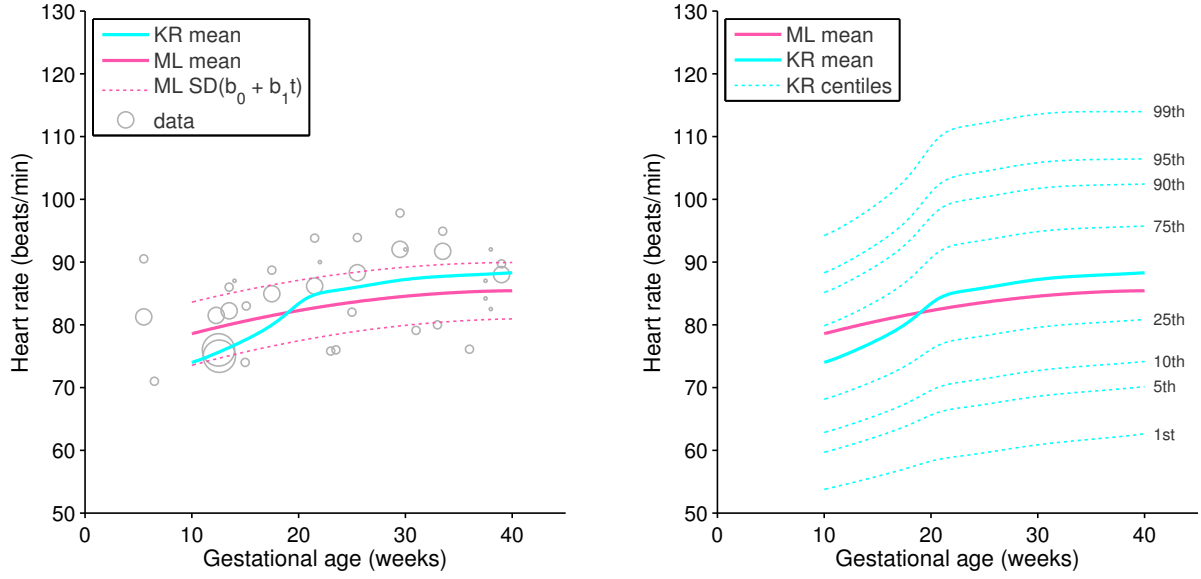
Previous studies have also reported that maternal age, BMI [82] and variation in the definition of the study population [83, 84] – particularly with respect to the inclusion of women who developed complications during their pregnancy – can affect the normal ranges of BP. However, due to incomplete reporting and the use of different summary measures of such population characteristics, we were not able to investigate the effect of these potential covariates.

2.3.2 Changes in heart rate

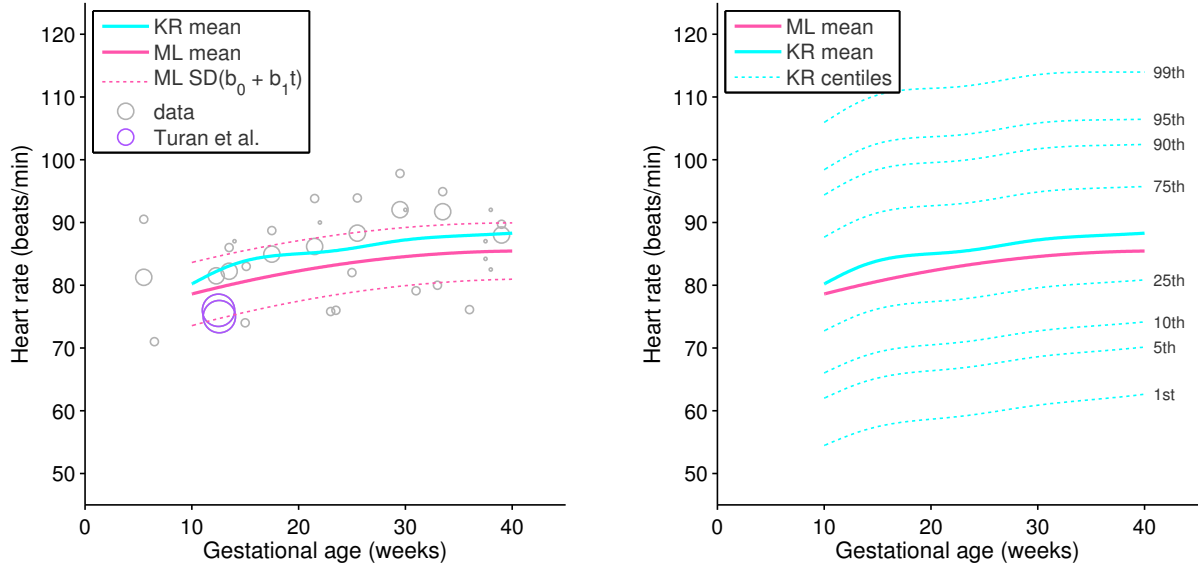
Of the studies that reported measurements of HR, 4 studies were longitudinal and 5 were cross-sectional. In total, 7,143 measurements of HR from 5,337 women were reported at 35 measurement occasions. The longitudinal studies reported measurements from 5 independent groups of women, with summary measurements reported at between 3 and 8 (median 4) measurement occasions per group. The cross-sectional studies all reported measurements at a single measurement occasion.

Figure 2.4a shows the extracted HR data and trajectories computed using MLAD and WVBKR. Both methods estimated an increase in mean HR from 10 weeks to 40 weeks of gestation. The mean estimated by WVBKR was approximately 2 bpm higher than the MLAD estimated mean. Figure 2.4a shows that the WVBKR estimate predicted a steep increase in HR before 20 weeks of gestation. This was caused by two data points from a single study by Turan *et al.* [85], shown in Figure 2.4b, which due to its large sample size contributed a weighting w_i that was more than 20 times greater than the median of all weightings $\mathbf{w}$. The effect of this is exaggerated by the limited number of data points. As this steep initial increase in HR is likely to be an artefact of the dataset and WVBKR method rather than a true reflection of physiological change, we adjusted the weighting to $w_i = \text{median}(\mathbf{w})$. This adjustment resulted in a closer resemblance between the mean trajectories predicted by the two methods (Figure 2.4b).

The methods of measurement of HR were reported as electrocardiography (3 studies, 6 measurement occasions), echocardiography (1 study, 2 measurement occasions), other



(a) Results from the meta-analysis of HR



(b) Adjusted HR results. Data points from study by Turan *et al.* highlighted in purple.

Figure 2.4: Results from the meta-analysis of HR. **[left]** The extracted HR data (grey circles, sizes reflect sample size) with mean fitted using MLAD (solid pink line), standard deviation of the within-study means (dotted pink lines) and mean fitted using WVBKR (solid light blue line). **[right]** Mean fitted using WVBKR (solid light blue line), associated centiles (dotted blue lines) and mean fitted using MLAD (solid pink line).

continuous waveforms (2 studies, 20 time points), reading from an automated BP monitor (1 study, 3 measurement occasions) or not specified (2 studies, 3 measurement occasions). Due to the limited data available, subgroup analyses were not deemed feasible.

2.3.3 Quality assessment

Overall quality assessment scores per study ranged from 5.6% to 84.2% (median 40.0%). Figure 2.5 summarises the score per assessment criterion. Of the quality assessment criteria, we identified three areas to be particularly relevant for the purpose of our analysis; these were sample selection (criteria 2 and 6), determination and reporting of gestational age (criteria 10 and 11) and suitability and reporting of measurement equipment and technique (criteria 13 and 19).

With respect to sample selection, 14 studies (38.9%) clearly stated that women were recruited either consecutively or at random (criterion 2) and 12 studies (33.3%) met our requirement for a low bias specification of inclusion and exclusion criteria (criterion 6). In order to score low bias for inclusion and exclusion criteria, the following criteria had to be met (number and percentage of studies with low-bias scores in brackets): exclude women with pre-existing cardiac disease or hypertension (25 studies, 69.4%); exclude women recruited on the basis of a diagnosis (27 studies, 75.0%); include women who subsequently developed abnormal fetal outcomes (28 studies, 77.8%) or pregnancy complications (25 studies, 69.7%) and; not present separate analysis for groups of women based on fetal or maternal outcomes (26 studies, 72.2%).

With respect to determination of gestational age (criterion 10), only 8 studies (22.2%) reported the use of ultrasound before 14 weeks of gestation; this is the gold standard method for pregnancy dating. Other studies used either ultrasound without specification of the gestational age at the point of measurement (6 studies, 16.7%), last menstrual period (3 studies, 8.3%) or did not report any method (19 studies, 52.8%). Only 4 studies (11.1%) reported gestational age in the exact form of days or weeks+days (criterion 11).

For the studies of BP, only 7 studies (21.9%) reported the use of suitable measurement

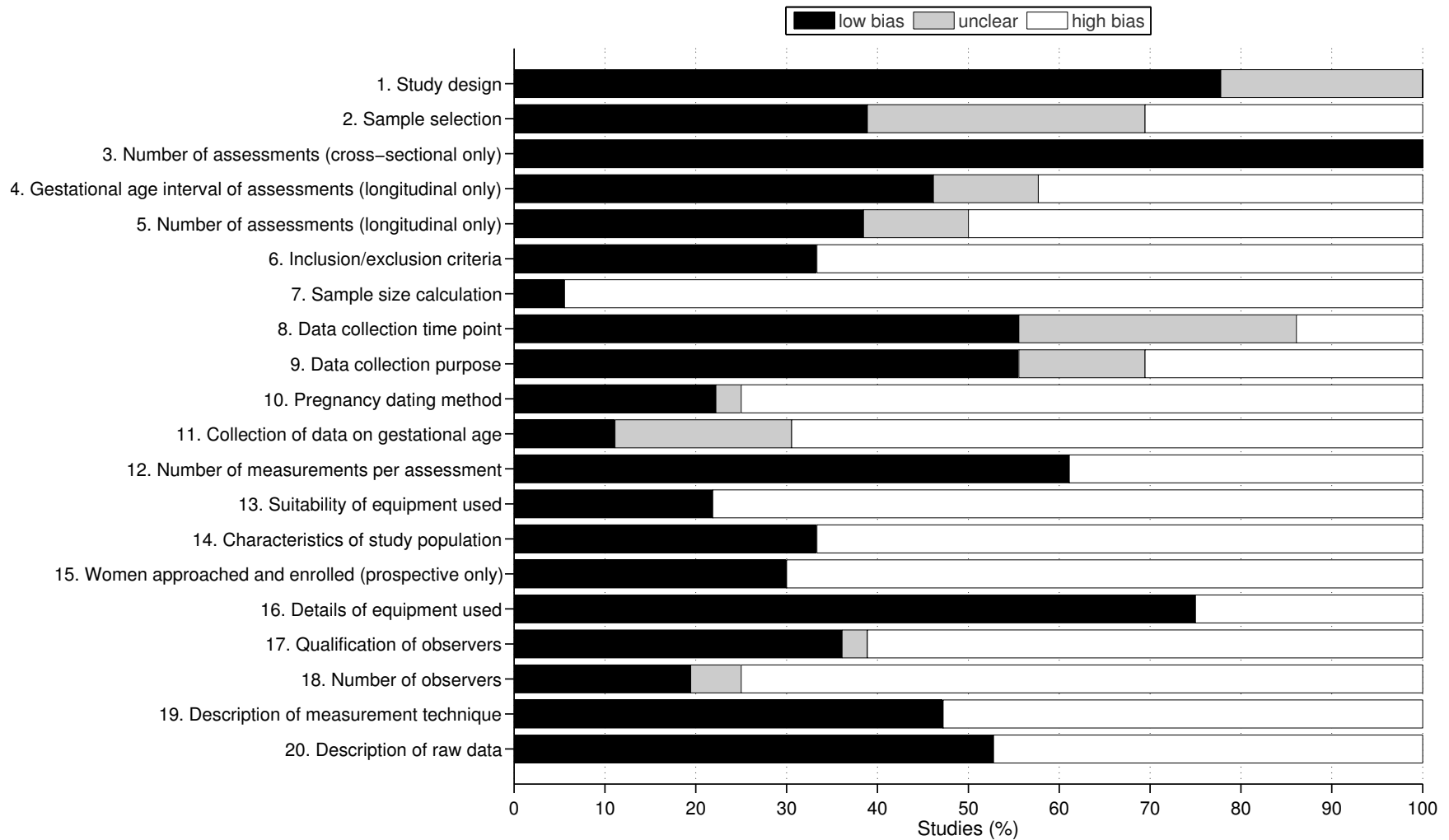


Figure 2.5: Scores from the quality assessment per question. The bars show the percentage of studies that were assigned a low bias (black), high bias (white) or unclear (white) score.

equipment (criterion 13), in either the form of a calibrated mercury sphygmomanometer (5 studies, 15.6%) or an automated BP monitor approved for pregnancy (2 studies, 6.3%). Only 17 studies (47.2%) provided an adequate description of the measurement technique used (criterion 19), including only 14 studies (43.8%) which reported use of an appropriately sized BP cuff. Monitors that are not approved for use in pregnancy have previously been reported to underestimate BP, particularly towards the end of the pregnancy [57]. This may explain our findings of lower BP from automated measurements in Section 2.3.1.

2.4 Discussion

Our synthesis of the existing evidence base of values of BP and HR associated with normal pregnancy shows that vital sign values change dynamically throughout the antepartum period. Their normal ranges thus depend on gestational age. However, the normal values reported in the literature vary greatly between studies. This makes it difficult to conclude what the exact expected gestation-specific values are.

Current guidelines for clinical assessment of vital signs in pregnancy do not account for the expected gestation-specific changes. To illustrate this, Figure 2.6 shows the centiles computed using WVBKR, superimposed on the current guidelines. The NICE guideline for hypertension in pregnancy recommends that treatment for high BP in pregnancy is initiated at 150/100 mmHg [10]. This threshold is well above the estimated 99th centile of both SBP and DBP. As a measurement of BP is the most common screening method to detect or predict hypertensive disorders in pregnancy, under-diagnosis of hypertension could result if the threshold is set too high. The MOET guideline states that a HR persistently > 100 bpm is to be defined as tachycardia (abnormally high HR) and interpreted as a *'red warning flag'* for serious infection [17]. In mid-pregnancy, a HR of 100 bpm corresponds to the estimated 90th centile. As serious infection can occur at any time during pregnancy, this threshold should vary with gestational age.

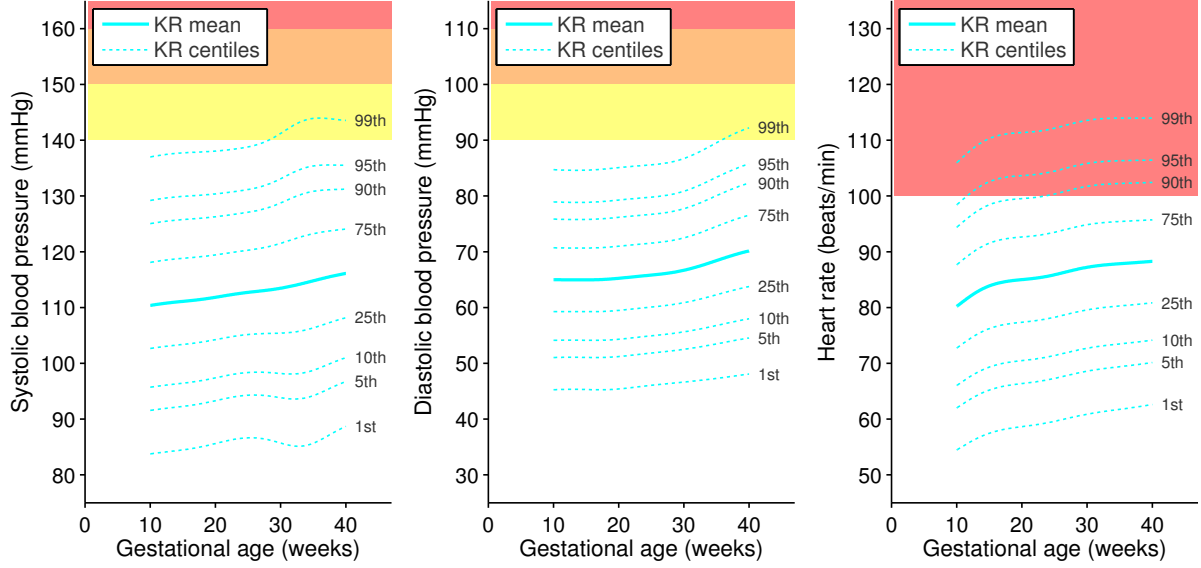


Figure 2.6: Computed centiles for SBP [left], DBP [middle] and HR [right] superimposed on current thresholds of normality. BP thresholds for diagnoses for mild hypertension, moderate hypertension requiring treatment and severe hypertension are shown in yellow, orange and red, respectively [10]. HR threshold for tachycardia is shown in red [17].

We have shown that different methods of measurement of BP are likely to result in considerably different BP values. This result is potentially confounded by the fact that many of the studies did not use a method of measurement that was suitable for use in pregnancy, as concluded from the quality assessment. The method of measurement should nevertheless be accounted for when designing cut-offs for early warning scores. Moreover, we found a progressive increase in BP from the late 1960's until today. This increase could be a consequence of increasing maternal age and BMI, as recently described by the Office of National Statistics [86]. An increase in maternal age could potentially also explain why we found a greater increase in DBP compared to SBP with study year (as a stiffening of arteries with age typically has a greater effect on DBP) [87]. However, as maternal age and BMI were poorly reported by the included studies, we were not able to correct for this in our analysis. Based on this, we advise others to exercise caution when interpreting the results of retrospective analyses of potentially outdated data sets, such as that recently presented by Macdonald-Wallis *et al.* [79], as such data are not necessarily suited to guide the design of normal ranges for use in current clinical practice [88].

We have not presented a sensitivity analysis and assessment of heterogeneity. For systematic review purposes, these analyses are considered to be integral parts of a quantitative data synthesis [67,89]. Details for how this would be done are presented in the published systematic review protocol, and will be included in the publication of the results of this systematic review.

The methods used for synthesising the data in the meta-analysis had several limitations. The main limitation of WVBKR was that all data were treated as independent. While this is unlikely to have affected the mean estimate, it may have affected the estimation of centiles [90]. If we hypothesise that women with abnormal physiological patterns are represented by the extreme centiles, this limitation may be critical for our ability to use these ranges to inform normal thresholds of vital signs. The main limitation of the MLAD model was that variations in the length of periods of gestational age were not accounted for. This may have limited our ability to accurately estimate the exact location of the observed dip in BP.

From the quality assessment, we found that most of the studies of BP and HR were of poor quality. In terms of sample size, more recent, higher-quality studies, such as Mahendru *et al.* [41], Van der Graaf *et al.* [91] and Grindheim *et al.* [46] achieved a relatively high quality score but reported measurements for small sample sizes and did not present centiles. Larger studies, such as MacDonald-Wallis *et al.* [79] reported data for larger sample sizes, but were retrospective in nature, which resulted in a low quality score. We also found large variations between studies in the definition of inclusion and exclusion criteria used to define *healthy* (or *low-risk*) pregnant women. This could potentially be explained by the differences in aims of the different studies, as their inclusion and exclusion criteria reflect whether the aim of each study was to present reference ranges or standard charts. The difference between these two concepts will be discussed in Chapter 5. Nevertheless, there does not appear to be any unanimous definition of what defines a *healthy* pregnant woman.

Two key quality aspects of concern were the methods used to determine gestational

age and the methods of measurement of BP. Both of these aspects directly affect the accuracy of the estimated centiles. The use of inaccurate methods of pregnancy dating and poor reporting of gestational age creates uncertainty about the shape of the vital sign trajectories (such as the location of the dip in BP). The use of BP measurement methods that are not suitable for use in pregnancy creates uncertainty about the values of the estimated mean and centiles of the vital signs at any given time point.

This systematic review is the first reported attempt at synthesising the available evidence base for changes in vital signs during pregnancy. The protocol for the systematic review has been published [69] and the main findings are currently being written up for publication. We were not able to identify a sufficient number of studies that reported on the other vital signs used in routine clinical assessment (SpO_2 , RR and Temp) or on the values of vital signs in the intrapartum and postpartum periods to include in a meta-analysis. Based on this, we conclude that there is a need for a study to establish normal trends of all of the six vital signs across the three stages of pregnancy. The 4P study, which is introduced in the following chapter, aims to establish this knowledge.

Chapter 3

The 4P study and collection of vital sign data

The systematic review presented in the previous chapter illustrated the limited availability of high-quality vital sign data required to describe normal trends in maternal vital sign physiology throughout pregnancy. From this, we conclude that there is an urgent need for a study to establish an evidence-based definition of normality in pregnancy.

This chapter introduces the reader to the 4P (Prediction of Physiology Patterns in Pregnancy) study. The 4P study aims to collect large numbers of vital sign measurements of HR, SBP, DBP, SpO₂, Temp and RR from 1,000 women longitudinally in pregnancy, from the early antepartum period until 2 weeks postpartum. These data will be used to design reference ranges for changes in vital signs with gestational age. The first section of this chapter outlines the design of the 4P study. In the second section, we describe the 4P digital health system that has been designed to enable the collection of large numbers of vital sign measurements. In the final sections, we provide a definition of the *low-risk* population whose vital sign data will be further investigated, summarise the data and present the quality checks performed to validate the data; this is an essential pre-requisite for determining accurate vital sign reference ranges.

3.1 Study protocol

The 4P study was designed in two phases. The first phase of the study was conducted in the John Radcliffe Hospital, Oxford University Hospitals (OUH) NHS Trust, from September 2012 to December 2015¹. During this phase, the women recruited to the 4P study were participants from the INTERBIO–21st study, for which they were required to attend regular ultrasound scans for the duration of their pregnancy [92]. The second phase of the 4P study commenced in July 2015 and is ongoing. It includes three data collection centres across the UK: OUH NHS Trust (Oxford), Newcastle Hospitals NHS Foundation Trust (Newcastle) and Guy’s and St Thomas’s Hospital NHS Trust (London)². As the investigation presented in this thesis will use the data collected during the first phase of the 4P study, the discussion will focus on data collection during this phase only.

3.1.1 Study participants

Women who attended the hospital for antenatal care from < 14 weeks of gestation were eligible for recruitment irrespective of their risk profile for adverse pregnancy or neonatal outcomes³. The inclusion criteria, presented in Table 3.1, were defined to ensure that the recruited women represented a typical sample of the Oxfordshire population of pregnant women in terms of age, ethnicity, general health and pregnancy complications (or lack thereof). The number of women recruited to the study at any one time was limited by the availability of the research midwives for data collection. Hence every month, recruitment was paused when a pre-defined monthly recruitment target of 25 women was reached.

Determination of gestational age: Accurate dating of pregnancy is imperative to ascertain that the timings of physiological changes are correctly measured. All participants of the 4P study had their gestational age determined at study entry using an ultrasound measurement of crown-rump length, performed between 9⁺⁰ and 14⁺⁰ weeks^{+days}.

¹Oxford Research Ethics Committee (08/H0606/139)

²National Research Ethics Service committee South East Coast - Brighton and Sussex (14/LO/1312)

³Delivery normally occurs at a gestational age of 40 weeks.

Table 3.1: Inclusion criteria for the 4P study

Inclusion criteria
Ultrasound dating scan < 14 weeks
BMI < 35 kg/m ²
Must be over 18 years old
Naturally conceived pregnancy
Singleton pregnancy
Must intend to deliver at the John Radcliffe Hospital, Oxford University Hospitals NHS Trust

With an estimated accuracy of 3 to 8 days, this is considered to be the gold standard for pregnancy dating [93]. The gestational age estimate was cross-checked with the date of the last menstrual period. In the case of disagreement, a second measurement was made.

3.1.2 Data collection stages

Data were collected throughout all three stages of pregnancy; antepartum, intrapartum and postpartum. Figure 3.1 summarises the approximate time points of data collection during each of these stages.

Antenatal stage: After the initial dating scan at 9 – 14 weeks, each woman was scheduled to attend 6 further visits at 5-weekly intervals, corresponding to 14 – 18, 19 – 23, 24 – 28, 29 – 33, 34 – 38 and 39 – 42 weeks of gestation. During each visit, a trained research midwife recorded the woman’s vital signs (HR, SpO₂, Temp, SBP, DBP and RR) using the 4P digital health system (to be described in Section 3.2). The time points varied around the nominal 5-weekly time points based on the availability of the women and of the midwife for their appointments.

Intrapartum stage: During labour, the woman’s vital signs were monitored at regular intervals and recorded on a range of standardised hospital charts, according to standard hospital procedures. These data were subsequently double-entered into a database by two research midwives through the 4P secure web interface (see Section 3.2.3).

Stages of data collection	Antepartum period (weeks)							Postpartum period (days)							
	9 - 14	14 - 18	19 - 23	24 - 28	29 - 33	34 - 38	39 - 42	0	2	4	6	8	10	12	14
Antepartum	.	.	.	.	.	.	.								
Intrapartum															
Postpartum															
Trimesters	1	2			3										

Figure 3.1: Timeline for data collection by the research midwives for the 4P study. The dots represent approximate time points of data collection appointments.

Postnatal stage: For up to two weeks after delivery, each woman was asked to make daily measurements of her vital signs (HR, SpO₂, SBP, DBP and Temp) at home. Women who agreed to monitor at home received a set of equipment similar to that used in the clinic. Correct use of the equipment was taught during a 10-minute training session with a research midwife. In addition, a research midwife visited the woman at home one and two weeks after delivery to obtain vital sign recordings on those two days. These recordings were used to validate the measurements made by the woman on the other days and to obtain measurements of RR on those two days.

3.2 System for collection of vital sign data

To enable a high-quality data collection process that included validation, review, security and confidentiality of the data, the 4P digital health system was designed as part of the work carried out for this thesis. The system, depicted in Figure 3.2, was developed in close collaboration with midwives and doctors from the John Radcliffe Hospital (OUH NHS Trust) according to the system requirements presented in Table 3.2. The code for the software applications was implemented by post-doctoral Research Associates in the Biomedical signal processing and m-health group (Department of Engineering Science, University of Oxford).

From the system requirements, written specifications were produced and working prototypes were built and tested with focus groups in the clinical and home settings.



Figure 3.2: The 4P digital health system used for data collection by midwives and women during the ante- and postnatal phases, respectively.

Table 3.2: System requirements for collection of vital sign data in the 4P study

	Requirements
Req-1	Electronic collection of all six vital signs (HR, SpO ₂ , SBP, DBP, Temp, RR)
Req-2	Devices and sensors chosen from those regularly used in standard hospital settings
Req-3	Suitable for data collection by healthcare professionals in the clinic and during home visits
Req-4	Suitable for data collection by pregnant women (non-expert users) at home
Req-5	Capable of real-time transmission and display of home monitoring data
Req-6	Full compliance with requirements for patient safety and security of data storage

The feedback obtained from healthcare professionals and women in the focus groups enabled the development of the system according to the end-users' needs.

3.2.1 Hardware

The 4P digital health system consisted of commercial medical devices regularly used in the hospital setting (Req-1 and Req-2). These peripheral devices were interfaced to a computer tablet (Samsung Galaxy Tab2, Android OS 4.0) running a custom-designed Android application (described in Section 3.2.2). Validation ranges were put in place to ensure that data input into the Android tablet using the on-screen keyboard were within a physiologically meaningful range.

Heart rate and oxygen saturation: HR and SpO₂ were measured with a Bluetooth-enabled pulse oximeter (Nonin 3150 WristOx2) using a finger probe. The values of HR and SpO₂, together with the PPG waveform (Section 1.3.2), were transferred to the Android tablet with a sampling frequency of 75 Hz. Any error flags and pulse quality metrics output by the pulse oximeter were also recorded. For each clinic visit, a single, robust value for both HR and SpO₂ was calculated from a section of data, as described in Section 3.4.1.

Systolic and diastolic blood pressure: SBP and DBP were collected as discrete measurements and typed into the Android tablet. The acceptable ranges of SBP and DBP were set to 80 – 200 and 40 – 110 mmHg, respectively. We were not able to find a Bluetooth-enabled BP monitor that was certified for use in pregnancy. The BP monitor used in the study (Microlife 3BTO-A(2)) has previously been certified specifically for use during pregnancy according to the guidelines of the British Hypertension Society [94]. The appropriate cuff size was selected based on the participant's arm circumference, as measured by the study midwife.

Temperature: Temp was measured using a tympanic infrared thermometer (Genius 2, Covidien), used in the majority of wards in the OUH NHS Trust. The readings from

the thermometer were typed into the Android tablet. The range of accepted values of Temp was set to $34.5 - 41^{\circ}\text{C}$. In the clinic, duplicate measurements were obtained with a second infrared thermometer (Fora IR20b, ForaCare), with transfer of the data to the Android tablet via Bluetooth. Duplicate measurements were deemed necessary because the Bluetooth-enabled thermometer was not certified for use in hospitals⁴.

Respiratory rate: The reference RR was collected as a discrete measurement. The midwife counted chest-wall movements for 15 seconds and multiplied this count by four to obtain an estimate of the RR in rpm. The count was typed into the Android tablet. In parallel, electronic measurements of RR were obtained from detection of chest-wall movement using the in-built accelerometer and gyroscope of a smartphone (Samsung Galaxy Nexus, Android OS) and from the PPG. These methods, and the subsequent application of signal processing techniques to extract an electronic measurement of RR, will be discussed in more detail in Chapter 4.

3.2.2 Tablet computer software

The tablet computer software was designed in two different versions, to comply with the system requirements (Table 3.2); one application optimised for use by healthcare professionals in the antenatal hospital clinic (Req-3) and one for use by women at home during the postnatal period (Req-4). The applications were written in Java for Android.

The clinic application was designed for use by midwives (expert users) in the antenatal clinic. It allowed vital signs to be recorded in any order, so that the midwife was able to fit data collection between her other tasks (for example, scanning the baby). The user interface, depicted in Figure 3.3a, was kept simple with minimal amounts of text and buttons to press. The pulse oximeter finger probe was attached at the beginning of the clinic visit and used for data acquisition throughout the clinic visit. Both the PPG waveform and the error flags output by the pulse oximeter were displayed to indicate improper sensor positioning and/or motion so that action could be taken by the midwife

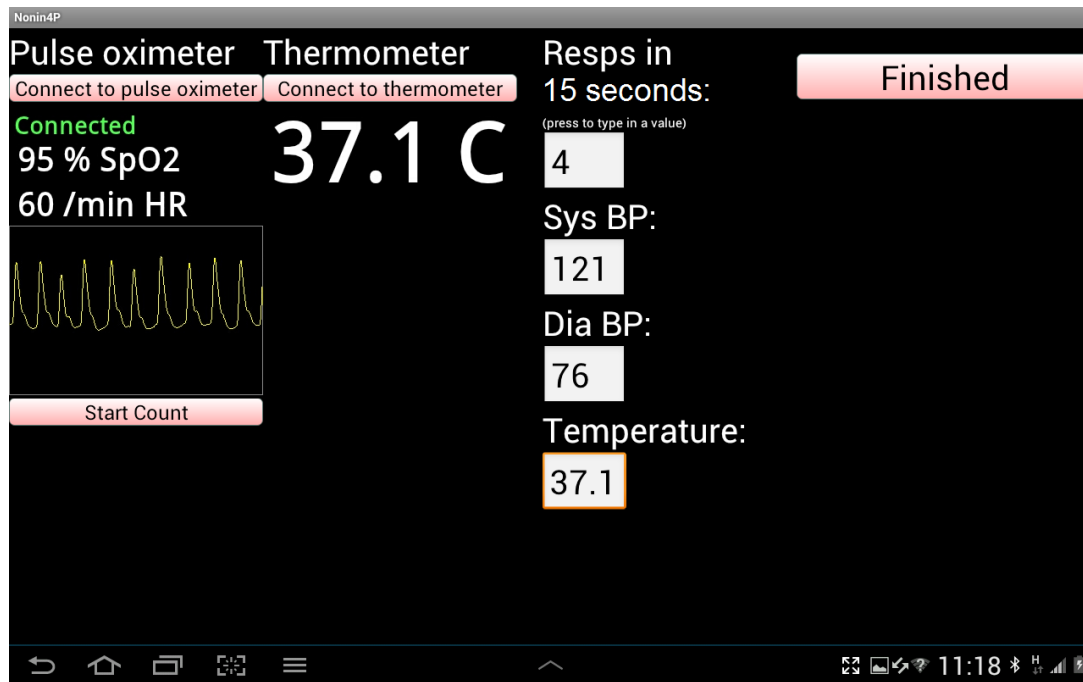
⁴The Genius 2 thermometer was only in use from 1st March 2012.

to correct any such problems, as recommended by current guidelines [43]. This was important because motion or improper positioning of the finger probe could corrupt the PPG waveform.

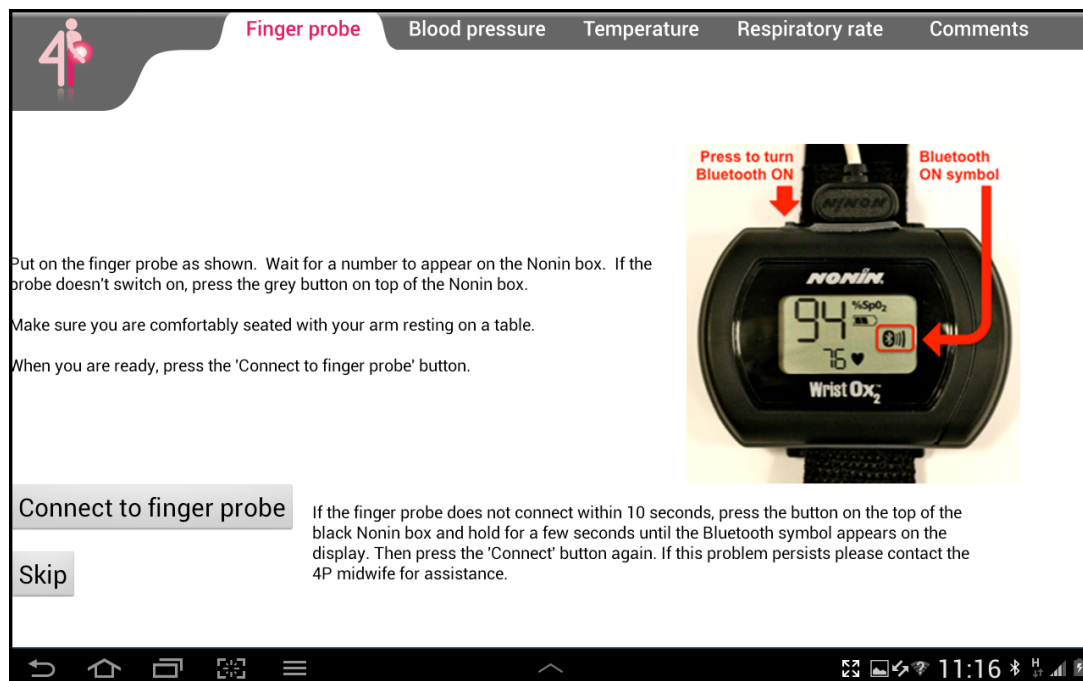
The second application was designed for home monitoring. It included step-by-step instructions in the form of schematics and videos to guide the pregnant woman (non-expert user) through data collection. An example screenshot is shown in Figure 3.3b. The pulse oximeter recording duration was fixed to one minute and a status bar displayed the progress and remaining time of the recording. Any error flags generated by the pulse oximeter were displayed in the form of messages to report improper sensor position and/or motion artefacts. The software allowed women to submit free-text comments together with their recorded datasets. No RR measurements were made by women during home monitoring. During the weekly visits to the participant's home, RR measurements were recorded manually by the midwives (and in parallel by the smartphone, as discussed in Chapter 4).

3.2.3 Web interface for follow-up of home monitoring

The research midwives were able to review the vital sign data from the women who measured their vital signs at home and plan their weekly visits to the participants' homes using the 4P secure web interface, through which it was possible to view completed datasets from the database in real-time (Req-5). To allow for efficient monitoring of several women simultaneously, the midwives were able to switch between a summary screen that displayed the data from all participants and individual screens that presented detailed records of each participant's recordings. This allowed the midwives to identify and deal with issues that the women would experience, such as problems with the equipment or questions about their measurements. Although this was initially only intended for use during the postpartum period, it was found to be so useful that data from both the ante- and postpartum periods were displayed. Figures 3.4a and 3.4b show screenshots of the summary screen and an example patient profile, respectively.

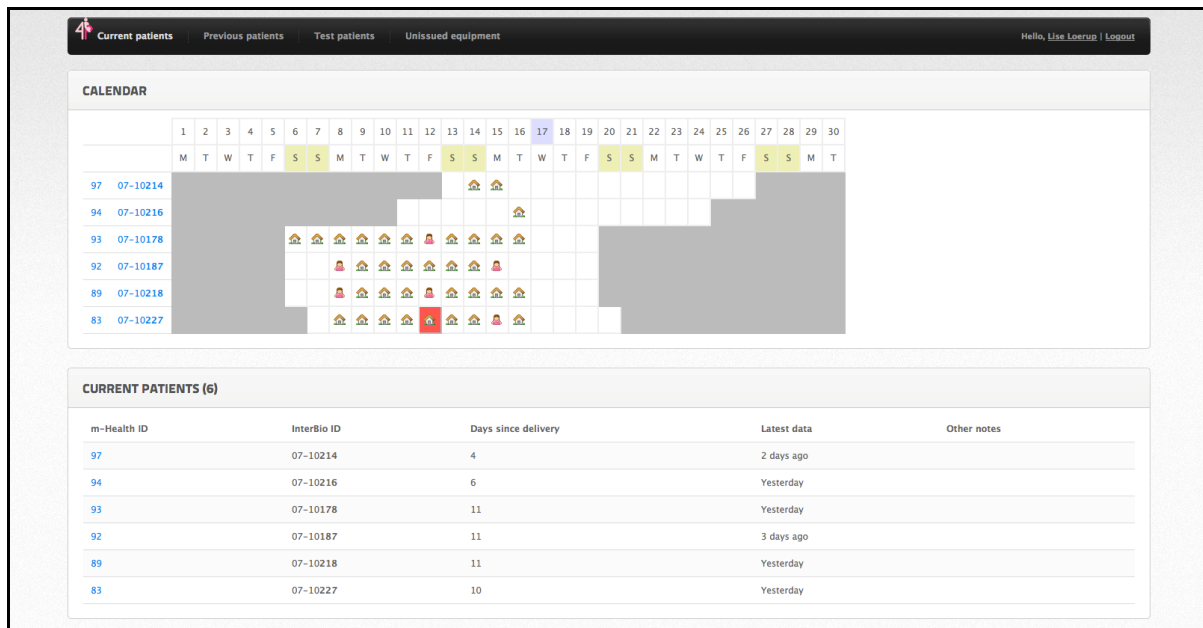


(a) The main interface of the clinic application displays the PPG waveform and instantaneous values of HR and SpO₂ as output by the pulse oximeter. Temp is transferred from the tympanic thermometer via Bluetooth. The count of breaths is input using the onscreen keyboard.

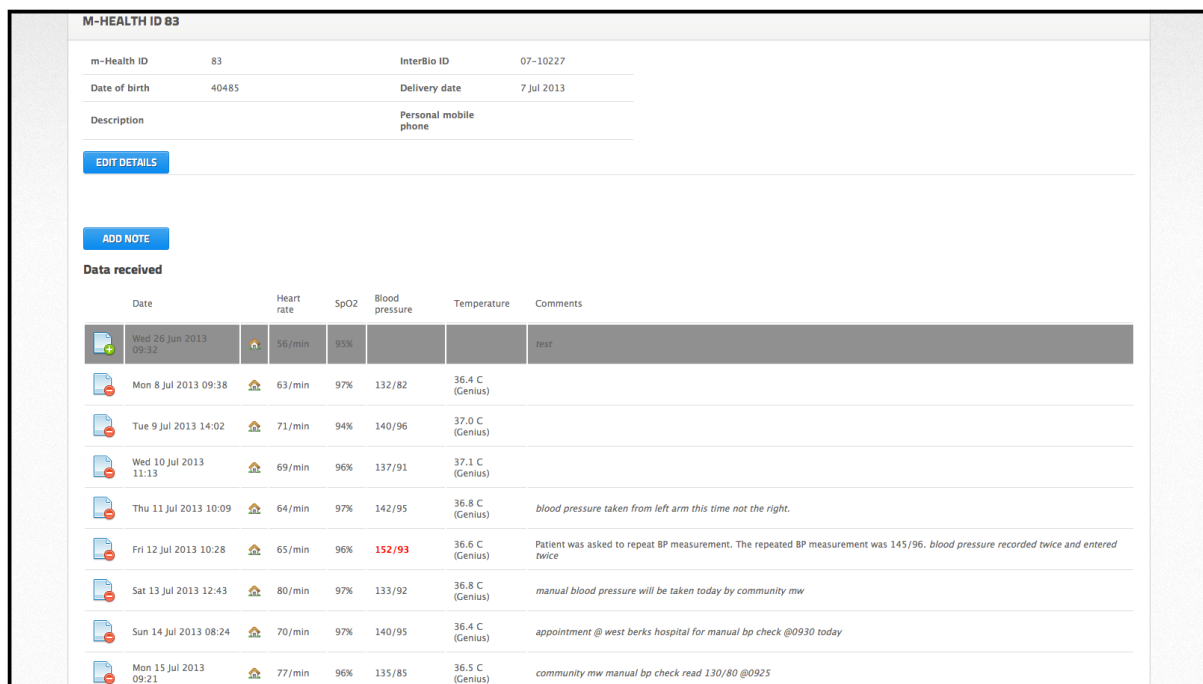


(b) The home monitoring application which includes pictures and instructions to guide the women through all the steps of data collection. This screen shows the user how to measure her BP. The readings are entered into the text boxes using the on-screen keyboard.

Figure 3.3: Screen shots of the 4P tablet computer applications



(a) The web interface summary screen shows the number and timings of datasets submitted from those women monitoring their vital signs at home. The different symbols in the calendar (top of page) indicate that datasets have been submitted from a woman monitoring at home (house symbol) or from a midwife during a home visit (female symbol). The red square indicates that a BP reading that requires follow-up has been submitted.



(b) Through the web interface, the midwives can view each participant's profile. This contains the time, date and values of all their home monitoring data (HR, SpO₂, Temp and BP), as well as any free text comments submitted by the participants. A high BP reading is highlighted in red. Text in Roman font is automatically generated by the system; text in italics are typed in by the participant on her tablet.

Figure 3.4: Screen shots of the 4P web interface

3.2.4 Other variables

Demographic information (maternal age, weight, height), pregnancy history (parity), pregnancy complications and medical history were collected from the maternity notes and electronic hospital records. These data were stored together with the vital sign data in the 4P database.

3.2.5 Data transmission and storage

Upon completion of a recording session, the data stored in the Android tablet and smartphone were automatically transmitted, either via a secure Wi-Fi network (in the hospital) or via the 3G network (from the participant's home), to the 4P study server for storage in the 4P database. The 4P database and web interface were hosted on a secure server that resided behind the NHS firewall (Req-6). The data were anonymised by assignment of an unique identifier (m-health ID), which was stored together with the date and time of the recording and the vital sign data. No patient identifiable data were collected or stored in the online database. All of the application settings (m-health ID and passcode) were hidden from the user and were password-protected. A paper list linking the names of the women and the m-health ID was securely stored in a locked location only accessible to selected members of the data collection team. For the purpose of retrospective analysis, the data were subsequently extracted from the 4P database and matched using the m-health ID and the timestamp of the recording.

3.3 Description of the data used for this thesis

For completeness, and to explain the motivation for the design of a digital health system for data collection, we have described the process of collection of vital sign data for all three stages of pregnancy. However, for the remainder of the investigation presented in this thesis, we will only use the data collected during the antepartum period.

3.3.1 Study population

Between 1st August 2012 and 23rd November 2015, 518 women completed the antepartum monitoring phase of the 4P study. For the investigations presented in this thesis, we will only include 439 women who at the time of recruitment were *healthy* or *low-risk*, i.e. did not have a medical condition or were not taking medications that may have affected her vital sign patterns. The criteria used to define low-risk women are presented in Appendix A.1.2. The baseline information for the women recruited to the 4P study is presented in Table 3.3, together with the possible reasons for exclusion from the low-risk group.

Table 3.3: Demographic information of women in the 4P study

	Healthy women <i>n=439</i>	Excluded women <i>n=79</i>
Baseline information, <i>median (IQR)</i>		
Age at recruitment (years)	31.6 (5.9)	30.5 (7.8)
Weight at recruitment (kg)	64.2 (16.2)	66.5 (19.2)
BMI at recruitment (kg/m ²)	23.6 (5.4)	24.9 (7.6)
Gestational age at recruitment (days)	87 (14)	87 (9)
Parity, <i>n (%)</i>		
Parous	251 (57.3)	49 (62.0)
Nulliparous	188 (42.8)	30 (38.0)
Chronic condition at recruitment, <i>n (%)</i>		
Cardiac disease	–	21 (26.5)
Hypertension	–	14 (17.7)
Malignancy (cancer)	–	3 (3.8)
Chronic respiratory disease	–	40 (50.6)
Renal disease	–	12 (15.2)

Table 3.4 summarises the delivery and pregnancy outcomes for women included in the analysis. A subgroup of these women met the stricter inclusion and exclusion criteria for the INTERGROWTH–21st project and were consequently defined to be *very healthy* [95]. In brief, these women had fewer risk factors (such as normal BMI and non-smoking) and did not develop any complication or condition during the pregnancy that required hospitalisation (such as severe pre-eclampsia). In Chapter 5 we will investigate whether the vital sign patterns of these women differed from those of the general analysis population.

Table 3.4: Pregnancy outcomes for women in the 4P study

	Analysis population <i>n=439</i>	IG-like women[†] <i>n=47</i>
Delivery information, <i>median (IQR)</i>		
Gestational age at delivery (days)	280 (15)	282 (13)
Birth weight (g)	3368 (675)	3508 (631)
Birth outcome, <i>n (%)</i>		
Live birth	434 (98.9)	47 (100)
Miscarriage / still birth	5 (1.1)	—
Pre-term delivery	33 (7.5)	1 (2.1)
Pregnancy complications, <i>n (%)</i>		
Gestational diabetes	16 (3.6)	1 (2.1)
Pregnancy-induced hypertension	25 (5.7)	3 (6.4)
Pre-eclampsia	5 (1.1)	1 (2.1)
Severe pre-eclampsia	1 (0.2)	—
Conditions during pregnancy, <i>n (%)</i>		
Cardiac disease	2 (0.5)	—
Diabetes	2 (0.5)	—
Endocrinological condition (incl. diabetes)	22 (5.0)	1 (2.1)
Respiratory disease	6 (1.3)	1 (2.1)
Infection (requiring antibiotics/antivirals)	113 (25.7)	10 (21.3)
Infection (other)	47 (10.1)	1 (2.1)
[†] Women who met the inclusion and exclusion criteria for the INTERGROWTH–21 st study [95]. These women (IG-like) represent a subset of the analysis population.		

3.3.2 Vital sign measurements

During the antepartum period, a total of 2,645 sets of vital sign measurements were collected, with a median (IQR) of 6 (1) datasets per woman. Figure 3.5 show the number of datasets collected per woman across all study participants.

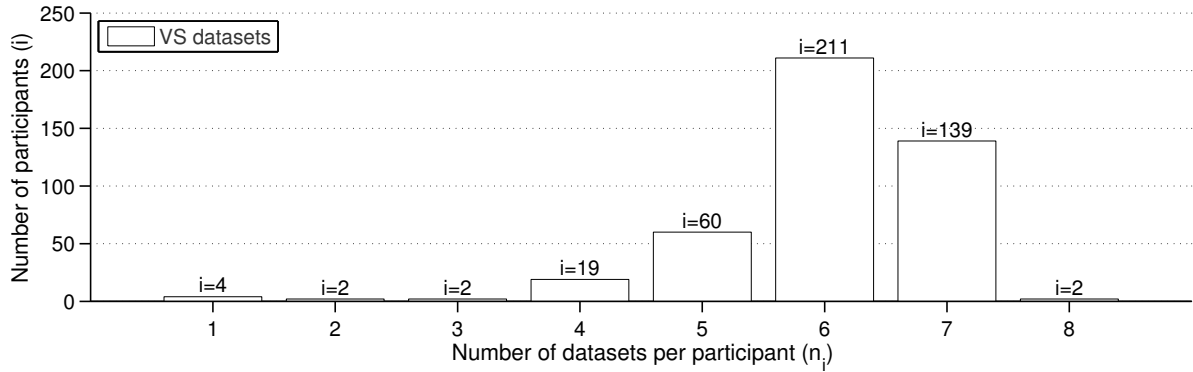


Figure 3.5: Number of datasets recorded per participant

Figure 3.6 shows the number of recorded sets of vital sign measurements per week of gestation. The peaks in the distribution roughly correspond to the pre-specified 5-weekly intervals, with most women having their first visit to the hospital between 12-13 weeks of gestation, as per the usual programme for antepartum care in the JR Hospital.

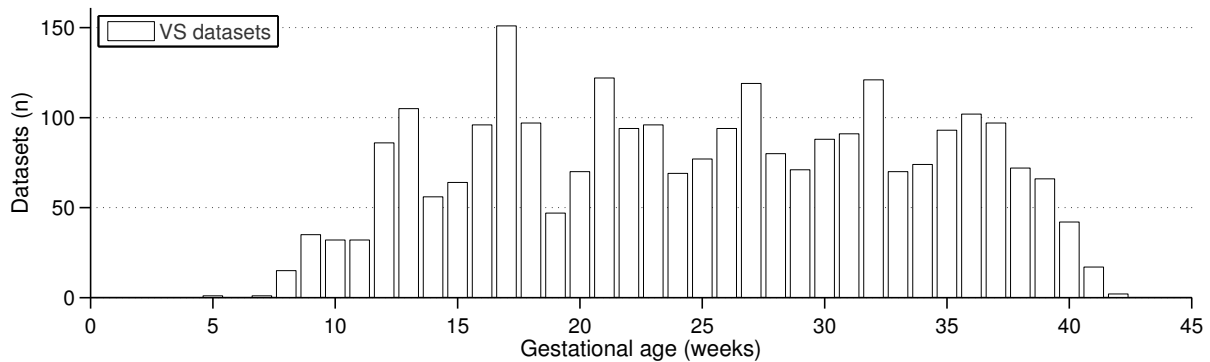


Figure 3.6: Number of datasets recorded per week of gestation

Table 3.5 summarises the number of measurements obtained for each vital sign. Although the majority of observation sets contained measurements of all vital signs, some measurements were occasionally missing. There were several reasons for missing measurements. Occasionally,

Table 3.5: Summary of collected vital sign data

	Datasets <i>n = 2645</i>
Vital sign measurements, <i>n</i> (% of total)	
Number of complete datasets	2001 (75.7)
Heart rate and SpO ₂	2632 (99.5)
Systolic blood pressure	2626 (99.3)
Diastolic blood pressure	2626 (99.3)
Temperature (Bluetooth)	2560 (96.8)
Temperature (manual)	2150 (81.3) [†]
Respiratory rate (manual)	2499 (94.5)
[†] Thermometer only used after 1 st March 2013.	

the research midwives reported difficulties with the Bluetooth connectivity of the pulse oximeter or Bluetooth thermometer, or that some manually-recorded measurements had simply not been entered onto the system. At other times, some measurements were not made at all.

3.4 Validation of vital sign measurements

The design of gestation-specific normograms of vital signs requires a single, robust value for each vital sign. Although some validation procedures, such as validation ranges based on the physiologically plausible ranges of each vital sign, were incorporated in the data collection system, erroneous or outlier measurements did occur. Whereas outlier data points resulting from abnormal physiology should be identified and included in the analysis, outlier data points that are caused by erroneous measurements (e.g. due to user error or problems with the equipment) should be removed. The removal of outlier values is primarily important for the estimation of extreme centiles. In the following sections, we present the strategy for data validation that was designed in order to ensure that a single, robust value for each vital sign measurement was obtained.

3.4.1 Heart rate and SpO₂

The pulse oximeter automatically outputs values of HR and SpO₂ computed as mean values over the past 4 beats from the PPG waveform⁵. The data are output at a frequency of 75 Hz, with the values of HR and SpO₂ updated at a rate of 3 Hz [53]. Table 3.6 summarises the characteristics of the outputs from the pulse oximeter.

Table 3.6: Characteristics of the pulse oximeter outputs

Output values	Update frequency	Description
PPG waveform	75 Hz	Infrared signal
HR	3 Hz	Mean over last 4 beats
SpO ₂	3 Hz	Mean over last 4 beats

Because the pulse oximeter finger probe was used for the duration of the data collection sessions, the recording duration and quality varied across the recordings. In general, several factors may cause artefacts in the PPG waveform that could undermine measurement quality. Examples include incorrect probe positioning, which can lead to a noisy signal, or motion artefacts or physical barriers to the penetration of light (such as nail varnish), which can cause noise or even a complete loss of signal to occur [43]. Poor-quality data may lead to inaccurate values of HR and SpO₂ [53]. To compute a single robust value of HR and SpO₂, we designed a three-step processing strategy to reject poor quality data:

1. reject sections of invalid data (automatically output by the pulse oximeter as SpO₂=125% or HR=511 bpm);
2. reject sections of data that were deemed to be of poor quality; and
3. reject entire recordings if the proportion of good quality data remaining after steps 1 and 2 was less than 20%.

At each step, rejected data were labelled by means of a signal quality index (SQI). Invalid data (step 1) that result from a *persistently* poor-quality signal were output as invalid values for HR and SpO₂ by the pulse oximeter. Such data were labelled as valid (SQI(valid)= 1) or invalid (SQI(valid)= 0) data. To identify and reject *intermittent*

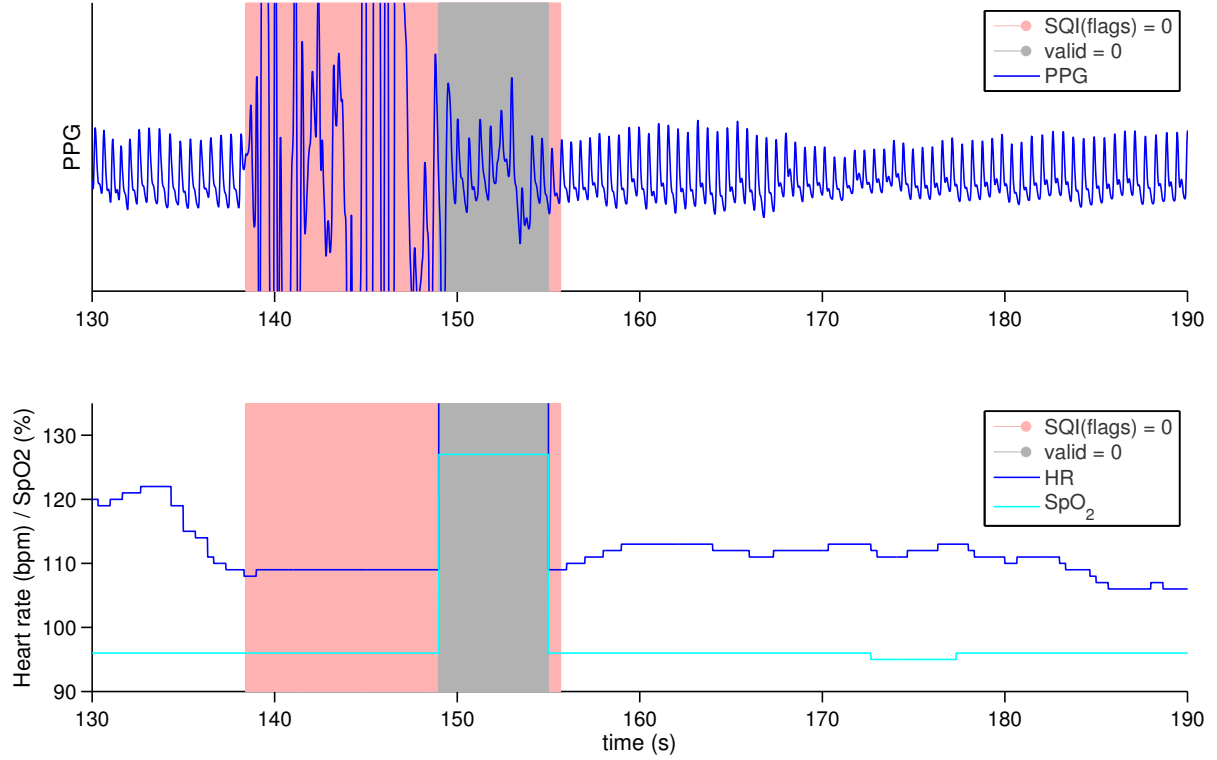
⁵The pulse oximeter and PPG waveform were described in Section 1.3.2.

segments of poor quality data (step 2) we designed a signal quality index (SQI(flags)) based on the alerts output by the pulse oximeter. This was used to classify sections of good (SQI(flags)= 1) and poor (SQI(flags)= 0) quality data. This SQI is described in greater detail in Appendix C.2.1. Finally, if the recording did not contain at least $\geq 20\%$ of good quality data (an empirically determined threshold), the recording was deemed to be unreliable and it was rejected (step 3). For the remaining recordings, the final value of HR and SpO₂ was computed as the mean of the good-quality segments of the recording.

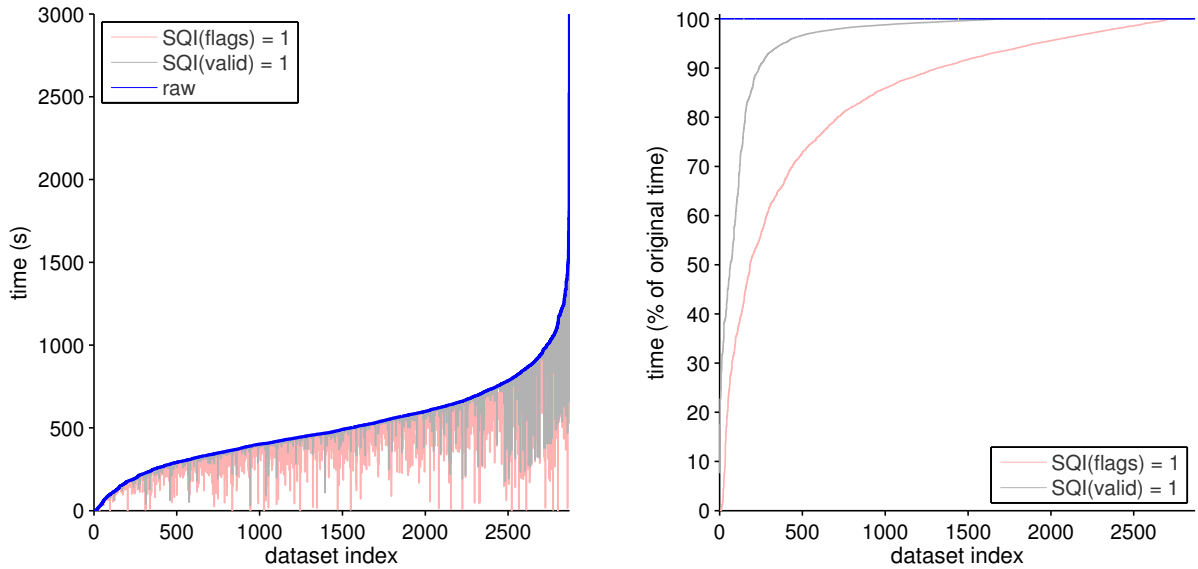
Figure 3.7a shows an example PPG waveform and the associated values of HR and SpO₂ output by the pulse oximeter. The sections of data rejected by SQI(valid)=0 and SQI(flags)=0 are highlighted. Figure 3.7b shows the total monitoring time for each recording using the pulse oximeter and effective monitoring time, defined as the amount of time that values of HR and SpO₂ were classified with SQI(valid)= 1 AND SQI(flags)= 1. As expected, more data were rejected using SQI(flags), particularly for the longer recording sessions.

Figure 3.8 illustrates the effect of each step of the processing strategy for all the recorded values of HR and SpO₂. Figure 3.8a shows the values of HR and SpO₂ computed after rejection of invalid data (step 1). Values of SpO₂ < 90% are indicative of hypoxia and rarely observed in healthy individuals [27]. Some of the very low (< 90%) measurements of SpO₂ were therefore hypothesised to result from poor-quality data. Figure 3.8b shows the values of HR and SpO₂ computed after rejection of poor-quality data (step 2). We see that while the number of poor quality SpO₂ readings has been reduced, a few measurements of SpO₂ < 90% remain. Figure 3.8c highlights the recordings of HR and SpO₂ that were rejected due to an insufficient amount of good quality data (step 3). This eliminates most of the poor quality recordings of SpO₂. The remaining recordings were considered to be robust and of sufficiently good quality for inclusion in the analysis. For reference, the recordings of SpO₂ < 90% that were not excluded from the analysis using this strategy are shown in Appendix C.2.2.

This three-step strategy, based on signal quality of the PPG, appeared to be effective



(a) Example of the signal quality indices applied to a 60-second segment of a pulse oximeter recording. **[top]** The PPG waveform with sections of artefactual data detected by the signal quality indices and classified as invalid (grey) and SQI= 0 (pink) highlighted. **[bottom]** The corresponding output values of HR (blue) and SpO₂ (light blue). Note the step change in the output values that occur when the artefacts in the PPG persist.



(b) The impact of signal quality indices on the time of monitoring with the pulse oximeter for each dataset. **[left]** Representation of the original (raw) monitoring time per dataset (shown in blue and sorted in ascending order), with the monitoring times after exclusion of invalid data (grey) and sections with SQI= 0 (pink). **[right]** The proportion of data excluded per dataset after application of the signal quality indices (each sorted in ascending order).

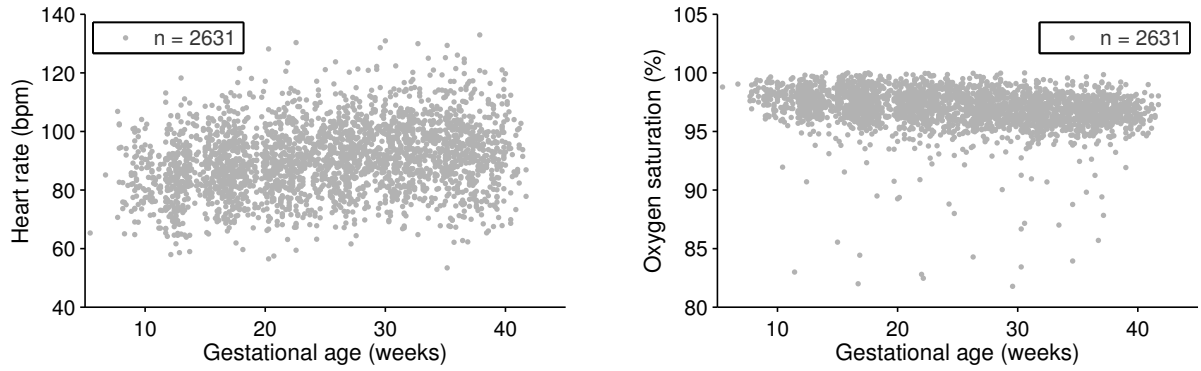
Figure 3.7: The effect of signal quality indices applied to the pulse oximeter data

for rejecting recordings of SpO₂ that produced erroneous SpO₂ values. Recordings of SpO₂ < 90% are not expected in healthy individuals and visual inspection confirmed that the eliminated recordings were corrupted by artefacts. For HR, on the other hand, the range of physiologically plausible values is much larger. Given that HR is estimated from prominent peaks in the PPG, it is also more likely to be accurately computed, even from a signal of poor quality. It was therefore decided to use the strategy presented here only for rejection of SpO₂ measurements and not for HR measurements.

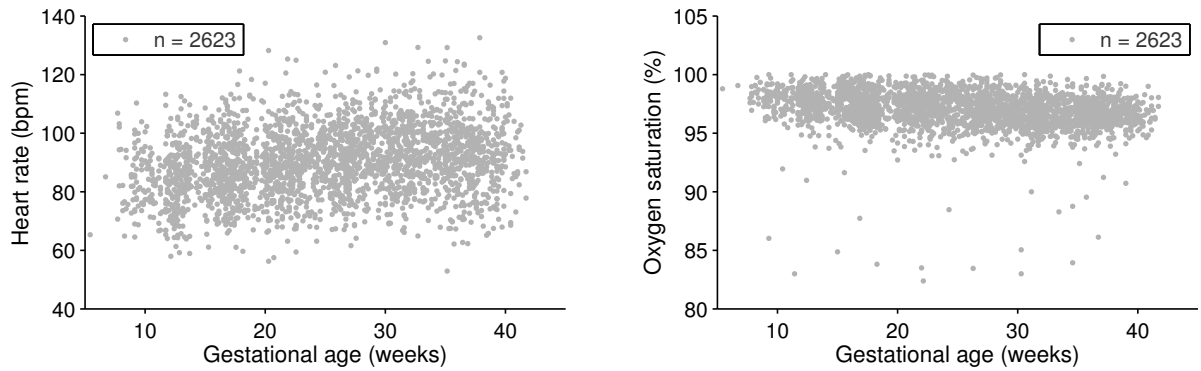
A statistical comparison of the distribution of SpO₂ computed before and after rejection of poor-quality data showed that the medians and inter-quartile ranges of the distribution at each stage were not statistically different. The duration of these recordings and the overall mean values of HR and SpO₂ are summarised in Table 3.7.

Table 3.7: Duration and quality of pulse oximeter data

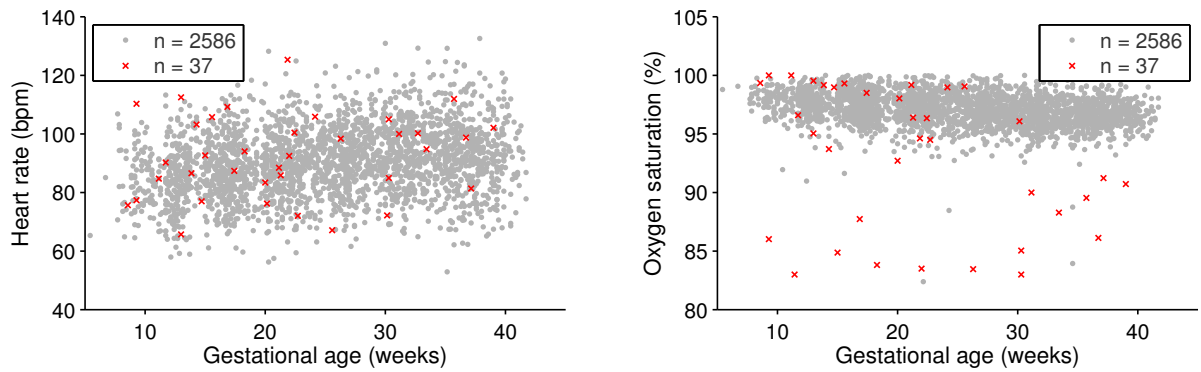
		Validation criteria		
		valid output $n = 2632$	SQI(flags) $n = 2631$	SQI(flags) $\geq 20\%$ $n = 2586$
Metric, <i>mean (SD)</i>				
Duration (s)	< 0.0001	458 (279)	401 (265)	404 (261)
Heart rate (bpm)	—	90 (16)	—	—
SpO ₂ (%)	0.5	97 (2)	97 (2)	97 (2)
†A Student’s two-tailed t-test was used to compare between valid output and SQI(flags) $\geq 20\%$				



(a) Values of HR [left] and SpO_2 [right] computed as the mean of data from the pulse oximeter with $\text{SQI}(\text{valid}) = 1$.



(b) Values of HR [left] and SpO_2 [right] computed as the mean of data from the pulse oximeter with $\text{SQI}(\text{flags}) = 1$.



(c) Values of HR [left] and SpO_2 [right] computed as the mean of data from the pulse oximeter with $\text{SQI}(\text{valid}) = 1$ (grey dots). Rejected data points that were not based on a minimum of 20% good quality data are highlighted (red crosses).

Figure 3.8: The three steps for validation of HR and SpO_2 measurements

3.4.2 Blood pressure

BP was taken as a single measurement and manually entered into the tablet software. In the antenatal clinic, the measurements were also entered into the INTERBIO–21st database. This made it possible to cross-check both databases for disagreements. A total of 25 records were found to contain disagreements, corresponding to an error rate of $< 1\%$. We therefore consider the manual entry of BP measurements to be reliable. Figure 3.9 shows the estimated values of SBP and DBP with gestational age.

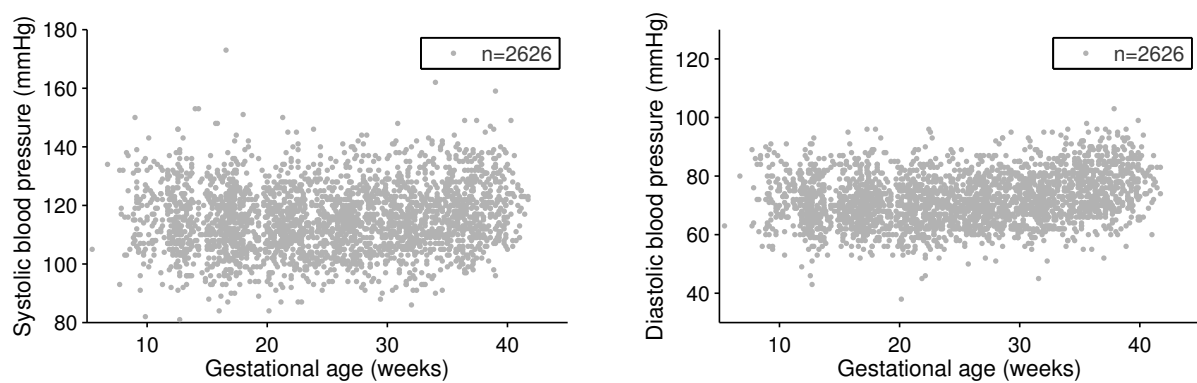


Figure 3.9: Measurements of SBP [left] and DBP [right] plotted against gestational age.

In the systematic review, we highlighted that the choice of measurement device may affect the measured values of BP. For the 4P study we used an automated monitor rather than a gold-standard manual measurement, as the automated monitor was more suitable for use in the home setting.⁶

A main limitation of the 4P study was that we only used a single measurement of BP. The European Hypertension Society [39] recommends use of a duplicate measurement to counteract a potential white coat effect; a phenomenon whereby a BP measurement is artificially high due to the patient’s feeling of anxiety in a medical environment [96, 97]. Brown *et al.* [97] found that the effect was particularly evident for measurements of DBP, with 26% of mildly hypertensive day assessment unit patients experiencing a diastolic white coat effect of ≥ 10 mmHg.

⁶The functionality of an automated BP monitor is explained in Appendix C.1.1

3.4.3 Temperature

Temp was measured with two different thermometers, as described in Section 3.2. We refer to the output Temp measurements from the Genius 2 and Fora IR20b thermometers as $Temp(man)$ and $Temp(BT)$, respectively. Pairs of measurements from the same recording session were compared by means of a Bland-Altman plot, shown in Figure 3.10. This plot displays the measurement difference against the average measurement of the two techniques.

Bland-Altman plots are frequently used to test the equivalence of two methods of measurement in the absence of a gold-standard method [98]. Two methods of measurement can be used interchangeably if the width of the 95% confidence interval (roughly equivalent to the mean bias (b) $\pm 2SD$ of the differences between measurements) is not clinically important [99]. Here, $b = 0.2^{\circ}\text{C}$ and $SD = 0.4^{\circ}\text{C}$ (in Figure 3.10, these values are illustrated by solid and dotted lines, respectively). This implies that $Temp(BT)$ is likely to measure within 1°C above or 0.6°C below $Temp(man)$. This difference is clinically significant and the two thermometers should therefore not be used interchangeably.

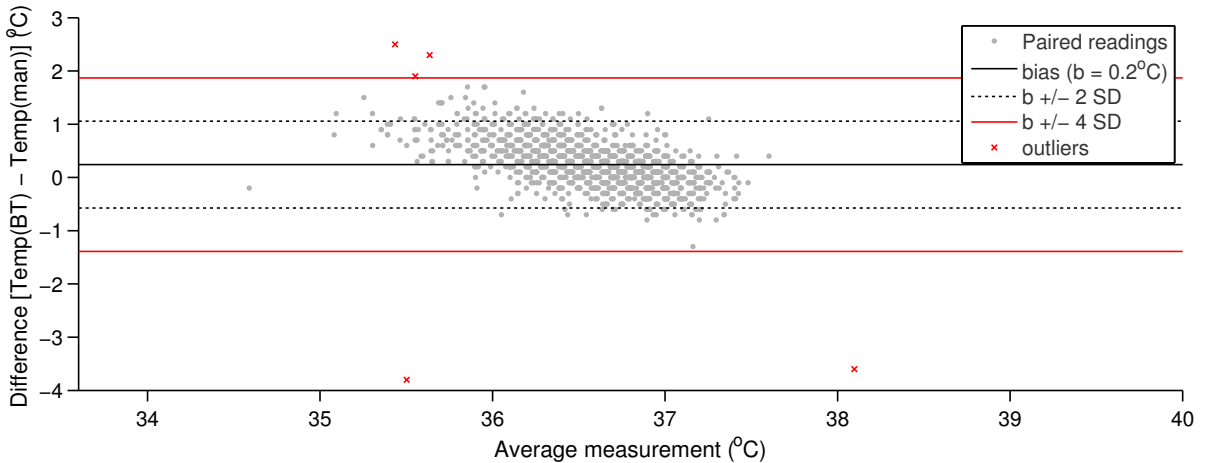


Figure 3.10: Bland-Altman comparison of paired measurements (grey dots) from the Bluetooth ($Temp(BT)$) and manually-entered ($Temp(man)$) thermometers. The lines show the bias (solid black), 95% confidence interval (dotted black) and outlier thresholds (red). Outlier values are labelled with red crosses. To display overlapping data points, a small amount of noise has been added to the x-axis.

Bland-Altman plots can also be used to identify outlier measurements. In the INTERGROWTH-21st study, outlier values were identified as pairs of measurements with a measurement difference that exceeded the limits $b \pm 4SD$ [100] (in Figure 3.10, these thresholds are illustrated by solid red lines). Figure 3.11 shows the Temp measurements plotted against gestational age. The outlier measurements have been labelled. From this figure, it is evident that outlier measurements were mainly caused by extreme (and most likely erroneous) measurements of Temp(man).

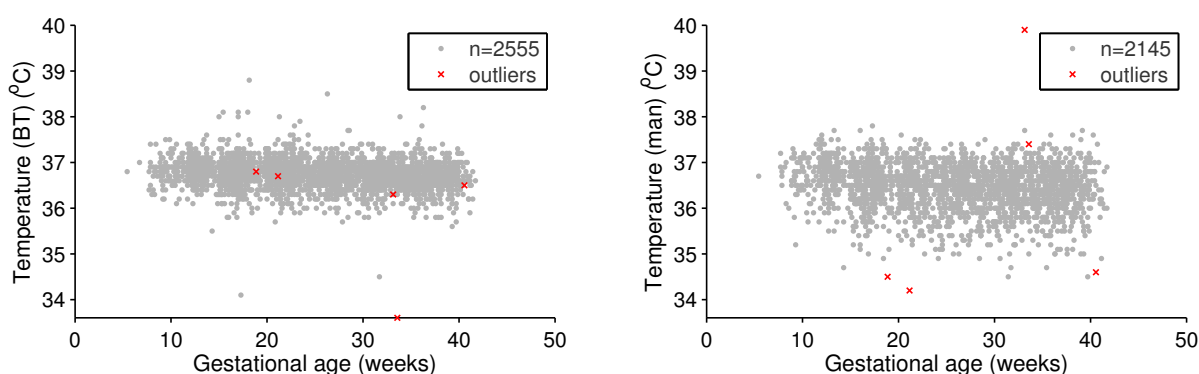


Figure 3.11: Measurements of temperature plotted against gestational age for the Bluetooth thermometer [left] and manually-entered thermometer [right].

We have collected a few potential explanations for these erroneous measurements. The bigger probe of the Genius 2 thermometer made it more difficult for the research midwives to correctly position the probe in the patient’s ear channel. Moreover, the thermometer caps had sharp edges which was more likely to cause discomfort to patients and made misalignment of the probe more likely. If the probe was misaligned in the ear channel and not pointing directly at the ear drum during the measurement, an erroneously low temperature reading may have resulted. Based on this, we have chosen to use the Temp readings Temp(BT) from the Fora thermometer for the remainder of the analysis in this thesis.

3.4.4 Diurnal variation of vital signs

Diurnal variation of vital signs has been reported for several vital signs, in particular for measurements of BP. In the 4P study, data were collected throughout the day, thereby enabling an analysis of the diurnal variation of the vital sign measurements. Figure 3.12 shows the number of datasets recorded per daytime hour in the 4P study, with the time of the measurements grouped to the nearest hour. For the analysis we only considered the time points with a minimum of 150 measurements which corresponded to measurements taken between 7:30 and 15:30.

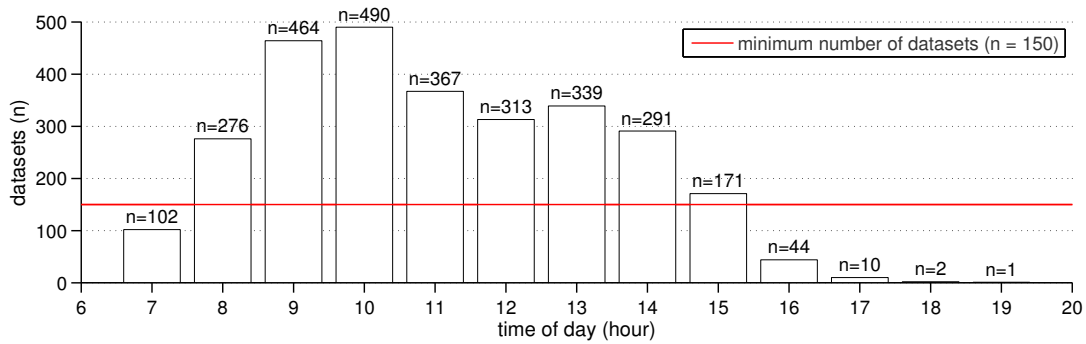


Figure 3.12: Number of datasets recorded per hour of the day

For each vital sign, the ranks of vital sign measurements at different hours of the day were compared using a Kruskal-Wallis test⁷. We did not account for any potential covariates, such as the gestational age. A summary of the vital sign data collected at the different hours of the day is shown in Table 3.8 and Figure 3.13. For all vital signs, there was a statistically significant ($p < 0.05$) difference between the ranks at different times of the day. However, it is important to note that a *statistically* significant difference does not necessarily imply a *clinically* significant variation. As we can see from Table 3.8, the amount of change in each vital sign that could be attributed to the time of day of the measurement was actually very small and unlikely to impact clinical decision-making (unless a woman is borderline abnormal).

⁷The Kruskal-Wallis test (also called a one-way ANOVA on ranks) is a non-parametric method for comparing two or more independent samples of equal or different sample sizes. Since no distribution is assumed, we compare ranks (rows of data from the same sample ordered by value).

Table 3.8: Diurnal variation of vital sign data (the measurements are grouped to the nearest hour of measurement)

Vital sign	Nearest daytime hour							
	8:00	9:00	10:00	11:00	12:00	13:00	14:00	15:00
HR (bpm)								
Number, n	274	459	485	367	309	335	291	169
Median	91	89	89	90	90	92	91	91
IQR	16	17	16	15	15	15	16	19
$p < 0.05$								
SpO ₂ (%)								
Number, n	274	459	485	367	309	335	291	169
Median	97	97	97	97	97	97	97	97
IQR	2	2	2	2	2	2	2	2
$p < 0.05$								
SBP (mmHg)								
Number, n	273	462	486	363	311	339	288	169
Median	115	116	113	115	116	117	118	116
IQR	14	16	14	16	15	14	15	14
$p < 0.0001$								
DBP (mmHg)								
Number, n	273	462	486	363	311	339	288	169
Median	72	72	71	72	72	72	74	75
IQR	12	12	12	12	14	12	12	11
$p < 0.05$								
Temp(BT) (°C)								
Number, n	266	452	474	353	303	327	281	162
Median	36.7	36.7	36.7	36.7	36.7	36.8	36.8	36.8
IQR	0.4	0.4	0.4	0.4	0.3	0.4	0.3	0.3
$p < 0.0001$								
RR (rpm)								
Number, n	260	446	465	351	292	315	273	159
Median	16	16	16	16	16	16	16	16
IQR	4	4	4	8	8	8	8	8
$p < 0.05$								

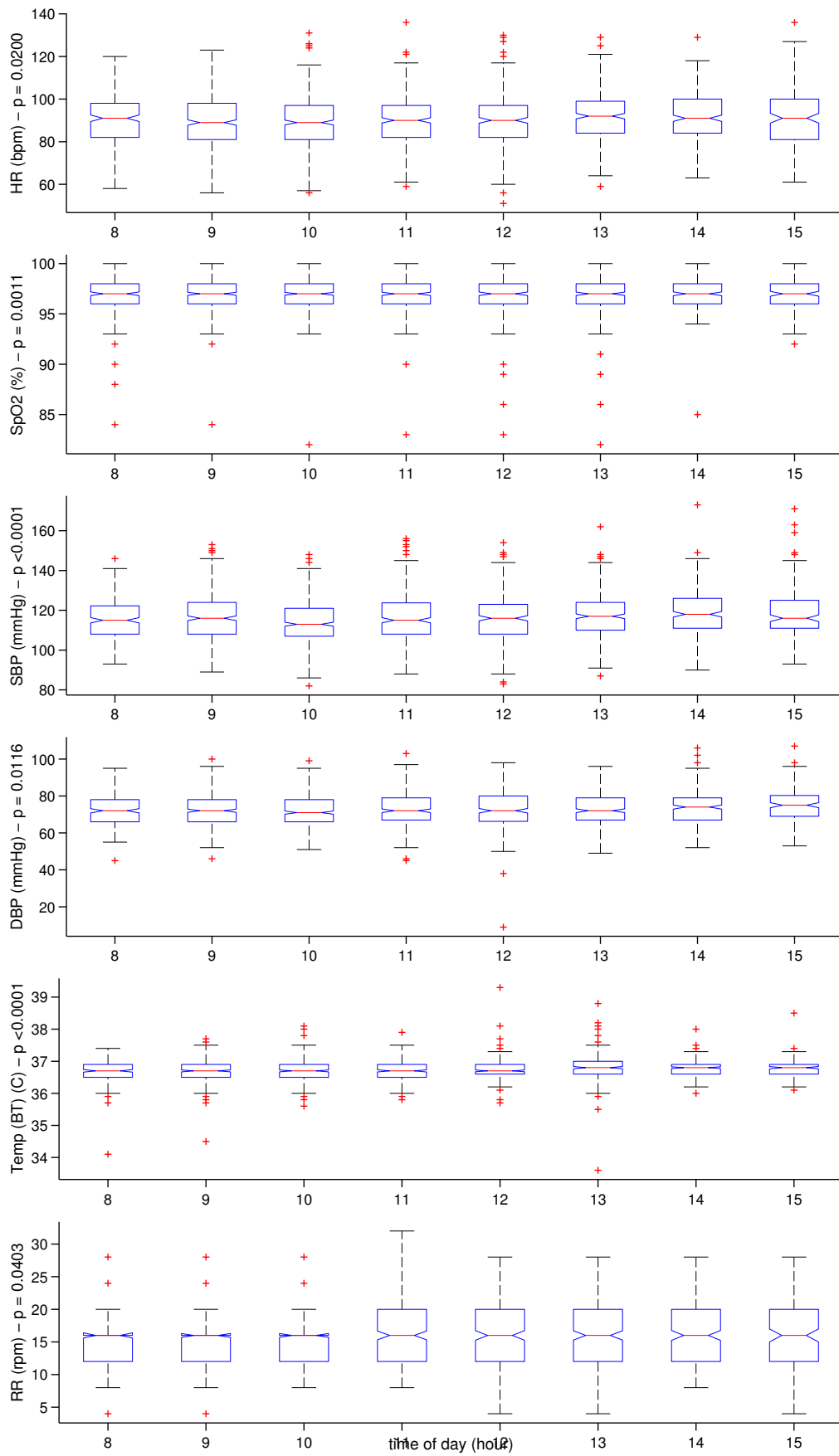


Figure 3.13: Box plots of vital sign data plotted against daytime hour.

3.5 Discussion and future improvements

The first stage of the 4P study has been completed with the collection of a large number of vital sign measurements. To collect the vital sign data, the 4P digital health system was developed in an iterative process together with healthcare professionals and pregnant women. This user-centered approach enabled the design of a system that was found to be reliable and easy to use. We have not formally assessed user satisfaction among the women and healthcare professionals who used the 4P system. In the next stage of the 4P study, we have sought ethical approval for distribution of user satisfaction questionnaires. These questionnaires will be based on the System Usability Scale [101] and provide a more formal assessment of user satisfaction.

The electronic data collection enabled prompt review and validation of the data. This resulted in a high proportion of complete datasets that contained only a relatively small number of outlier values. Prevention of error has previously been emphasised to be the best form of data quality assurance [100]. The transfer of data via Bluetooth, or where not available, the use of validation ranges incorporated into the tablet computer software, may have contributed to the high data quality and helped to ensure that data were correctly entered at the time and place of measurement.

Throughout the data collection process, routine validation procedures, such as regular data reports and validation checks, were in place to ensure the consistency and completeness of the submitted data. The electronic system for data collection enabled queries about missing data, duplicate records and outlier values to be promptly resolved and, if necessary, manually updated in the database. This contributed to the low rates of missing data reported in the previous section. The importance of ongoing, prompt quality control procedures has been highlighted by other large-scale data collection studies such as the INTERGROWTH-21st study [100, 102].

In this chapter, we also presented procedures for validation of the electronically collected vital sign measurements (namely for HR, SpO₂ and Temp) and strategies for

removal of outlier values. The difference between outliers caused by measurement error and by physiological abnormality was highlighted, and consequently outliers were only removed if there was evidence of measurement error. The identification and removal of outliers is an essential part of pre-processing in order to yield a dataset suitable for statistical analysis [100]. The number of outlier values removed was relatively small. A validation of the measurements of RR, which were manually collected, will be presented in the next chapter.

During the next phase of the 4P study, which commenced in July 2015, a more thorough validation of the vital sign measurements will be performed, including an examination of the effects of duplicate measurements of all six vital signs.

Chapter 4

Electronic measurement of respiratory rate

Of the six vital signs recommended for routine clinical assessment, respiratory rate (RR) remains the only vital sign that is not currently measured electronically [61, 62]. At the same time, it has been shown to be one of the most sensitive indicators of critical illness [103]. As part of the data collection for the 4P study, we sought to validate two methods for electronic measurement of RR, which rely on different physiological phenomena associated with breathing. The first method was measuring the chest-wall movement from breathing using the in-built accelerometer of a smartphone. The second method was measuring the respiratory-induced HR variability using the PPG signal from a pulse oximeter.

This chapter begins with a discussion of the requirements for developing an electronic measurement method for RR. This will be followed by a review of the existing literature on RR estimation using accelerometry and PPG signals. We will then develop some of these methods for application to the 4P data set and discuss the merits and limitations of the different approaches. Finally, we will combine the RR estimates from accelerometry and PPG to obtain a single 'fused' estimate of RR, which we will use in Chapter 5 to investigate normal values of RR in pregnancy.

4.1 Current measurement of respiratory rate

Because it is partly subject to conscious control, RR is often considered to be the most difficult vital sign to measure accurately [62]. Whereas the other vital signs are measured electronically, RR is most often measured manually during routine clinical assessment.

4.1.1 Measurable effects of respiration

Most current methods for measurement of RR are based on the observation of one of three main effects of respiration: measurement of airflow characteristics through the nose or mouth (such as temperature, pressure or chemical composition), measurement of the physical movement of the chest wall in response to the expansion and contraction of the lungs, and measurement of HR variability induced by a neurological (autonomic) response [104,105].

Airflow characteristics

Airflow measurement includes the use of miniature microphones to measure respiratory sounds, nasal thermistors to measure temperature changes, nasal pressure transducers to measure airflow pressure or face masks to measure gaseous exchange.

The main problem with airflow measurement is that the sensors may be uncomfortable for patients as they need to be applied in close proximity to the airways, which may cause some obstruction of the airflow. Moreover, the hardware is often complex and expensive, which render these methods unsuitable for use outside the clinic.

Physical movement

Inspiration and exhalation lead to expansion and contraction of the thorax (rib cage) as the lungs physically fill with air and deflate as we breathe. This is the movement observed by a nurse who counts the number of breaths. It can also be tracked using the use of impedance bands or strain gauges. Respiratory inductance plethysmography

(RIP) is based on this principle, whereby two bands measure movement of the rib cage and abdomen, respectively. Alternatively, motion sensors (accelerometers or gyroscopes) placed on the thorax can be used to measure chest wall movements.

An indirect measurement of chest wall movement can be obtained from the PPG. Expansion and contraction of the chest-wall leads to momentary changes in thoracic pressure and stroke volume that cause changes in perfusion baseline and peripheral blood volume (and therefore pulse amplitude, defined as the peak-to-trough distance). These effects can be extracted from the PPG [106]. The magnitude of these effects depend on the tidal volume¹ and mode of breathing. The mode of breathing is typically described as either thoracic or abdominal, depending on whether inspiration is achieved predominantly either by expanding the thorax or by expanding the abdomen (achieved by lowering the diaphragm), respectively. Of the two modes of breathing, thoracic breathing typically produces more pronounced effects on the PPG in the form of a greater baseline wander and more pronounced intensity variation [54].

Heart rate variability

Respiration has a direct effect on the variability of HR through a phenomenon known as respiratory sinus arrhythmia (RSA) [107,108]. RSA is a response of the parasympathetic nervous system that causes HR to increase during inspiration and decrease during expiration. Specifically, RSA arises when an increase in the activity of stretch receptors in the lungs leads to inhibition of the cardiac vagal (parasympathetic) neurons during inspiration, which results in an elevation of HR. During expiration, this mechanism is reversed. Since RSA is a neurological response, it is commonly observed in healthy adolescents and adults but diminishes in the elderly, the critically ill, and those suffering from chronic disease [54,108].

Heart rate variability (HRV) can be measured as the change in interval between peaks in the PPG. The HRV spectrum, computed from the time series of inter-beat intervals,

¹Tidal volume is defined as the amount of air inspired per respiration

has three main components: high frequency (HF), low frequency (LF) and very low frequency (VLF). RSA is reflected in the HF component (0.15 to 0.40Hz) of the HRV spectrum. The LF and VLF components are not related to respiratory activity. While the LF component (0.04 to 0.15 Hz) reflects vasomotor activity through baroreflexes, the origins of the VLF component remain disputed [109–111]. Figure 4.1 shows an example HRV spectrum with the HF and LF peaks labelled. This HRV spectrum was computed using a 5th order AR model² from the PPG signal shown in Figure 1.5, which was recorded from one of the 4P participants at 16 weeks of gestation.

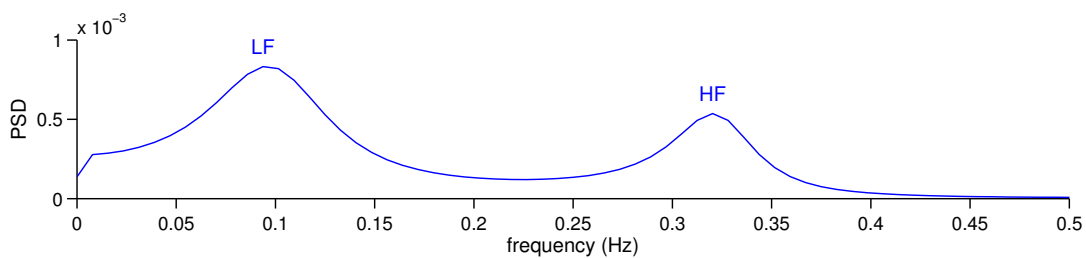


Figure 4.1: Example heart rate variability frequency spectrum with LF and HF peaks

4.1.2 Requirements for a novel electronic method

A novel method that allows for an electronic and objective measurement of RR is needed [64]. To be suitable for use with a digital health system, with applications both in the clinic and in the home setting, the proposed solution should fulfill certain requirements³. First of all, it should allow for an electronic measurement of RR that can be digitally transmitted, processed and stored. Secondly, the measurement technique should be as unobtrusive as possible to make patients less aware of the measurement and help avoid conscious adjustment of respiratory patterns. Thirdly, the solution should be cost-efficient. This implies that any hardware should be selected so as to keep the complexity and cost at a minimum. Finally, the equipment should be easy to operate by non-expert users as well as by healthcare professionals.

²The AR model will be introduced in Section 4.3.6

³The complete set of requirements for the 4P digital health system was presented in Section 3.2.

By means of the inbuilt motion sensors (accelerometer and gyroscope) of a smartphone and a pulse oximeter, we can obtain two independent, electronic measurements of RR; one which originates from the physical movement of the chest-wall and the other which depends on HRV. Both of these devices are readily available, off-the-shelf devices that fulfill the above requirements. Accelerometry and PPG data analysis will therefore be the focus of our methods to estimate respiratory rate.

4.2 Review of literature

In this section, we present previous work on the development of methods for estimation of RR from accelerometry and PPG data.

4.2.1 Respiratory rate from accelerometry data

The use of accelerometers for measurement of respiratory activity is a relatively recent field of research. Some of the earliest studies were published by Phan *et al.* [112, 113]. They used a bi-axial accelerometer, strapped around the subject's chest, to obtain motion signals on two axes that were summed to obtain a single waveform. To remove undesired frequency components they developed an adaptive filtering approach. This was based on extraction of the dominant frequency by spectral analysis of the raw signal, which was subsequently used to design a bandpass filter of fixed width, with the pass-band centered on the dominant frequency. The motivation for this approach was that deep breathing was observed to result in more noisy signals than shallow breathing. The cause for the higher level of noise was hypothesised to be that deep or abdominal breathing resulted in less movement of the upper thorax, which may have complicated the capture of chest wall movement in these patients. They only extracted a respiratory waveform for comparison with a reference waveform (obtained using an RIP band), without estimating a RR.

Later work has primarily used tri-axial accelerometers, which are now readily available and allow measurement along orthogonal axes regardless of subject position. Jin *et*

al. [114] compared three different algorithms for axis fusion of the signals from a tri-axial accelerometer. Their objective was to obtain a respiratory waveform with the highest possible signal-to-noise ratio. The most effective method for axis fusion was based on Principal Component Analysis (PCA). Liu *et al.* [115] extended this work and proposed a method for measurement of RR in subjects in different postures (sitting, lying and standing) or during activity (at rest, walking and running). To achieve this, they developed an adaptive filtering approach for which the amount of physical motion detected by the accelerometer was used to adaptively select the parameters of a band-pass filter according to a pre-defined set of thresholds. The signals were subsequently fused using PCA. Spectral (Fourier) analysis was used to estimate RR. Their algorithm was tested in a proof-of-concept study with five subjects, using expired CO₂ as a reference. They reported large errors of 4 to 15 respirations per minute, even for subjects at rest.

To the best of our knowledge, only one group has reported testing their system in a clinical setting [116]. The system was validated in 19 patients who were recovering from surgery, using the signal from a pressure-monitoring nasal cannula as a reference. By matching the two waveforms, breath-by-breath, and discarding signal segments of inadequate quality, the authors were able to match instantaneous RR within 2 rpm on 86% of occasions. To obtain a respiratory waveform, they used a tri-axial accelerometer to track the axis of rotation of the chest wall, which they converted to angular rates of breathing motion. The RR was estimated by detection of the onset of each breath using a two-step threshold, as described by Bates *et al.* [117]. The group also used this technique to attempt analysis of RR during speech [118] and during light activity [119] with limited success, as well as for detection of abnormal breathing patterns [120].

Whereas all of these studies used purpose-built accelerometry systems to record chest-wall movements, only a few studies have reported the use of the in-built accelerometer in a smartphone. The use of a smartphone accelerometer provides a particularly attractive prospect, because both the motion detection and signal processing required to extract a RR estimate can be performed with a single device. Alternatively, the signal can be

digitally transmitted for storage and offline processing. An early proof-of-concept test was described by Ono *et al.* [121], who used an iPod Touch mounted on a stand that simulated the chest-wall movement of a patient. More recently, Pechprasarn and Pongnumkul [122] used an iPhone 4 to collect and analyse the respiratory signal. The phone was resting on the chest of their subject, who was placed in a supine position. RR was obtained from the dominant spectral peak in a Fourier transform of the respiratory signal. However, their study was severely limited in that they only examined a single test subject, with selection of a specific segment of the signal for analysis (they did not report how this segment was selected).

4.2.2 Respiratory rate from the PPG

The first reports of derivation of RR from the PPG were published in the early 1990s, inspired by the work done to extract RR from the ECG [123]. However, as PPG measurements, in comparison to ECG, are less obtrusive, easier to perform in an ambulatory setting and require less costly equipment, researchers have continued to develop methods for RR extraction from the PPG. Here, we have attempted to highlight some of the key methods.

Respiratory activity causes both *amplitude modulation* (AM) and *frequency modulation* (FM) of the PPG. AM is observed as a combination of variation in the intensity (baseline wander) and variation in the amplitude (pulse peak-to-trough distance). Karlen *et al.* [106] have proposed the terms respiratory-induced intensity variation (RIIV) and respiratory-induced amplitude variation (RIAV), respectively, in order to distinguish between these two forms of AM. However, most of the older papers that describe amplitude modulation refer to the RIIV waveform. Frequency modulation (FM) is observed as variation in the time interval between beats in the PPG. It is also referred to as respiratory-induced frequency variation (RIFV) [106]. Figure 4.2 illustrates these three forms of respiratory-induced variation in the PPG.

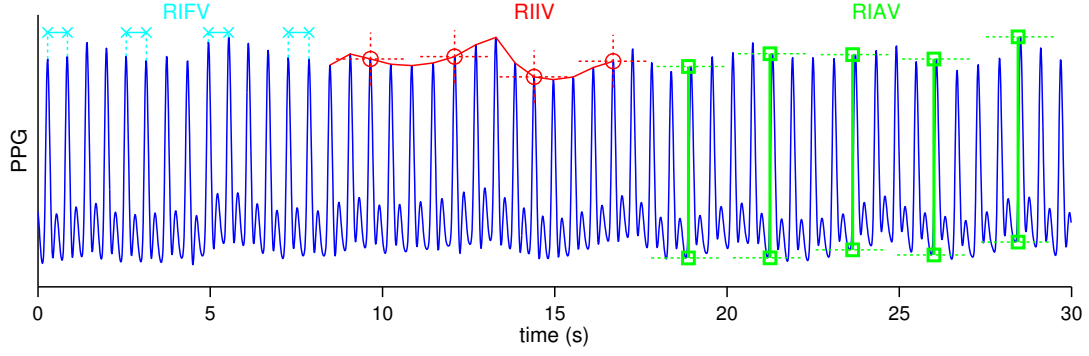


Figure 4.2: Illustration of respiratory-induced variation in the PPG. The RIFV (light blue), RIIV (red) and RIAV (green) are obtained from the pulse intervals (peak-to-peak distance), the pulse intensity (peak baseline wander) and the pulse amplitude (peak-to-trough distance), respectively.

Extraction of respiratory rate from amplitude modulation

A range of methods have been investigated for analysis of AM-derived PPG signals. The use of digital filters represents the simplest approach. The main purpose of filtering is to eliminate the cardiac-synchronous oscillations while preserving the lower-frequency respiratory waveform. Both low-pass and band-pass filters have been used, either with a uniform fixed band-width [124–126] or with a cut-off frequency selected from a bank of low-pass filters [127]. From the resulting AM-derived waveform, breath detection has been performed either manually [124, 125, 127] or by use of an automated peak-detection method [126]. Following peak detection, RR has been calculated either as an instantaneous rate (from the reciprocal of peak-to-peak intervals) or as a mean rate (from counting a number of breaths over a specified time window).

The use of spectral techniques, in which RR is obtained from analysis of the frequency content of the PPG, has also been reported. This includes use of the Fast Fourier transform (FFT) [128] and the short-time Fourier transform [129]. Fourier analysis can be applied directly to the raw PPG waveform; however, the best results have been obtained when filtering the raw PPG signal prior to spectral analysis. RR is then identified as the dominant peak in the frequency spectrum. To further improve the accuracy of spectral analysis, pre-filtering can be tailored to the individual recordings. This was described

in [130], where the cut-off frequency of a LPF was set to half the median heart rate (HR estimates were obtained from an initial Fourier analysis applied to the raw PPG signal). This was found to be particularly effective for estimation of low RR.

More recently, the use of another spectral estimation technique, the autoregressive (AR) model, has become popular. The reason for the popularity of the AR model is that a reliable frequency estimate can be obtained from relatively short signal segments containing very few respiratory cycles. This property is particularly advantageous for the analysis of biosignals, which are often non-stationary⁴. Moreover, depending on the selected model order, only a small set of relevant frequency components are identified, together with the corresponding spectral powers [131]. The model order of the AR model has either been determined empirically [132, 133] or optimised for each recording, by criteria such as Akaike's Information Criterion [134, 135]. However, information criteria should be used with caution as their application has been reported to result in the selection of artificially high model orders, which could lead to over-fitting of the model or pole splitting [134].

Selection of a respiratory frequency from the AR model can either be performed from the frequency spectrum or from the pole-zero plot. In the pole-zero plot, the respiratory frequency is given by the pole angle and the appropriate pole has in most cases been selected based on pole magnitude [132–134]. However, just as a decreased signal-to-noise ratio may cause selection of the incorrect peak in the frequency spectrum, it can also cause the wrong pole to be selected. AR models will be discussed in more detail in Section 4.3.6.

Other methods that have been developed for extraction of a respiratory frequency include various variations of wavelet transforms [136–141]. However, compared with the approaches described above, these methods are more complex and computationally demanding. A few papers have also reported on the use of principal component analysis (PCA) [142] and independent component analysis (ICA) [143]. However, as these methods

⁴Stationarity implies that the statistical properties of the signal do not vary with time.

require both the red and infrared PPG waveform as inputs we will not consider them further.

Extraction of respiratory rate from frequency modulation

In order to derive a FM signal from the PPG, the beat-to-beat interval of the cardiac-synchronous pulsatile component of the PPG must be extracted. This is done by identification of salient points in the PPG and extraction of the associated timestamps. In the literature, salient points have generally been identified as the pulse peaks. The FM waveform, also known as the tachogram or RIFV waveform, is then constructed by taking the time-interval between consecutive peaks (y-axis) and locating it at a consistent point in time, for instance at the mean time of the two points (x-axis). This results in an unevenly sampled waveform that must be interpolated and resampled to enable further processing [75, 144, 145]. Since both the FM and AM-derived waveforms represent a respiratory signal, most of the feature-based and spectral methods developed for AM analysis may also be applied to FM analysis. An excellent review of the work to date has been published by Schäfer and Vagedes [123].

More recent publications have also investigated the fusion of RR estimates using information from both the amplitude- and frequency-modulated signals. One of the first examples of a fusion algorithm was published by Orphanidou *et al.* [146], who investigated the fusion of amplitude (RIIV) and frequency (RIFV) modulated signals from the ECG. After obtaining each waveform (by peak detection), they band-pass filtered the signal with three different filters with pass-bands of 0.1 – 0.5 Hz, 0.1 – 0.6 Hz and 0.2 – 0.6 Hz and applied an AR model to each waveform, selecting the pole with the highest magnitude amongst the poles from all three waveforms. This process was repeated for both the RIIV and RIFV waveforms and the final RR was extracted from the angle of the pole with the highest overall magnitude. This method was observed to perform well across a range of RR, thereby eliminating a commonly observed problem of methods underperforming at lower or higher RR.

Karlen *et al.* [106] proposed a 'Smart Fusion' method for estimation of RR. For this method, RR was estimated separately using an FFT from the RIIV, RIAV and RIFV waveforms derived from the PPG. A single fused estimate was subsequently obtained by taking the mean of the RR estimates produced by each of the three methods, discarding cases where the standard deviation of the three estimates differed by more than 4 rpm. Their fusion approach estimated RR with a root-mean-square error of 3 ± 4.7 rpm, which was a marginally lower error than the error for the individual estimates. In doing so, 41.5% of the measurements were excluded. The reference measurement was a capnometry waveform and the method was tested on patients under general anaesthesia.

4.2.3 Measurement of respiratory rate in pregnancy

To the best of our knowledge, methods for RR estimation from accelerometry and PPG data have not been tested in pregnancy. When using the accelerometer to measure breathing, the fact that the subject is pregnant may complicate data collection. The woman's growing abdomen and bust may prevent optimal positioning of the accelerometer and subject. It is not advisable to use a strap or adhesives to secure the accelerometer, as recommended by Liu *et al.* [115] and Bates *et al.* [117]. Further, the pregnant woman should not be placed in a supine position, as lying on the back may cause compression of the inferior vena cava and a temporary but significant dip in blood pressure.

The majority of studies which estimated RR from PPG were performed in non-pregnant adults or children. As pregnancy is associated with substantial and dynamic changes to the cardiovascular system, pregnancy may affect the modulations of the PPG to either facilitate or complicate RR estimation. The effect of pregnancy on HRV is not well established, and the results of studies disagree as to whether the HF component of HRV, which contains RR information, actually increases or decreases in pregnancy.

4.3 Methods for collection and analysis of respiratory waveforms

4.3.1 Experimental set-up

To record the chest-wall movement, we used a smartphone (Samsung Galaxy Nexus, Android OS 4.1) with a custom-designed application to collect data from the inbuilt motion sensors (accelerometer and gyroscope). The woman was seated in a comfortable position with her upper body resting at 45° relative to the horizontal and the smartphone placed on her chest, as shown in Figure 4.3.

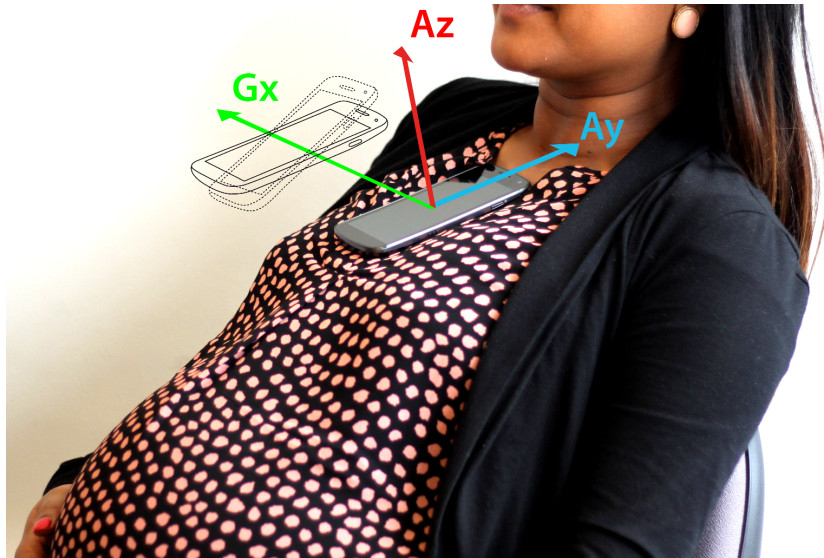


Figure 4.3: Position of the smartphone on the chest of a woman during the recording

The PPG waveform was recorded with the pulse oximeter and the 4P tablet application, as described in Section 3.2. During the antenatal appointments, the motion sensors and the pulse oximeter were used to record data for the duration of the appointment. During this time, a research midwife performed a count of breaths for 15 seconds, which was multiplied by four to obtain RR in rpm. The timestamp of the count was registered in the smartphone and tablet applications, as the research midwife pressed a button on the devices' screens at the start of count. This activated a timer that beeped when the 15

seconds had passed, to notify the midwife to stop the count. The timestamp of the count was saved together with the data for use in the analysis.

4.3.2 Window selection

The use of longer signal segments will generally produce more accurate results because they contain more respiratory cycles, which mitigate the effects of poor signal quality and artefacts. However, if the signal segment is too long, the assumption of stationarity of the signal may not hold. For each recording, although the counting period for the reference standard was only 15 seconds, we selected a longer signal segment of 30 seconds (centered on the counting period) for analysis.

4.3.3 Reference respiratory rate and error metric

The reference RR was computed from the 15-second count of breaths by the research midwives. This reference standard has an associated quantisation error of up to four respirations per minute (from the scaling used to convert the count to a RR in rpm), which can also amplify any counting error. Moreover, as the measurement is subjective it is likely to be subject to intra- and interobserver variation. It is therefore not an ideal reference standard, but it reflects the current clinical practice.

To report the results of the various estimation methods, we have used the mean absolute error (MAE) and the proportion of estimates that agreed to within four rpm of the reference RR. The formula for MAE is given by Equation 4.1, where n denotes the total number of RR estimates, RR_i , and the reference RR is denoted RR_{ref} .

$$MAE = \frac{1}{n} \sum_{i=1}^n |RR_i - RR_{ref}| \quad (4.1)$$

4.3.4 Processing of the accelerometry waveform

The accelerometer captures movement along three orthogonal axes (A_x , A_y , A_z). The gyroscope captures rotation about the same axes (G_x , G_y , G_z). The main components of chest wall movement, resulting from inhalation and exhalation, are expected to occur in the sagittal plane. With reference to Figure 4.3, this corresponds to movement along the A_y (blue) and A_z (red) axes of the accelerometer and rotation of the chest-wall about axis G_x (green). However, depending on the exact position of the woman and the smartphone during the recording, some respiratory activity may also be captured by the other three channels. The signals from all six channels were therefore recorded and used in the analysis.

Pre-processing

For the purpose of pre-processing, the six accelerometer and gyroscope channels were treated independently. The steps involved in pre-processing the signals are summarised in Figure 4.4a. The phone was set to sample data from the sensors at the maximum possible rate, which resulted in the raw output signal from the sensors being sampled at a non-constant rate between 90 and 1000 Hz. The raw signal was subsequently detrended (by removal of signal mean), resampled to 100 Hz and interpolated using cubic splines.

High frequency noise in the signal, which may result from movements, speech or even heart vibrations, was attenuated by application of a Butterworth Infinite Impulse Response (IIR) low-pass filter (pass-band frequency 0.6 Hz (36 rpm), stop-band frequency 1.2 Hz (72 rpm) and attenuation 20 dB outside the stop-band). The filtered signal was subsequently downsampled to 4.0 Hz. This frequency was chosen because it is well above the Nyquist frequency⁵, while low enough to ensure a good resolution around the expected range of respiratory frequencies. To attenuate any baseline-wander at frequencies below the minimum respiratory frequency, the signal was filtered a second

⁵The minimum sampling frequency required to avoid aliasing, which is twice the highest frequency present in the signal.

time using a Butterworth IIR high-pass filter (pass-band frequency 0.15 Hz (9 rpm), stop-band frequency 0.05 Hz (3 rpm) and attenuation 20 dB outside the stop-band). For both the high-pass and low-pass filters, the signal was filtered twice (forwards and backwards) to ensure a zero-phase shift, hence the overall attenuation in each case is 40 dB. The filter responses are shown in Figure 4.5.

Fusion of accelerometer channels

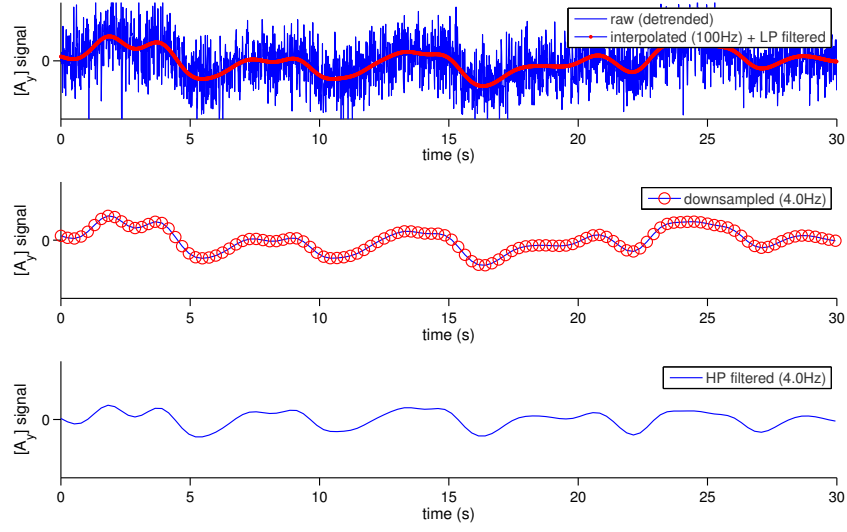
To combine the three output channels from the accelerometer and gyroscope we applied Principal Component Analysis (PCA) to the filtered waveforms as shown in Figure 4.4b. PCA is commonly used in dimensionality reduction and axis fusion can be considered as a dimensionality reduction problem. Mathematically, it rotates the sensor coordinate system to obtain a set of orthogonal, linearly-uncorrelated variables, the principal components, where the first principal component accounts for the largest variance in the data, hence represents the rotated axes with the highest signal-to-noise ratio. The 'fused' outputs from the accelerometer and gyroscope, respectively, will be referred to as A_P and G_P .

4.3.5 Processing of the PPG waveform

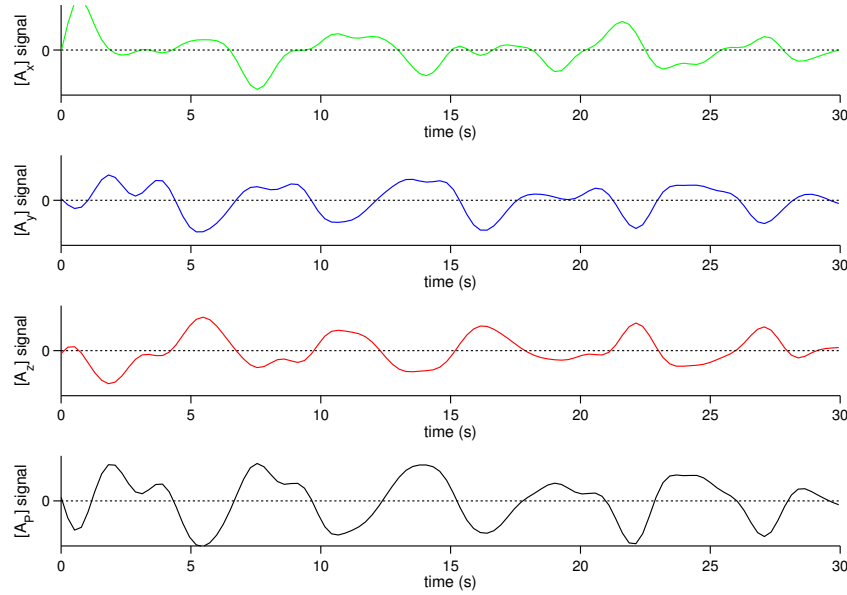
The pulse oximeter sampled the PPG at a frequency of 75 Hz. The RIIV, RIAV and RIFV waveforms were subsequently extracted from the PPG. The first step for computation of these waveforms is the detection of peaks in the PPG. This, and the subsequent steps for computation and processing, are described below.

PPG peak detection and computation of respiratory waveforms

The peak detector described here was designed based on that originally described by Fleming [75], with some minor modifications. If the parameters of the peak detector are correctly set, only valid peaks (heart beats) are counted and local extrema in the PPG (such as the dicrotic notch illustrated in Figure 1.5) are rejected. Peaks are only deemed 'valid' if the following rules are met:



(a) Processing steps for single-channel accelerometer signal. **[top]** Raw signal with the interpolated and low-pass (LP) filtered signal superimposed. **[middle]** Downsampled signal. **[bottom]** High-pass (HP) filtered signal



(b) Fusion of single-channel accelerometer signals using PCA. **[top 1-3]** Normalised single-channel accelerometer outputs A_x , A_y and A_z . **[bottom]** PCA output

Figure 4.4: Example of the pre-processing and fusion of accelerometer and gyroscope signals. The y-axes have been manually scaled for comparison, hence units are not provided.

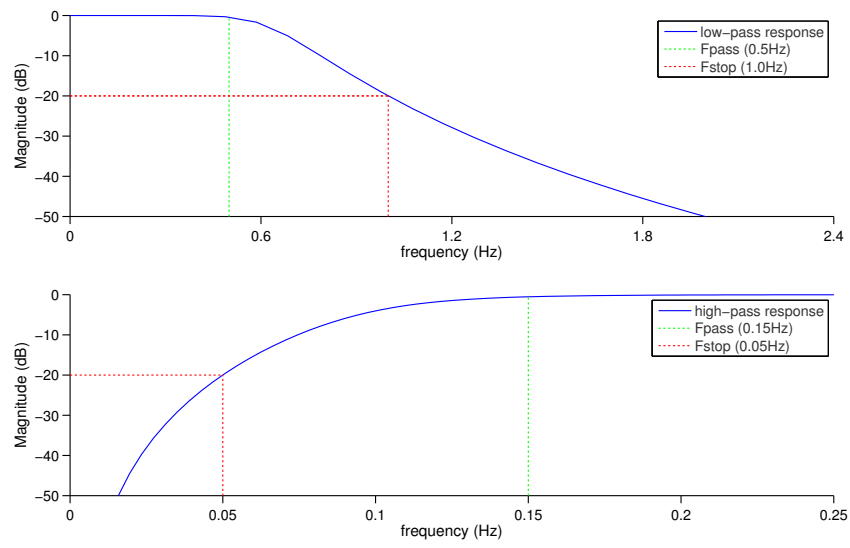


Figure 4.5: Filter responses with the pass-band (F_{pass}) and stop-band (F_{stop}) frequencies highlighted. **[top]** Low-pass filter response. **[bottom]** High-pass filter response

1. The appropriate change in gradient has occurred as given by Equation 4.2.

$$\frac{d^2}{dt^2}[X] \begin{cases} < 0, & \text{if peak} \\ > 0, & \text{if trough} \end{cases} \quad (4.2)$$

2. The extremum is of the opposite type to that of the last detected extremum (i.e. a trough must be followed by a peak and vice versa).
3. The signal value at the extremum is above a set threshold for a peak, and below a set threshold for a trough

The thresholds for rejection of local extrema were set empirically based on the mean (μ) and standard deviation (σ) of the PPG signal for that 30-second segment. Valid peaks were only detected above $\mu + 0.2\sigma$ and valid troughs below $\mu - 0.2\sigma$, where the threshold of 0.2 was set empirically. A refractory period (a minimum time between consecutive peaks) was not deemed necessary because the thresholds eliminated most local extrema in the signal, in particular those caused by the dicrotic notch.

The RIIV, RIAV and RIFV waveforms were subsequently computed using the positions of peaks and troughs in the PPG. The steps involved in the computation of these waveforms are described below.

RIIV: The RIIV waveform was computed by plotting the value of the PPG peak against the timestamp of the peak.

RIAV: The RIAV waveform was computed by plotting the amplitude of the PPG (the difference in amplitude of consecutive pairs of troughs and peaks) against the timestamp halfway between the pairs.

RIFV: The RIFV waveform was derived from a beat-by-beat time series (tachogram). This was computed by plotting the time interval between successive peaks in the PPG against the timestamp halfway between the two peaks. Outlier values in the tachogram were detected using a threshold and removed by linear interpolation of the two neighbouring values. The upper and lower thresholds were based on the expected maximum and

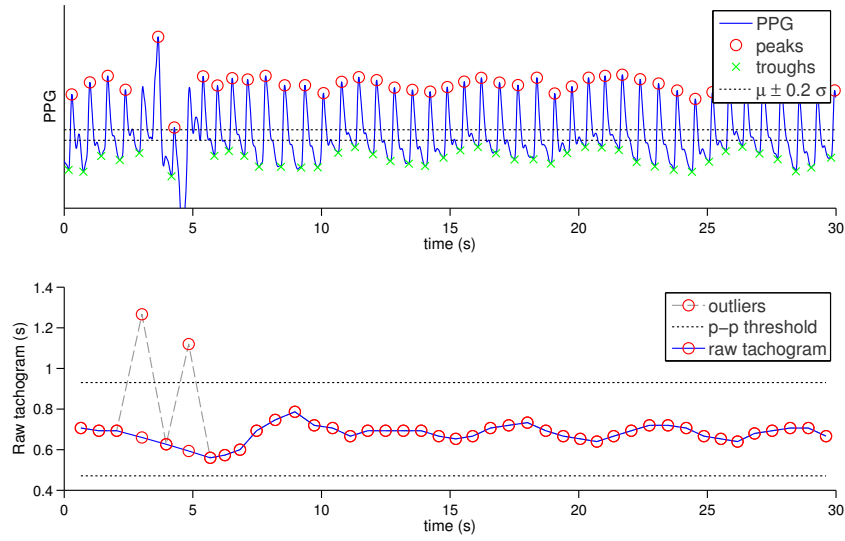
minimum time interval between peaks in the signal, derived from the minimum and maximum HR values output by the pulse oximeter, respectively. Outliers may be caused by noise in the PPG, such as that due to subject movement, or by incorrect registration of peaks and troughs by the peak detector. Such outliers may corrupt the waveform and thus make it more difficult to determine RR. The process of removal of outlier values in the tachogram is illustrated in Figure 4.6a.

Processing of the RIIV, RIAV and RIFV waveforms

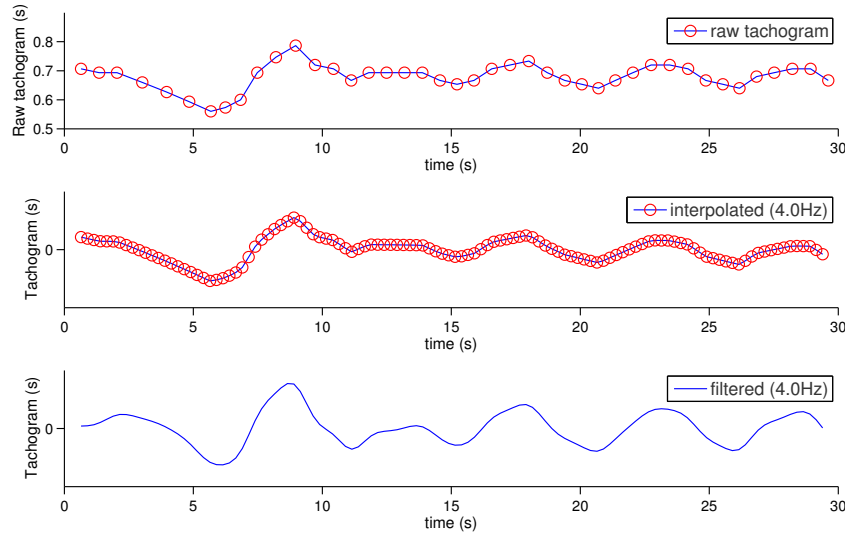
Each of the RIIV, RIAV and RIFV waveforms is an irregularly sampled time series. As most signal processing techniques require an evenly sampled signal, each waveform was linearly interpolated to 4.0 Hz. A Butterworth IIR high-pass filter (pass-band frequency 0.15 Hz, stop-band frequency 0.05 Hz and attenuation 40 dB outside the stop-band) was subsequently applied, in order to attenuate any baseline wander. High-pass filtering is particularly important for the RIFV waveform, in order to attenuate the LF and VLF frequency components that are not related to respiratory activity. The waveforms were also low-pass filtered (Butterworth IIR low-pass filter with a pass-band frequency 0.6 Hz, stop-band frequency 1.2 Hz and attenuation 40 dB outside the stop-band) to attenuate any sharp peaks, from which we obtained a smooth waveform for analysis. With both filters, the signal was filtered twice (forwards and backwards) to ensure a zero-phase shift. These steps have been summarised in Figure 4.6b.

4.3.6 Algorithms for respiratory rate estimation

To obtain RR estimates from each respiratory waveform, a series of RR estimation algorithms were investigated. The selected algorithms included both feature-based methods (peak detection) and frequency-based methods (Fourier analysis and autoregressive models). The output of each algorithm is an estimate of the respiratory *frequency* (in Hz), which is multiplied by 60 to obtain a respiratory *rate* (in rpm). The range of observed reference RR was 4 – 32 rpm, hence the outputs were restricted to a range of plausible values of



(a) Example of computation of tachogram from a PPG signal. **[top]** A 30-second segment of PPG signal with peaks (red) and troughs (blue) identified. **[bottom]** Tachogram computed from PPG. The outlier values (grey) result from artefacts in the PPG. These are identified using a threshold (black) and corrected by linear interpolation of neighbouring values.



(b) Example of processing of tachogram waveform. **[top]** Raw (unevenly sampled) tachogram. **[middle]** Tachogram linearly interpolated to 4 Hz. **[bottom]** Filtered tachogram.

Figure 4.6: Example of the computation and processing of a tachogram derived from the PPG to obtain a respiratory waveform

3 – 33 rpm (0.05 – 0.55 Hz). This range will from now on be referred to as the domain of RR (*RR domain*). Outputs outside this range were deemed invalid.

Peak detection

Individual breaths were identified using a peak detector, which was based on the same principles as those described in Section 4.3.5, with two minor changes:

1. [PPG and accelerometry waveforms] a refractory period was used to define the minimum time between two consecutive breaths; and
2. [accelerometry waveforms only] the threshold for rejection of local extrema (criterion 3) was set to the mean μ of the signal for both peaks and troughs, with valid peaks $> \mu$ and valid troughs $< \mu$.

The refractory period was set empirically to 1.67 seconds corresponding to a maximum observed RR of 36 rpm (above the maximum expected RR of 33 rpm, to allow for irregularities in the duration of each breath). The difference in the algorithms applied to the PPG and accelerometry waveforms was necessary because the PPG waveforms contained more low-frequency components. This will be discussed further in Section 4.4. Only 'valid' peaks in the signal were counted as breaths and 'invalid' peaks were rejected as noise, as shown in Figure 4.7a. The mean and median respiratory frequency (f_R) were found from Equations 4.3 and 4.4, respectively. Here, t_{win} is the length of the signal (in seconds), Np is the number of valid peaks p and Δt_p is the time between valid peaks.

$$f_R = Np/t_{win} \quad (4.3)$$

$$f_R = \text{median}(\Delta t_p)^{-1} \quad (4.4)$$

Fourier spectral analysis

Dominant frequencies in the signal were also identified using the FFT. The FFT produces a discrete frequency spectrum, referred to as a periodogram, where prominent frequencies in the signal are characterised by higher power spectral densities (PSD). The resolution (Δf) of the periodogram for any given sampling frequency (f_s) is inversely proportional

to the number of sampling points (N_{win}), as expressed by Equation 4.5. To increase the frequency resolution, longer signal sequences (collected over a larger time window) would have been required.

$$\Delta f = f_s / N_{win} \quad (4.5)$$

A time window of 30 seconds, sampled at a frequency of $f_s = 4\text{Hz}$, is equivalent to a 120-point window. RR estimates were therefore obtained with a resolution of 0.33 Hz (equivalent to 2 rpm).

Following the application of the FFT, with a pre-processing Hamming window to minimise spectral leakage, the respiratory frequency (f_R) was identified as the dominant (highest magnitude) peak in the part of the spectrum that contained frequencies within the *RR domain*, as illustrated in Figure 4.7b.

Autoregressive modelling

Autoregressive (AR) models have become popular for analysis of biological data because of their ability to resolve frequency information from very few cycles or even fractions of a cycle of a periodic signal. The estimated AR spectrum is a continuous function of frequency, hence the resolution of the spectrum does not depend on the length of the analysed signal.

The AR model is defined by Equation 4.6, where the current value $x(t)$ of the time-series can be modelled as the linearly weighted sum of the preceding m values.

$$x(t) = \sum_{k=1}^m a_k x(t-k) + e(t) \quad (4.6)$$

The model order m is usually much smaller than the length of the sequence N . The error term $e(t)$ represents the error between the predicted and the current value at the point t . It is assumed here to be normally distributed with zero mean and variance σ^2 . When visualised as a system with input $e(t)$ and output $x(t)$, the z-transform of the transfer function H becomes

$$H(z) = \frac{1}{\sum_{k=1}^m a_k z^{-k}} = \frac{z^m}{(z - z_1)(z - z_2) \dots (z - z_m)}. \quad (4.7)$$

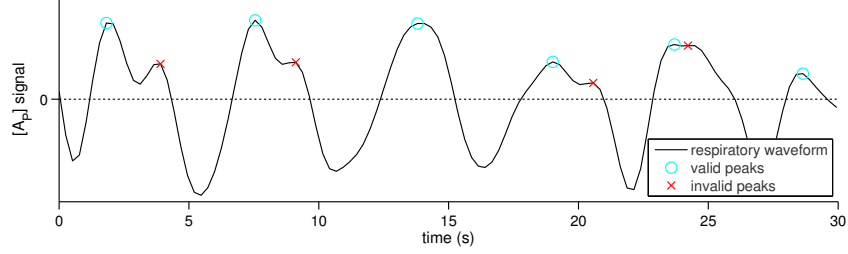
The denominator of $H(z)$ can be factorised into m terms, where $a_1, a_2, \dots, a_m$ and σ^2 are parameters to be estimated. Each root z_i of the denominator corresponds to a pole of $H(z)$. The poles of Equation 4.7 occur in complex conjugate pairs and define spectral peaks in the power spectrum of the signal. By rewriting these in the Euler form (Equation 4.8), we can obtain the frequencies f at which the spectral peaks occur.

$$z_i = e^{j2\pi f \Delta t}, \quad i = 1, \dots, m \quad (4.8)$$

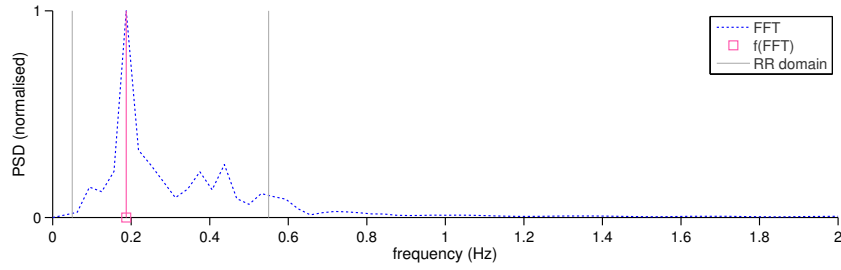
Here, Δt is the sampling interval of the signal $x(t)$. The frequency of each spectral peak is given by the phase angle of the corresponding pole on a pole-zero plot in the z -domain. The height of a spectral peak relates to the magnitude of that pole, with higher magnitude poles corresponding to higher magnitude peaks.

In order to accurately resolve frequencies in the *RR domain*, the respiratory waveform was downsampled to 2.0 Hz. The downsampling process enhances the frequency resolution of the pole-zero plot for low frequencies by allowing a given frequency range to be represented by a greater number of degrees. This is illustrated in Figure 4.7c, where the proportion of the pole-zero plot occupied by *RR domain* is shown. A downsampling frequency of 2.0 Hz ensured that all poles modelled frequencies up to 1.0 Hz (an upper limit imposed by the Nyquist criterion).

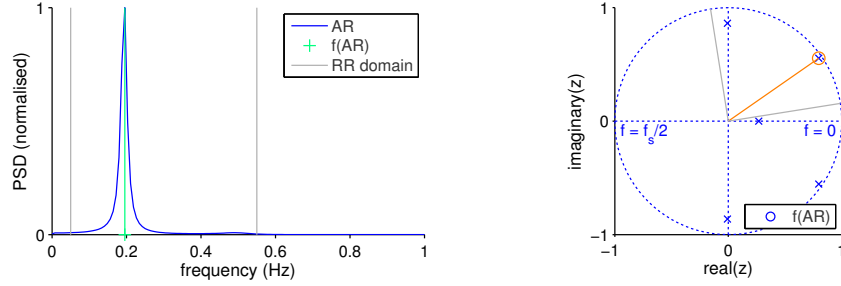
The model order m determines the number of poles that are available to model the frequency spectrum. Based on inspection of the mean absolute error with respect to the reference RR, a model order of $m = 5$ was found to provide the optimal fit. To compute the AR coefficients, we used the Maximum Entropy Method, which is implemented by the Burg Algorithm in Matlab [147]. The respiratory frequency (f_R) was identified in two ways:



(a) An example 30-second respiratory waveform with valid peaks (breaths) and invalid peaks (noise) labelled.



(b) The power spectral density (PSD) against frequency computed from a Fourier transform. The dominant frequencies (pink line) in the PSD is selected by maximum peak height. The limits of the *RR domain* are shown (grey lines).



(c) The power spectrum of a 5th-order AR model [left] and the associated pole-zero plot [right]. The 5th-order model gives two possible frequencies. The dominant frequencies (coloured lines) in the PSD and pole-zero plot are selected by maximum peak height and maximum magnitude, respectively. The limits of the *RR domain* are shown in both plots (grey lines).

Figure 4.7: Example application of the RR estimation algorithms using a 30-second ($N_{win} = 120$) respiratory waveform.

AR peak: The respiratory frequency is selected as the dominant peak in the AR power spectrum.

AR pole: The respiratory frequency is selected as the highest magnitude pole in the pole-zero plot.

4.3.7 Signal quality assessment

The feasibility of RR extraction has been observed to relate directly to the periodicity of the respiration waveform. An ideal respiratory waveform is one that resembles a periodic sinusoid, with a single dominant frequency that represents the RR [148]. The respiratory waveforms derived from both the accelerometry data and the PPG were subject to corruption as a result of motion artefacts. Motion artefacts are therefore an impediment to reliable estimation of RR.

To deal with motion artefacts, two main approaches have previously been described in the literature: *(i)* the development of signal processing techniques to suppress the motion artefact; or *(ii)* identification of good-quality segments and subsequent processing of only these segments. Shah *et al.* [145] have previously stated that an approach based on signal quality assessment will not only improve estimation accuracy, but also increase the confidence in it. As our aim was to obtain a single *reliable* estimate of RR, we also chose to investigate methods for signal quality assessment.

To assess signal quality, we devised two signal quality indices (SQI) that were applied to the pre-processed respiratory waveforms from either the motion sensors or PPG. Each SQI took a value between 0 and 1, where a higher value indicated a closer resemblance to an 'ideal' respiratory waveform.

SQI from spectral characteristics of the Fourier transform

SQI(fft) was computed as the power spectral density associated with the dominant (respiratory) frequency f_r normalised by the total power spectral density across the *RR domain*, as described by Equation 4.9.

$$\text{SQI}(\text{fft}) = \sum_{k=r_1}^{r_2} |S_k(f)| \bigg/ \sum_{k=R_1}^{R_2} |S_k(f)| \quad (4.9)$$

Here, $|S_k(f)|$ is the estimate of power spectral density. The limits r_1 and r_2 define the area under the spectral peak, where $r_1 = f_r - \Delta f$ and $r_2 = f_r + \Delta f$. The value Δf is the quantisation of the Fourier spectrum. Given that all of the analysed segments were of the same length, Δf was the same across all recordings. The limits R_1 and R_2 correspond to the upper and lower limits, respectively, of the *RR domain*. These areas were computed by the trapezium rule for numerical integration and are illustrated in Figure 4.8a.

SQI from autocorrelation

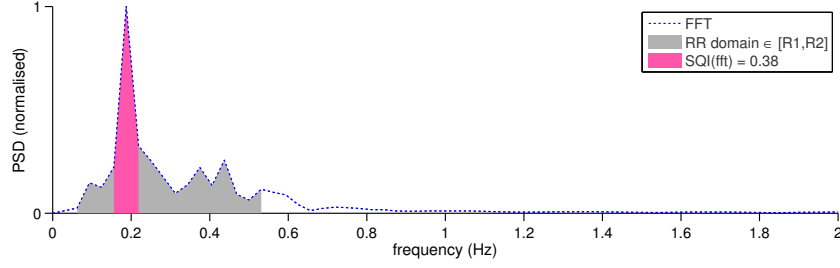
$\text{SQI}(\text{acf})$ is based on the sample autocorrelation function given by Equation 4.10. It is computed as the sample correlation r_k between pairs of samples k units apart and normalised by the overall mean $\bar{y}$ and overall sum of squares c_0 for the 30-second segment.

$$r_k = \frac{c_k}{c_0}, \quad \text{where} \quad c_k = \sum_{t=k+1}^n (y_t - \bar{y})(y_{t-k} - \bar{y}), \quad k = 1, 2, \dots \quad (4.10)$$

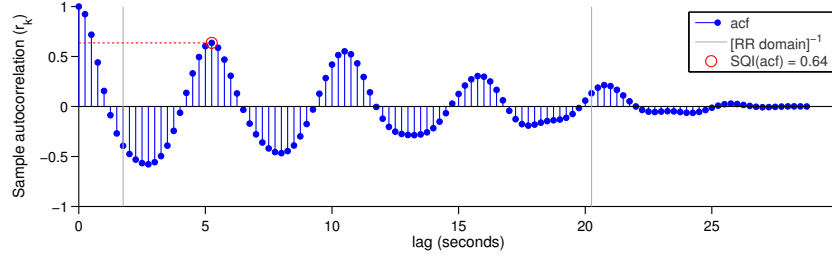
$\text{SQI}(\text{acf})$ was defined as the highest value of autocorrelation within the range of frequencies that represent the *RR domain*. Considering that each sample lag k represents a time lag of $\tau = f_s^{-1}$,

$$\text{SQI}(\text{acf}) = \max(r_k(\tau)), \quad \tau \in [RR \text{ domain}]^{-1} \quad (4.11)$$

Figure 4.8b illustrates the sample autocorrelation function and the selection of $\text{SQI}(\text{acf})$.



(a) A frequency spectrum with shaded areas that illustrate the total power spectral density of the *RR domain* (grey) and of the respiratory peak (pink).



(b) A sample autocorrelation function with the value of $SQI(acf)$ (red circle) and time lags that correspond to the limits imposed by the *RR domain* (grey vertical lines).

Figure 4.8: Example application of the two SQIs (a) $SQI(fft)$ and (b) $SQI(acf)$ computed for the respiratory waveform in Figure 4.7a.

4.3.8 Summary of methods for estimation of respiratory rate

We have described five different RR estimation algorithms for obtaining an estimate of RR from a respiratory waveform. Figure 4.9 summarises these algorithms and the required pre-processing steps. The five algorithms included both feature-based methods (PK:mean and PK:med, as shown in Figure 4.9) and frequency-based methods (AR:peak, AR:pole and FFT:peak, as shown in Figure 4.9). In the next section we will apply these algorithms to 4P data to obtain electronic estimates of RR. The results from RR estimation using accelerometry and PPG data are presented in the following sections.

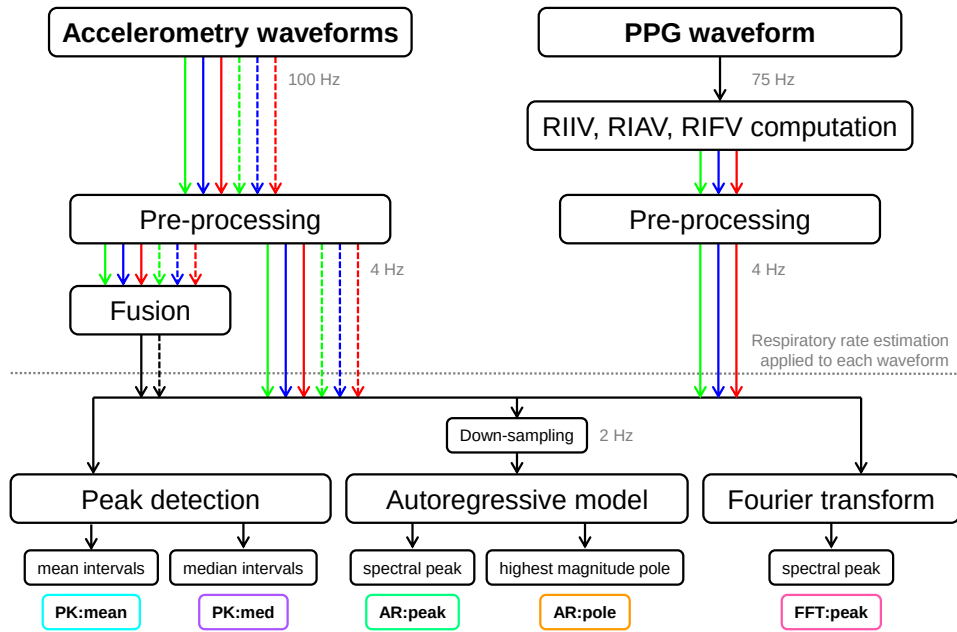


Figure 4.9: Summary of algorithms for the estimation of RR from accelerometry and PPG waveforms. The green, blue and red lines represent the different channels of the accelerometer (solid lines) and gyroscope (dotted lines), as well as the RIIV, RIAV and RIFV waveforms extracted from the PPG.

4.4 Respiratory rate estimates from the 4P data

Of the datasets collected in the 4P study during the antenatal period, 2,167 datasets contained both a reference RR and signals of adequate duration (minimum 25 seconds) from the accelerometer and pulse oximeter. It was not possible to detect peaks in 93 (4.3 %) of the PPG signals, hence the total number of RR estimates from the PPG were 2,074. In what follows, we will first discuss the accuracy of the estimates and potential sources of error, then compare the five RR estimation algorithms.

4.4.1 Respiratory rate estimates from accelerometry data

The processing of a single accelerometer recording results in a total of eight waveforms (signals) for analysis (six channels originally recorded by the accelerometer and gyroscope and two fused waveforms). Figure 4.10 shows, for each of the eight signals, the variation in mean absolute error (relative to the reference measurement described in Section 4.3.3) across the 2,167 recordings with different RR estimation algorithms.

Of the six channels originally recorded by the accelerometer and gyroscope, RR estimates from the A_z , A_y and G_x signals were associated with the lowest estimation errors. Based on our knowledge of the movement of the thorax during breathing, this result is expected: the thorax rises in the shape of a hemisphere (moving along the y and z axes) and then rotates upwards (rotating about the x axis) during inspiration. This movement is reversed during expiration. Figure 4.3 illustrated the alignment of the axes with the thorax.

Signals from the gyroscope resulted in estimates with a greater error than the signals from the accelerometer. The reason for this could be because the phone was simply resting on the chest of the woman; in order to properly capture rotation, the sensors (in this case, the smartphone) would need to be more firmly fixed to the thorax. These results agree well with observations made in a previous study by Yoon *et al.* [149].

The waveforms A_P and G_P , obtained using PCA, were associated with lower errors than the original constituent waveforms. Overall, the lowest mean absolute errors were associated with the fused accelerometer (A_P) signal. This result is consistent across the five RR estimation algorithms. Table 4.1 summarises the error metrics obtained for each of the five RR estimation algorithms, using the A_P signal as the input.

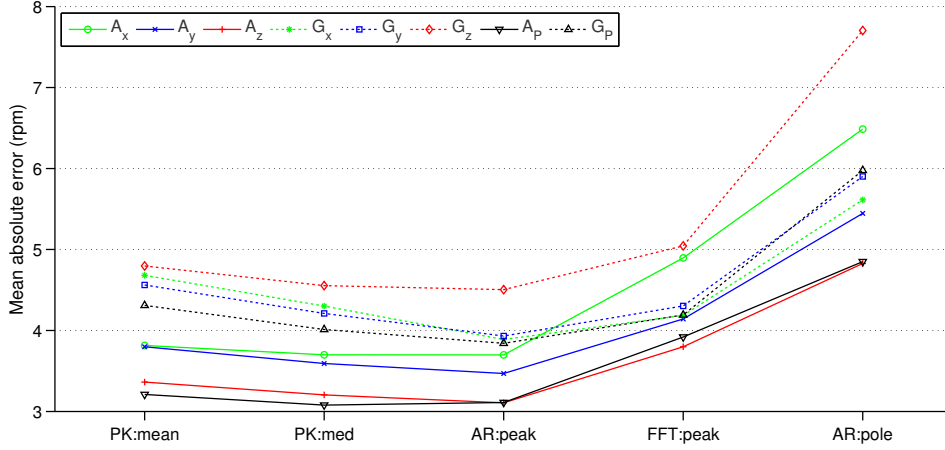


Figure 4.10: Variation of mean absolute error with different RR estimation algorithms using eight different signals from the accelerometer and gyroscope (as explained in the plot legend).

Potential sources of measurement error

In order to investigate potential reasons for the observed errors between the estimated and reference RR, we compared the performance of each algorithm at different reference RR using the A_P signal as input. Figure 4.11 shows, for each RR estimation algorithm, a plot of the estimated RR against the reference measurement. The mean estimated RR values are superimposed for reference RR between 8 and 28 rpm. The associated error metrics for the same set of reference RR are summarised in Table 4.1, where the results of the best performing algorithm at each reference RR is highlighted in bold. Additionally, we have provided three example recordings that will be used to illustrate our observations. These are presented in Figure 4.12. Each example includes a respiratory waveform (the A_P signal), with the associated frequency spectra, pole-zero plots, and estimated and reference RR shown below each waveform.

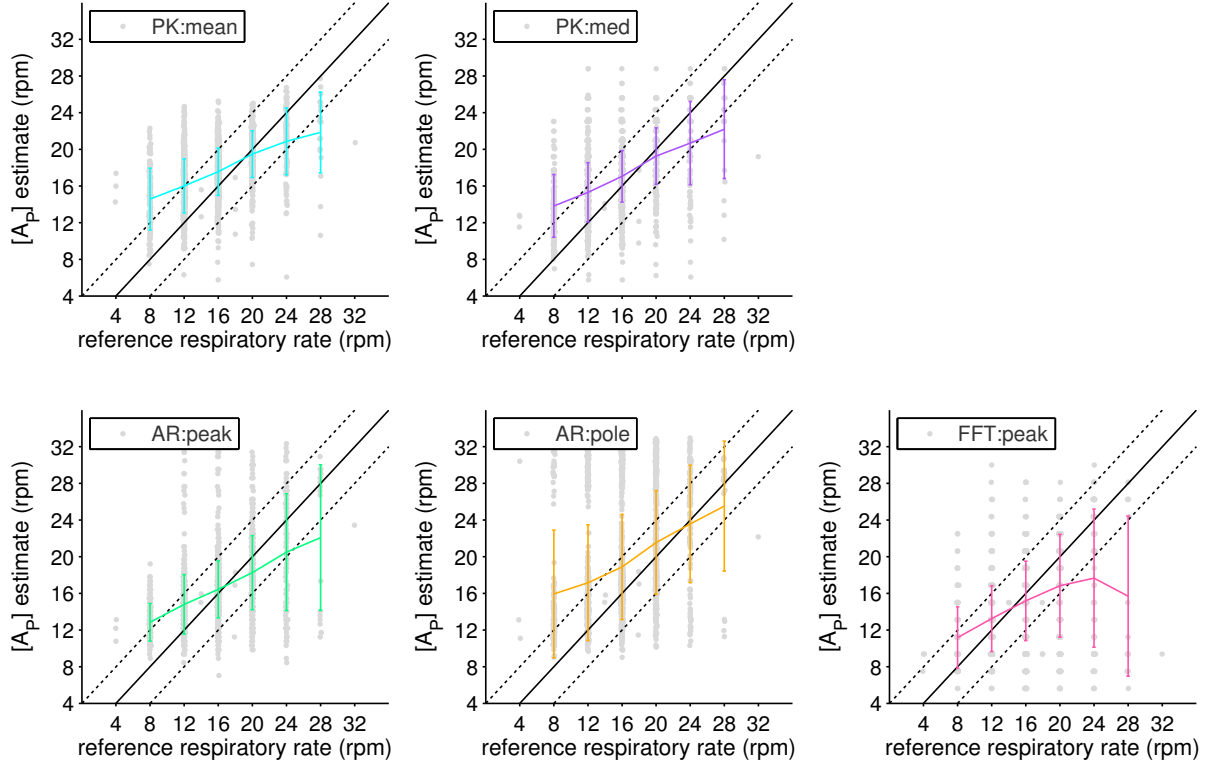
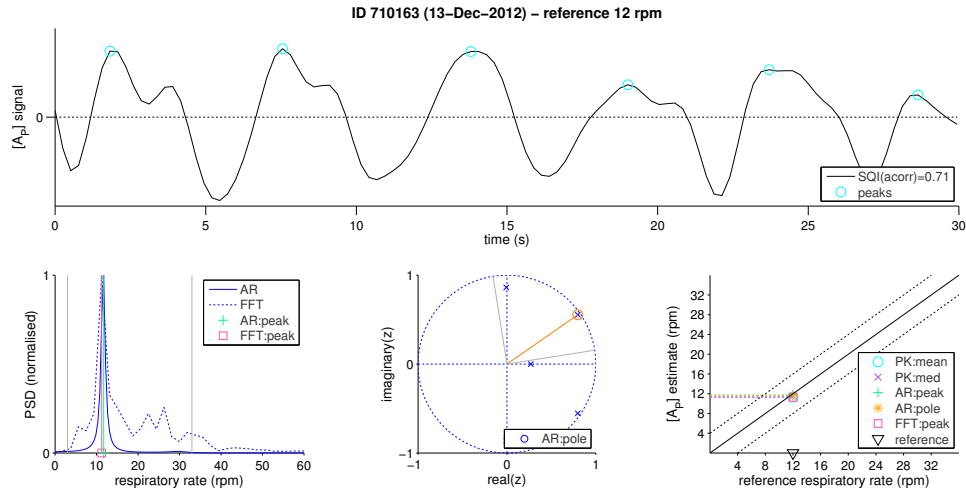


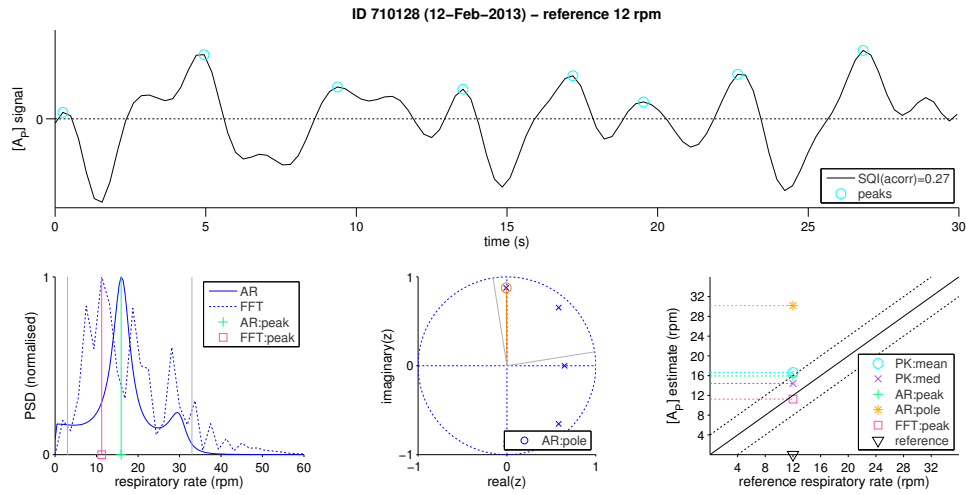
Figure 4.11: Variation of estimated RR from the A_P waveforms plotted against the associated reference RR (grey dots). The mean (SD) estimated RR at reference RR of 8, 12, 16, 20, 24 and 28 rpm are superimposed (coloured error bars). Estimates within ± 4 rpm of the reference RR fall within the dotted lines.

Table 4.1: Error metrics for the five RR estimation algorithms at selected reference RR for the A_P waveforms. The results of the best performing algorithm at each reference RR are highlighted in bold.

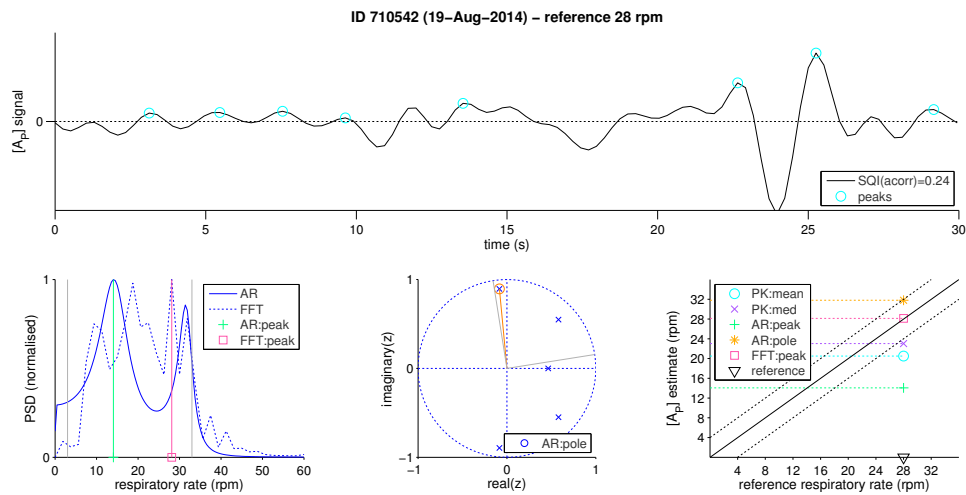
Input signal [A_P]		Reference respiratory rate (rpm)						
		All	8	12	16	20	24	28
Number of estimates, n		2167	141	619	848	419	112	19
PK:mean	Absolute error (rpm), <i>mean</i>	3.2	6.6	4.2	2.4	1.9	3.6	6.2
	Within ± 4 rpm, %	70	23	53	81	90	64	37
PK:med	Absolute error (rpm), <i>mean</i>	3.1	5.8	3.6	2.3	2.4	4.2	6.0
	Within ± 4 rpm, %	73	30	63	83	85	63	53
AR:peak	Absolute error (rpm), <i>mean</i>	3.1	4.9	3.0	2.2	3.5	5.9	6.8
	Within ± 4 rpm, %	74	40	78	87	62	45	58
AR:pole	Absolute error (rpm), <i>mean</i>	4.8	7.9	5.3	4.1	4.5	5.3	4.3
	Within ± 4 rpm, %	64	36	65	75	57	45	79
FFT:peak	Absolute error (rpm), <i>mean</i>	3.9	3.4	3.0	3.5	4.9	7.5	12.3
	Within ± 4 rpm, %	63	73	73	60	55	47	37



(a) Good quality recording with a low reference RR.



(b) Medium-poor quality recording with a low reference RR.



(c) Recording with a high reference RR containing a large artefact.

Figure 4.12: Examples of three A_P waveforms and estimated RR. Each figure (a-c) shows the respiratory waveform [top] and the associated power spectra [bottom left], pole-zero plot [bottom middle] and the estimated RR plotted against the reference measurement [bottom right].

From these data, we found that the estimation of RR is most accurate between 16 and 20 rpm. Moreover, we made two key observations about the accuracy of RR estimation from the A_P signal (i) at low RR and (ii) at high RR.

The error at low values of RR is thought to be a consequence of irregular breathing patterns. To illustrate this, Figures 4.12a and 4.12b show two example recordings from subjects with a slow reference RR. For the example waveform in Figure 4.12a, the periodic respiration cycles are easily identifiable and the five RR estimation algorithms produce similar RR estimates. Figure 4.12b, on the other hand, shows an example of an irregular breathing pattern. Here, the irregular duration of the breaths and the variability in the signal complicate the identification of breaths and cause greater variation in the RR estimates from the different algorithms. In particular, the AR:pole algorithm gives an erroneously high estimate. The difficulty of accurate estimation of low values of RR may result from the higher level of noise associated with deep breathing, as previously described by Hung *et al.* [112] and discussed in Section 4.2. This effect was potentially exaggerated by our experimental set-up, where the smartphone was positioned high on the pregnant woman’s chest to accommodate the growing abdomen.

At high values of RR, the accelerometer signals were prone to high-amplitude single artefacts, which most likely occurred as a result of movement. Such movement artefacts introduce undesired low-frequency components in the respiratory waveforms. To illustrate this, Figure 4.12c shows an example of a high RR waveform that has been corrupted by a large artefact. The low frequency component introduced by this artefact can be seen as a large peak in the associated AR frequency spectrum. Such artefacts were particularly prominent at high RR (> 20 rpm). This is likely to be because a higher RR is associated with more shallow breathing, which yields signals with a lower amplitude.

4.4.2 Respiratory rate estimates from PPG data

Following the same procedures as for the accelerometry data, Figure 4.13 shows the variation in mean absolute error with different RR estimation algorithms for the RIIV,

RIAV and RIFV waveforms. Across the five algorithms, the lowest errors were consistently obtained using the RIFV signal. As discussed in Section 4.1.1, frequency modulation of HR is a phenomenon that is particularly associated with young and healthy individuals [54]. It was therefore not surprising that the pregnant women in our sample, of whom the majority could be described as young and healthy, exhibited a significant amount of RSA. The commercial pulse oximeter outputs a filtered version of the PPG waveform, which may attenuate some of the amplitude variation in the PPG from which the RIIV and RIAV signals are computed [53]. Because of this, we have focused the remainder of our discussion on the RIFV signal.

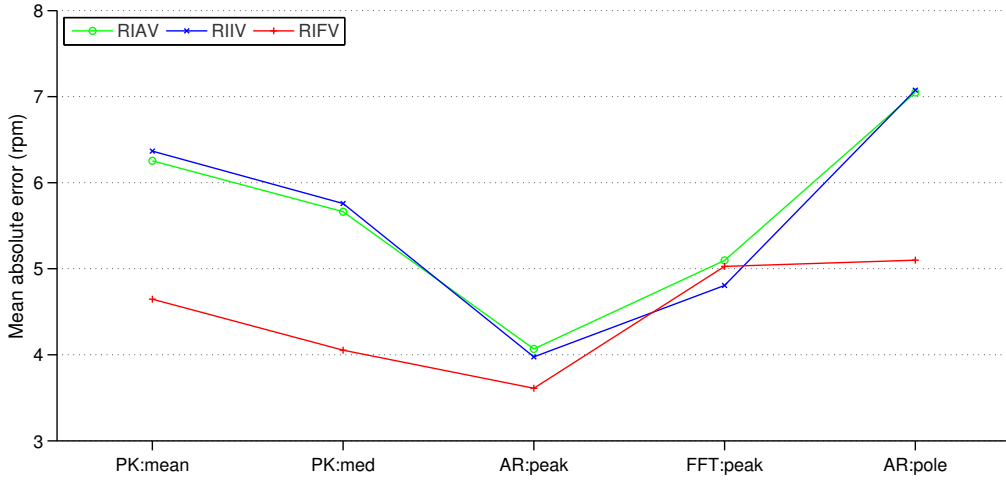


Figure 4.13: Variation in mean absolute error with different RR estimation algorithms for the RIIV, RIAV and RIFV waveforms.

Sources of error for PPG

The RR estimates against reference RR are plotted in Figure 4.14. The associated error metrics for a set of reference RR between 8 and 28 rpm are summarised in Table 4.2. Figure 4.15 provides three example recordings that will be used to illustrate our observations. For each example recording, the original PPG signal and the associated RIFV respiratory waveform are shown. The associated frequency spectra, pole-zero plots and the estimated and reference RR are shown below each respiratory waveform.

In comparison with the RR estimates derived from the A_P accelerometer signal, the

overall mean absolute errors associated with the PPG-derived RR estimates are slightly greater. In particular, a greater deviation between the estimated and reference RR is observed at higher RR (> 16 rpm) for the frequency-based algorithms and at lower RR (< 16 rpm) for the feature-based algorithms. From inspection of the respiratory waveforms and associated RR estimates, we observe three main sources of error that could have contributed to the large errors observed for high RR.

Firstly, errors can arise during the recording of the PPG signal (as discussed in Section 3.4.1) as the PPG is sensitive to movement or recording artefacts. Such artefacts may moreover not always be picked up by the SQI output by the pulse oximeter. Secondly, errors can arise from the computation of the PPG-derived waveforms, for example through a failure to identify the correct peaks in the PPG. In particular, we note that failures of the peak detector are frequently linked to artefacts in the PPG waveform. Finally, erroneous measurements may be due to the physiological characteristics of the subject that can complicate the measurement; as discussed in Section 4.2, the degree of RSA present in the PPG may vary with age or health status. Recently, Gandhi *et al.* [150] have observed that the parasympathetic modulation of HRV decreases with gestational age. This reduction in parasympathetic activity manifestes itself as a reduction in spectral power of the HRV HF frequency band (associated with RSA), which may cause the HF frequencies to be masked by lower frequencies. The three example recordings provided in Figure 4.15 serve to illustrate these sources of error. All three recordings have an associated high reference RR of 28 rpm.

Figure 4.15a shows an example of a good quality PPG recording. The extracted RIFV waveform resembles a periodic sinusoid, which we would expect from a good-quality respiratory waveform. We note that all of the five RR estimation algorithms gave similar estimates of RR, which agreed exactly with the reference measurement.

Figure 4.15b illustrates an example of the first and second sources of error. Here, a transient artefact in the PPG recording (seen between the 5 and 10-second timestamp) causes the PPG peak detector to fail, which results in an artefact in the RIFV waveform.

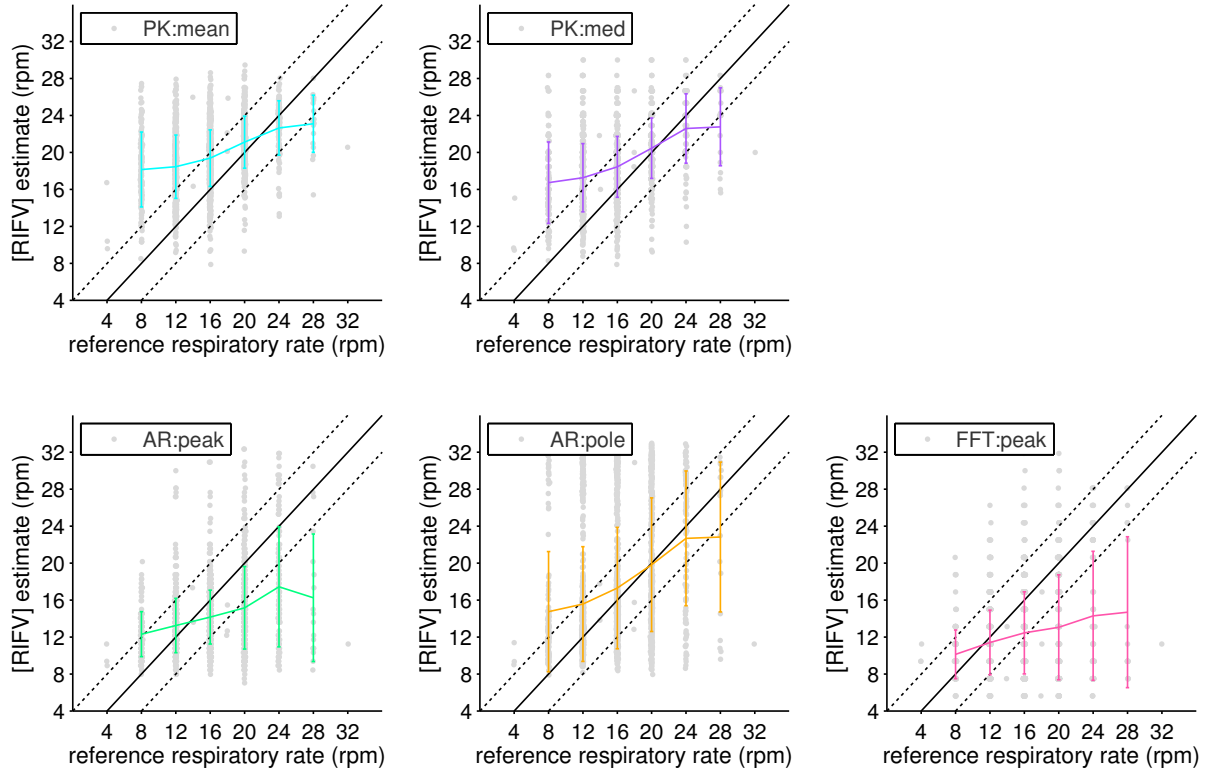
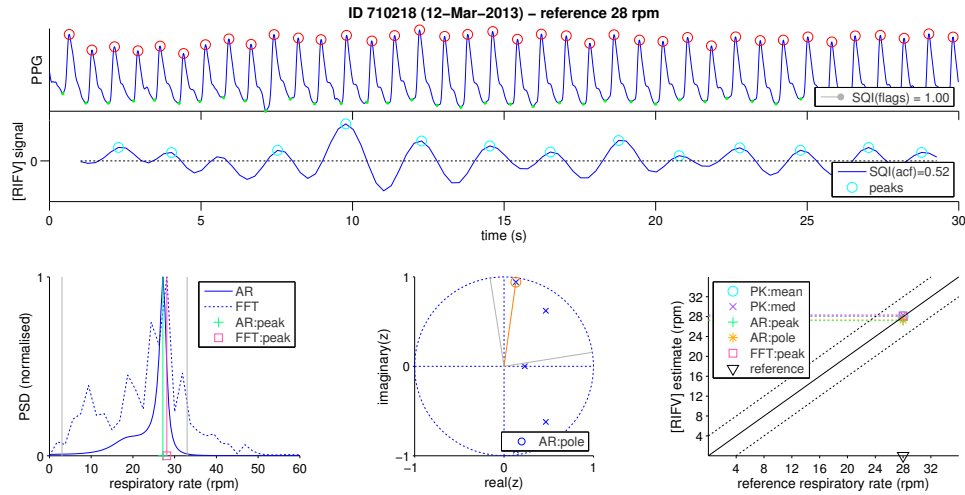


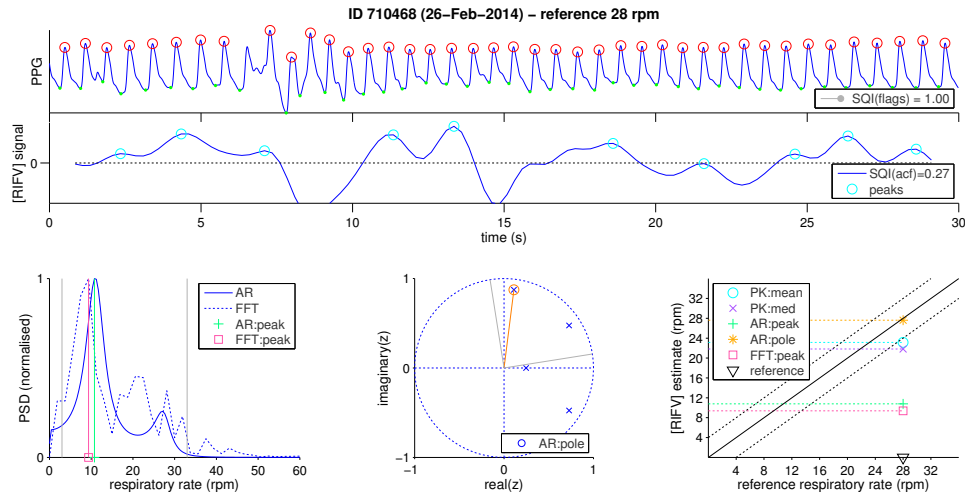
Figure 4.14: Variation of estimated RR from the RIFV waveforms plotted against the associated reference RR (grey dots). The mean (SD) estimated RR at reference RR of 8, 12, 16, 20, 24 and 28 rpm are superimposed (coloured error bars). Estimates within ± 4 rpm of the reference RR fall within the dotted lines.

Table 4.2: Error metrics for the five RR estimation algorithms at selected reference RR for the RIFV waveforms. The results of the best performing algorithm at each reference RR are highlighted in bold.

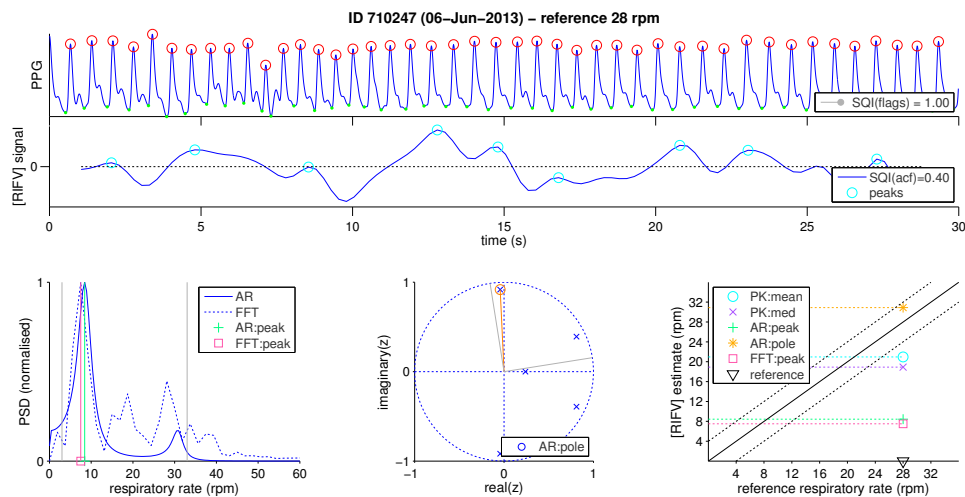
Input signal [<i>RIFV</i>]		Reference respiratory rate (rpm)						
		All	8	12	16	20	24	28
Number of estimates, n		2074	134	596	811	402	104	18
PK:mean	Absolute error (rpm), <i>mean</i>	4.6	10.1	6.5	3.8	2.4	2.3	4.9
	Within ± 4 rpm, %	51	6	24	59	83	83	39
PK:med	Absolute error (rpm), <i>mean</i>	4.1	8.7	5.4	3.2	2.5	2.9	5.3
	Within ± 4 rpm, %	64	16	46	73	84	79	61
AR:peak	Absolute error (rpm), <i>mean</i>	3.6	4.3	2.1	2.7	5.8	8.2	12.0
	Within ± 4 rpm, %	66	52	90	75	32	21	17
AR:pole	Absolute error (rpm), <i>mean</i>	5.1	6.8	4.2	4.7	6.4	6.3	6.7
	Within ± 4 rpm, %	58	46	77	64	30	29	56
FFT:peak	Absolute error (rpm), <i>mean</i>	5.0	2.5	2.9	4.8	7.7	10.2	13.3
	Within ± 4 rpm, %	49	84	67	42	31	27	28



(a) Good quality RIFV waveform



(b) RIFV waveform corrupted by artefacts in the PPG waveform.



(c) RIFV waveform with significant baseline wander

Figure 4.15: Examples of three different RIFV waveforms with the same reference RR of 28 rpm. Each figure (a-c) shows the PPG waveform [top], the RIFV waveform [middle], the associated FFT and AR power spectra [bottom left], pole-zero plot [bottom middle] and the estimated RR plotted against the reference measurement [bottom right].

This ultimately leads to erroneously low estimates of RR from the frequency-based methods AR:peak and FFT:peak.

Figure 4.15c illustrates an example of a respiratory waveform that is corrupted by low frequency baseline fluctuations, which could result from a low degree of RSA in the signal. Although the PPG recording appears to be of a good quality, the extracted RIFV signal contains low frequency fluctuations (baseline wander). In the associated frequency spectra, the low frequency components are seen to be greater than the peak at the expected respiratory frequency of about 28 rpm. To attenuate frequencies below the HF band, a high-pass filter with a higher cut-off frequency would be required. However, when the universal application of a HF filter with a higher cut-off frequency was attempted, low RR values were estimated less accurately. For PK:mean and PK:med, we also note the effect of not including a threshold in the peak detector, which allows peaks to be detected in spite of the low-frequency amplitude fluctuations.

4.4.3 Assessment of RR estimation algorithms

Up to this point, our discussion has focused on comparing RR estimates obtained from the accelerometry and PPG signals across a range of reference RR. In this section, we shift the focus to examine the accuracy of each of the five RR estimation algorithms in more detail.

Figures 4.10 and 4.13 show that the largest overall errors were associated with the AR:pole and FFT:peak algorithms. The AR:pole algorithm frequently produces estimates of RR that are too high. The errors associated with this algorithm are particularly large at low RR (< 16 rpm), as shown in Figures 4.11 and 4.14, for the A_P and RIFV waveforms, respectively (the AR:pole estimates are highlighted in orange). This may indicate that the estimation of pole magnitude is particularly sensitive to high-frequency noise. This is demonstrated by the poor-quality example in Figure 4.12b; here, the noise in the waveform causes selection of the high-frequency pole (based on pole magnitude) rather than the low-frequency pole that is close to the expected respiratory frequency (at an

angle of approximately 45°). For estimation of high RR, on the other hand, the AR:pole estimates are more accurate. As an example, consider the three examples of high RR presented in Figure 4.15, where the AR:pole estimates appear to be more accurate than the estimates from any of the four other algorithms.

The FFT:peak algorithm gives the most accurate estimates at low RR. However, because the FFT produces a quantised frequency spectrum, the estimates are less precise. The quantisation can be seen in Figures 4.11 and 4.14 (the FFT:peak estimates are highlighted in pink) in the form of a separation of points on the y-axis corresponding to the quantisation of the FFT of approximately $\Delta f = 2 \text{ rpm}$ ⁶.

Of particular interest, Tables 4.1 and 4.2 show a similar pattern of estimation accuracy at low and high RR that applies across the accelerometry and PPG-derived waveforms. At low RR ($\leq 16 \text{ rpm}$), the frequency-based AR:peak method is particularly accurate. The main advantage of AR:peak over FFT:peak is that the estimates are not quantised but instead estimated from a continuous spectrum, thus resulting in more precise estimates. At high RR ($> 16 \text{ rpm}$), PK:mean is the most accurate. In the case of the A_P signal, this could indicate that PK:mean is less affected by motion artefacts, which we have shown are likely to introduce low-frequency features in the respiratory waveform (Section 4.4.1). In the case of the RIFV signal, we showed that PK:mean was better suited to identify breaths in the presence of baseline wander (Section 4.4.1). However the use of a peak detector without a threshold to detect peaks may cause RR to be over-estimated at low RR. In the next section we will consider combining the AR:peak and PK:mean algorithms in order to obtain single RR estimates from the A_P and RIFV signals, respectively.

4.4.4 Combining the RR estimation algorithms

To obtain single measurements of RR from the A_P and RIFV signals, we computed the mean of the AR:peak and PK:mean estimates for each recording. We will refer to these combined estimates as RR:mean [A_P] and RR:mean [RIFV]. The RR:mean estimates

⁶The quantisation (Δf) of the FFT was discussed in Section 4.3.6

against the reference RR are plotted in Figure 4.16. The associated error metrics at a set of reference RR between 8 and 28 rpm are summarised in Table 4.3.

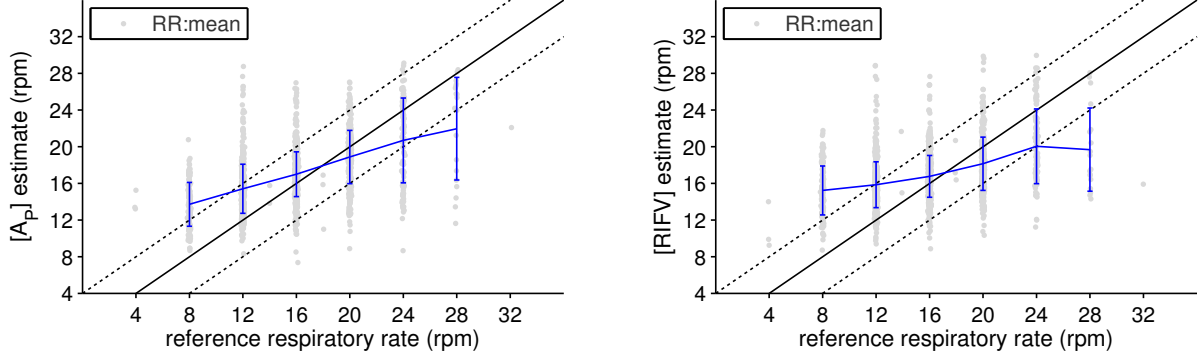


Figure 4.16: Variation of RR:mean estimates plotted against the reference RR (grey dots) for the A_P [left] and RIFV [right] signals. The mean (SD) estimated RR at reference RR of 8, 12, 16, 20, 24 and 28 rpm are superimposed (coloured error bars). Estimates within ± 4 rpm of the reference RR fall within the dotted lines.

Table 4.3: Error metrics associated with RR:mean at selected reference RR.

Input signal [A_P]		Reference respiratory rate (rpm)						
		All	8	12	16	20	24	28
Number of estimates, n		2163	141	617	845	420	112	19
RR:mean	Absolute error (rpm), <i>mean</i>	2.9	5.7	3.5	2.0	2.5	4.4	6.1
	Within ± 4 rpm, %	73	26	63	89	78	54	53
Input signal [$RIFV$]		Reference respiratory rate (rpm)						
		All	8	12	16	20	24	28
Number of estimates, n		2074	134	596	810	403	104	18
RR:mean	Absolute error (rpm), <i>mean</i>	3.2	7.2	3.9	1.8	2.8	4.9	8.3
	Within ± 4 rpm, %	69	11	56	92	74	39	22

Compared to the original methods AR:peak and PK:mean, RR:mean results in a lower overall mean absolute error for the A_P and RIFV signals. For good-quality signals, we have shown that the RR estimates from the two methods were likely to agree; consequently, RR:mean is close to both the AR:peak and PK:mean estimates. For poor-quality signals, we have shown that the AR:peak and PK:mean estimates were less likely to agree. In this case, RR:mean lies between these two measurements. Consequently, at low RR, the RR:mean errors are lower than the error associated with PK:mean, but

slightly higher than the error of AR:peak. At high RR, we observe the opposite: the RR:mean error is lower than the error associated with AR:peak but higher than the error associated with PK:mean. The effect of this is to pull the low and high RR estimates towards a mean value. In spite of this, the RR:mean estimates do *overall* perform better than any of the previous methods across the full range of reference RR.

Figure 4.17 shows the distributions of the RR:mean $[A_P]$ and $[RIFV]$ measurements compared to the distribution of the reference measurements. From this we make two key observations. Firstly, while the distribution of the reference measurement is quantised to 4 rpm, the electronic measurements provide continuous values of RR. Secondly, the distributions of the reference measurements appears to be slightly skewed towards lower values of RR compared to the distribution of the electronic RR estimates. This may be because the midwives tend to only count *completed* breaths (i.e. if the true respiratory rate is 17, the corresponding reference RR would be 16 because the midwife only counts four completed breaths during the 15-second period), something which results in a general under-estimation of the *true* RR.

We have shown that it is possible to obtain two single estimates of RR independently from two different sources, namely the accelerometry and PPG data. Figure 4.18 shows a comparison of the two methods, from which we found that 1,622 (78.3%) of the paired measurements agreed to within ± 4 rpm (indicated by the black dotted lines in the figure). To formally compare the two methods of measurement we use a Bland Altman plot.⁷ From the plot, we see that the width of the 95% confidence interval of the differences between measurements (roughly equivalent to $b \pm 2SD$, where b is bias and SD is the standard deviation) is approximately 14 rpm. The agreement between the methods is therefore not sufficient for the two methods to be used interchangeably.

We could look for a more sophisticated data fusion method to combine the two methods. However, we know that errors may arise because of artefacts in the measurement process that result in respiratory signals of poor quality. Consequently, investigating the

⁷The Bland-Altman plot was introduced in Section 3.4.3 for comparison of temperature measurements

use of an SQI to combine measurements from the accelerometer and PPG signals is an approach that is more likely to be beneficial. In the next section we will investigate whether the estimation accuracy can be improved by means of an SQI.

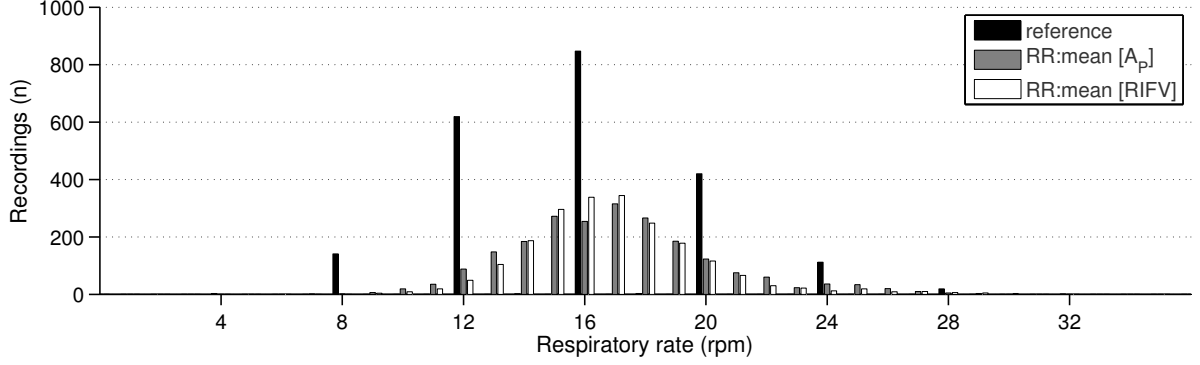


Figure 4.17: Distribution of reference and electronic RR estimates. Bars representing the reference, $RR:mean[A_P]$ and $RR:mean[RIFV]$ are shown in black, grey and white, respectively.

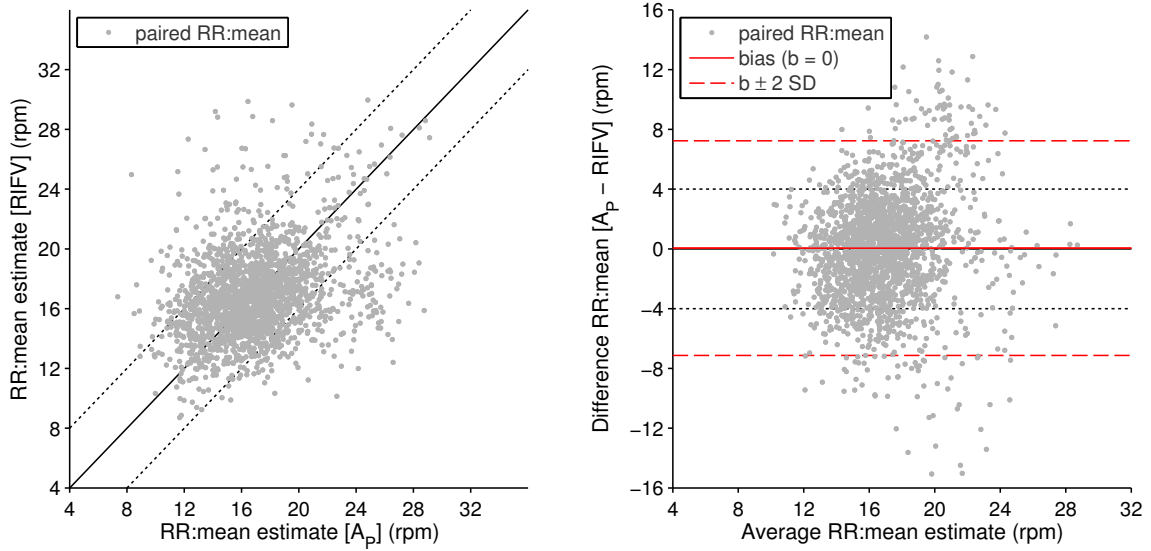


Figure 4.18: Comparison of the accelerometry and PPG derived RR estimates. Estimates within ± 4 rpm of the reference RR fall within the black dotted lines. **[left]** $RR:mean[RIFV]$ plotted against $RR:mean[A_P]$ (grey dots). **[right]** Bland-Altman plot showing the difference between $RR:mean[A_P]$ and $RR:mean[RIFV]$ plotted against their average. The bias (b) and $b \pm 2 SD$ are showed by red solid and dashed lines, respectively.

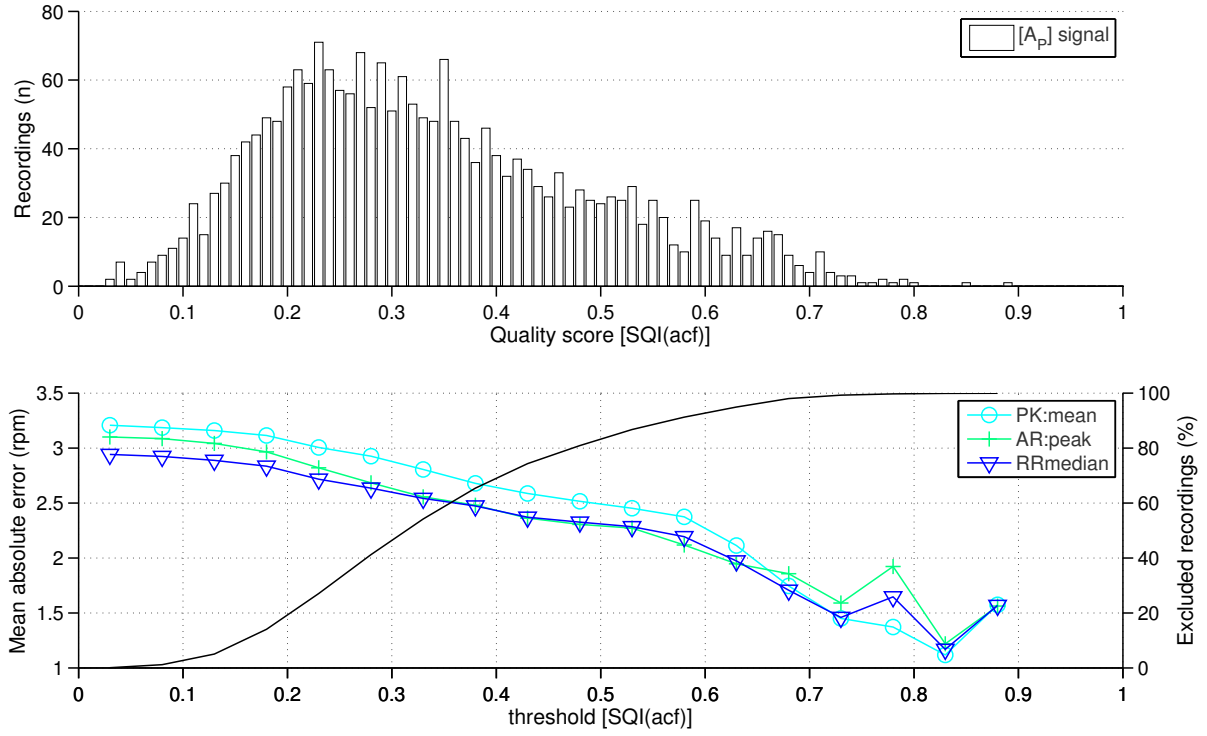
4.5 Improving estimation accuracy using a signal quality index

We hypothesise that the accuracy of RR estimation could be improved by only considering good-quality respiratory waveforms (where good quality is indicated by an SQI closer to 1). The two SQIs described in Section 4.3.7 gave similar results. Hence, we will only report the results obtained using the SQI based on autocorrelation, here referred to as SQI(acf). Because the RR:mean was computed by combining the AR:peak and PK:mean methods, we will assess the effect of accounting for signal quality on the accuracy of these two methods, as well as on the accuracy of the final RR:mean estimates.

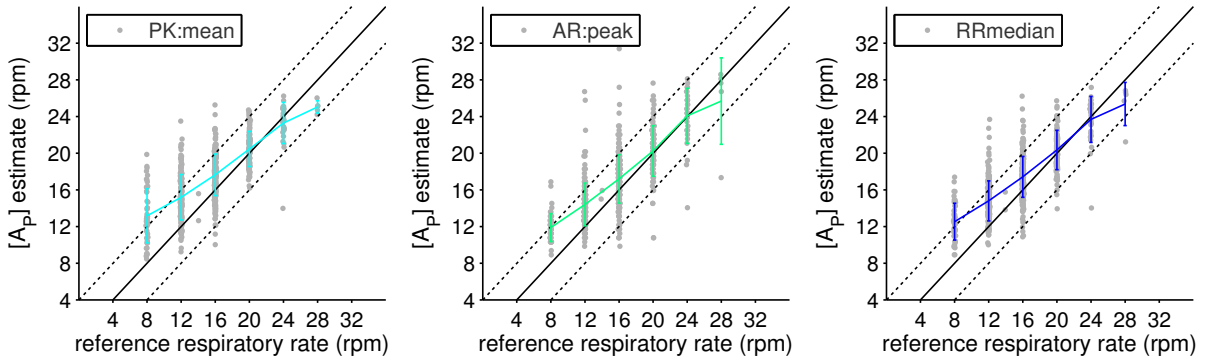
4.5.1 Effect of signal quality for accelerometry waveforms

For the A_P estimates, errors primarily arise from noise associated with deep breathing and from movement artefacts. This results in over- and underestimation of respiratory rate, respectively. Figure 4.19a shows the distribution of SQI scores for all 2,167 A_P signals and the variation in mean absolute error as the threshold of SQI for inclusion in the analysis is increased. As expected, the error decreases with an increasing threshold value of SQI, as this corresponds to exclusion of an increasing number of poor-quality waveforms.

Figure 4.19b and Table 4.4 show the RR estimates and improvement in mean absolute error obtained when only A_P waveforms with $\text{SQI}(\text{acf}) \geq 0.4$ are included. This threshold was chosen to illustrate how a threshold for acceptable waveform quality based on SQI contributes to reducing the number of outlier RR estimates. With a limit for inclusion at $\text{SQI}(\text{acf}) \geq 0.4$, a total of 1,506 (70%) recordings are excluded. For all three methods, the greatest improvement in mean absolute error is seen for recordings with a high RR. Based on our previous finding of a link between underestimation of RR and movement artefacts, we can conclude that the autocorrelation-based SQI is useful for identification of movement artefacts.



(a) **[top]** Distribution of the SQI(acf) quality score for all the recorded A_P waveforms **[bottom]** Variation in mean absolute error (coloured lines) with SQI threshold. The cumulative distribution of quality scores is superimposed (black line).



(b) Variation of estimated RR from the A_P waveforms ($\text{SQI}(\text{acf}) \geq 0.4$) plotted against the associated reference RR (grey dots). The mean (SD) estimated RR at reference RR of 8, 12, 16, 20, 24 and 28 rpm are superimposed (coloured error bars).

Figure 4.19: Effect of SQI(acf) on RR estimates derived from accelerometry data

Table 4.4: Improvement in mean absolute error (improvement if < 0) for RR estimates when accounting only for recordings of the A_P waveforms with $SQI(acf) \geq 0.4$.

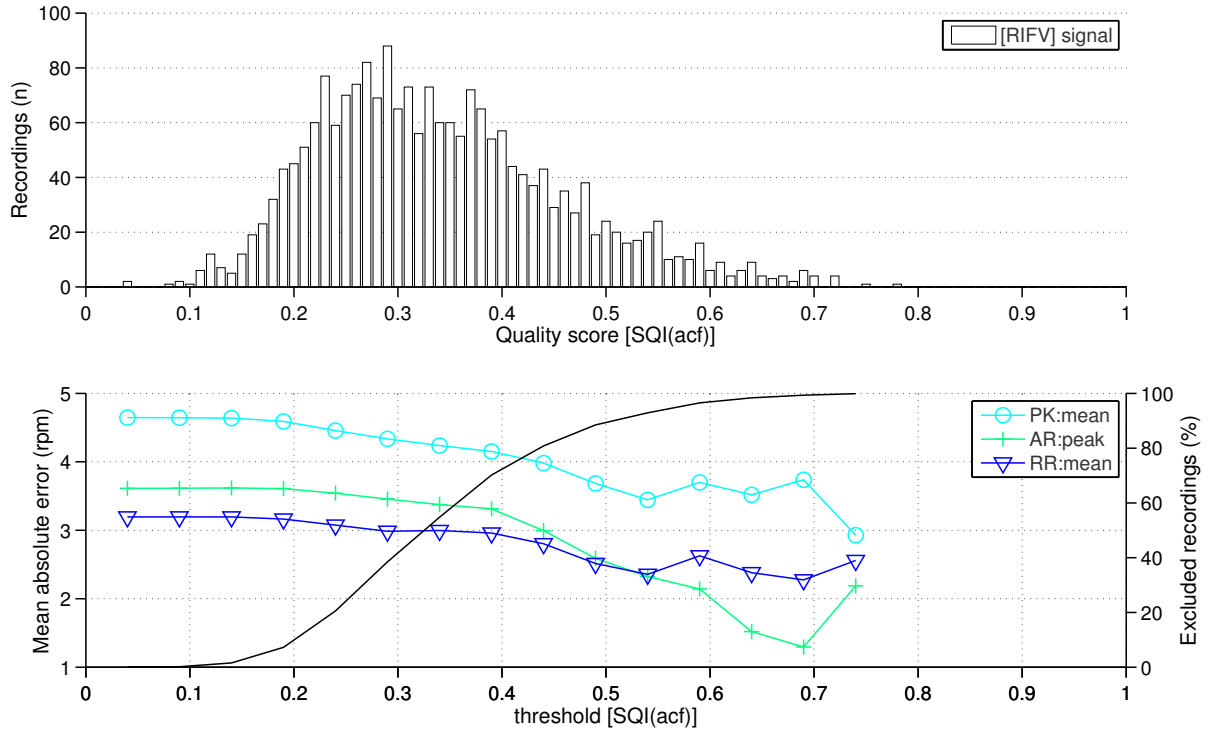
Input signal [A_P]	Reference respiratory rate (rpm)						
$SQI(acf) \geq 0.40$	All	8	12	16	20	24	28
Number of estimates, n	657	49	203	263	105	30	5
% of original recordings	30	35	33	31	25	27	26
Improvement in absolute error (rpm), <i>mean</i>							
PK:mean	-0.5	-1.4	-0.8	-0.2	-0.3	-2.1	-3.2
AR:peak	-0.7	-1.0	-0.5	-0.0	-1.5	-3.7	-4.2
RR:mean	-0.5	-1.2	-0.6	0.0	-0.7	-2.7	-3.5

4.5.2 Effect of signal quality for PPG-derived waveforms

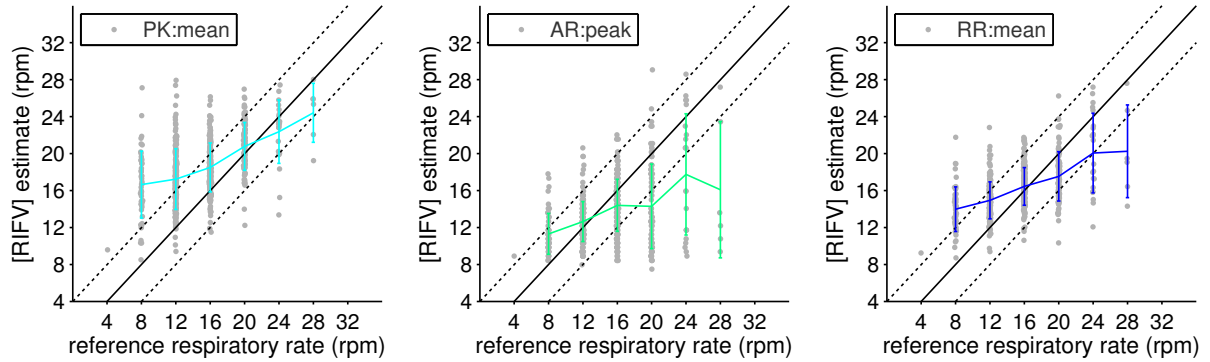
Artefacts in the RIFV waveform primarily result either from artefacts in the PPG recording or from the dominance of low-frequency components present in the HRV frequency spectrum. As artefacts in the PPG signal lead to artefacts in the derived RIFV waveforms, we focused on assessment of the respiratory waveform using $SQI(acf)$.

Figure 4.20a shows the distribution of SQI scores across all 2,074 RIFV recordings and the variation in mean absolute error as the threshold of SQI for inclusion in the analysis is increased. As for the A_P signal, an increase in the threshold for acceptable SQI scores results in a greater proportion of low-quality waveforms being excluded and a decrease in mean absolute error. We also note that the RIFV signals generally are of lower quality compared to the A_P signals (for comparison, the distribution of SQI for A_P signals was shown in Figure 4.19a). This is indicated by the overall lower values of SQI associated with RIFV signals. The trend of a decreasing error with increasing SQI threshold is the same for all three algorithms. With respect to the different RR estimation algorithms, PK:mean is associated with a higher mean error.

Figure 4.20b and Table 4.5 show the RR estimates and improvement in mean absolute error resulting from exclusion of RIFV waveforms with $SQI(acf) < 0.4$. Similar to the results for the A_P signal, the exclusion of low-quality recordings results in fewer RR estimates outside the ± 4 rpm limits of agreement.



(a) **[top]** Distribution of the SQI(acf) quality score for all the recorded RIFV waveforms **[bottom]** Variation in mean absolute error (coloured lines) with SQI threshold. The cumulative distribution of quality scores is superimposed (black line).



(b) Variation of estimated RR (RIFV waveforms, $SQI(acf) \geq 0.4$) plotted against the associated reference RR (grey dots). The mean (SD) estimated RR at reference RR of 8, 12, 16, 20, 24 and 28 rpm are superimposed (coloured error bars). Estimates within ± 4 rpm of the reference RR fall within the dotted lines.

Figure 4.20: Effect of SQI(acf) on RR estimates derived from PPG data

Table 4.5: Improvement in mean absolute error (improvement if < 0) for RR estimates when accounting only for recordings of the RIFV waveforms with $SQI(acf) \geq 0.4$.

Input signal [<i>RIFV</i>]	Reference respiratory rate (rpm)						
$SQI(acf) \geq 0.40$	All	8	12	16	20	24	28
Number of estimates, n	565	54	182	204	96	22	6
% of original recordings	27	40	31	25	24	21	33
Improvement in absolute error (rpm), <i>mean</i>							
PK:mean	-0.5	-1.5	-1.2	-0.8	-0.2	0.3	-1.3
AR:peak	-0.3	-1.0	-0.3	-0.2	0.5	-0.8	-0.1
RR:mean	-0.3	-1.3	-0.9	-0.2	0.2	-0.2	-0.6

In comparison to the A_P results, the improvement in mean absolute error after excluding RIFV signals with $SQI(acf) < 0.40$ is smaller. In fact, for higher RR, no improvement (or even a slight increase) in mean absolute error is observed. This is likely to be because the periodic, low-frequency baseline-wander in the RIFV signal occasionally masks the RSA (HF) frequencies and leads to RR being under-estimated. The periodic low-frequency fluctuations causes both a high autocorrelation value and a prominent peak in the frequency spectrum, which can be mistaken for the respiratory peak. Such artefactual frequency components are generally very hard to identify using the autocorrelation-based SQI (or any SQI based on frequency characteristics of the signal).

4.6 Computation of a single electronic estimate

We have shown that for both the A_P and RIFV waveforms, a higher SQI is associated with a more accurate RR estimate. To derive reference ranges of changes in RR with gestational age we require a single electronic RR estimate from each recording session. To obtain this, we 'fused' the RR:mean [A_P] and RR:mean [RIFV] estimates using the SQI according to the following rules:

1. If paired observations (from the same recording session) of RR:mean [A_P] and RR:mean [RIFV] existed, the observation with the highest SQI was selected.
2. If, for a given recording session, only one waveform existed, this single estimate was selected.

The resulting 'fused' RR:mean estimates are shown in Figure 4.21 and the associated errors at each reference RR are reported in Table 4.6. In the cases of paired measurements, the A_P estimates were selected for 1,075 (51.8%) of the recording sessions and the RIFV estimates were selected for the remaining 999 (48.2%) recording sessions. The slightly higher proportion of RR:mean [A_P] estimates was to be expected, given that A_P waveforms were associated with a higher SQI. Overall, the 'fused' estimate showed a very slight improvement (in the form of reduced error in comparison to the reference measurement) compared to using the RR:mean [A_P] estimate only.

In the next chapter, we will use this 'fused' electronic estimate of RR to compute reference ranges for the distribution and change in RR across gestational age.

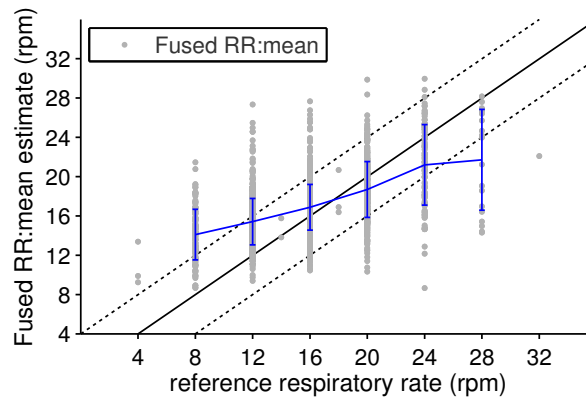


Figure 4.21: Variation of the 'fused' RR:mean estimates plotted against the reference RR (grey dots). The mean (SD) estimated RR at reference RR of 8, 12, 16, 20, 24 and 28 rpm are superimposed (coloured error bars). Estimates within ± 4 rpm of the reference RR fall within the dotted lines.

Table 4.6: Error metrics for the 'fused' RR:mean estimates at selected reference RR.

Input signal [<i>fused</i>]		Reference respiratory rate (rpm)						
		All	8	12	16	20	24	28
Number of estimates, n		2165	141	619	845	420	112	19
RR:mean	Absolute error (rpm), <i>mean</i>	2.9	6.1	3.5	1.9	2.5	3.9	6.3
	Within ± 4 rpm, %	74	21	63	91	80	59	47

4.7 Limitations and future improvements of respiratory estimation methods

The greatest limitation of this study is likely to have been the quality of the reference measurement, which was associated with a quantisation error of up to ± 4 rpm. Previous studies have reported using continuous signals for validation, such as an impedance pneumography waveform [116,117] or a validated capnography waveform [106,151]. However, the subjects of these studies were patients confined to beds in the hospital rather than out-patients. In the setting of our study, the application of such measurement techniques would not be feasible.

Instead, we propose two alternative methods for obtaining a more accurate reference measurement. The first method is the use of a 60-second count for validation of RR, which would be associated with a much smaller quantisation error of ± 1 rpm. The second method is to record the exact timing of each breath, and subsequently compute the reference RR from time intervals between breaths. This could be implemented for example by tapping on the touch screen of a tablet computer or mobile device at a fixed point in the breathing cycle (e.g. at the peak of each inhalation), as proposed by Karlen *et al.* [152]. The knowledge of the timestamps of each breath would further enable assessment of the regularity of breathing, which could be useful for a more detailed comparison of the algorithms.

A better reference RR would also enable further tuning of the RR estimation algorithms. The five RR estimation algorithms that we investigated are relatively simple. For the extraction of RR from the PPG in particular, many more sophisticated algorithms have recently been described in the literature, of which the majority are based on AR models. Some of these algorithms were tested on our dataset but without achieving notable improvements in the accuracy of RR estimation. This included the two-step pole search algorithm described by Shah *et al.* [145] and an algorithm described by Mason *et al.* [144] where all poles within 95% magnitude of the dominant pole are identified and the lowest

frequency is taken to be the RR. However, given that we reported mean absolute errors that were less than the potential quantisation error, any improvement may have been masked by the imprecision of the reference standard. This investigation should therefore be repeated with a more robust reference standard.

Finally, our findings indicated that the accuracy of RR estimation could be improved by accounting for signal quality. An assessment of signal quality would be most useful if conducted at the time of measurement. This could for example be done by implementation of an SQI in the tablet computer and smartphone software. Assessment of signal quality at the point of data collection would have two advantages; firstly, a signal quality index could be used to guide inexperienced users with regards to positioning of the probe and motion sensors; secondly, it would enable a new recording to be made straight away if the first recording was of unsatisfactory quality.

Chapter 5

Design of vital sign reference ranges

In this chapter we will describe the development and assessment of statistical models for design of gestation-specific reference ranges for observations of vital signs from the antenatal period. In order to decide on the appropriate method for design of reference ranges, we will assess several statistical models. Using HR as an example, we will highlight the assumptions, merits and limitations of each modelling approach and compare the fit of the models to the 4P data. At the end of the chapter we will present reference ranges for each of the six vital signs (HR, SpO₂, SBP, DBP, Temp and RR) and compare the vital sign patterns of samples of women with different characteristics.

5.1 The use of reference ranges in obstetric care

In obstetric care, reference ranges are used to assess a variety of indicators of maternal and child health. Of particular importance is the use of reference ranges for the assessment of normal fetal and child growth [70, 153]. Other applications include reference ranges for normal changes with gestational age of cervical length [154] and SBP and DBP [78, 79, 84, 155, 156]. As discussed in Chapter 2, we have not been able to identify previous attempts at constructing reference ranges for the other vital signs (HR, SpO₂, Temp and RR).

Studies to develop charts typically follow either a *descriptive* or *prescriptive* approach. A descriptive approach is used to make *reference ranges* that are built on data from a specified target population for whom the reference ranges will apply. Typically, these are based on an unselected group of women and therefore only a minimum number of exclusion criteria should be applied to exclude women with risk factors known to affect optimal health. A prescriptive approach is, on the other hand, used to make *standard charts*. These are based on a highly selected group of women, with strict inclusion and exclusion criteria, for example based on factors such as age, BMI and medical history. Such charts aim to define what the optimal characteristics of a target population should be [157]. The 4P study is based on a descriptive approach, where a minimum of exclusion criteria were set to define a target population. The selection criteria for the 4P analysis population were presented in Section 3.3.

5.2 Method identification

The selection of an appropriate model for derivation of reference ranges or *centiles* of a continuous response variable, such as a vital sign measurement, with time, is not trivial. Table 5.2 summarises the approaches developed by other studies identified in the systematic review in Chapter 2 to model centiles of BP in pregnancy. These represent only a small number of the models available to model centiles. A particularly popular application for centile models is for the study of patterns of fetal growth with gestational age or child growth with age. The Multicentre Growth Reference Study Group (MGRS) [90] has published a comprehensive review of the available methods. According to the MGRS, in order to select an appropriate centile model three key questions must be addressed: *(i)* what assumptions (if any) can we make about the shape of the distribution of observations at each time point?; *(ii)* which smoothing technique is appropriate in order to model the trend across time?; and *(iii)* how does the study design impact our choice of method? In what follows we will discuss each of these three aspects in turn.

Table 5.1: Models from the literature used for computation of centiles of BP in pregnancy

Publication	Study type (measurements per woman)	Modelling approach and shape parameterisation	Model fit assessment	Model covariates
Ochsenbein-Kolble <i>et al.</i> (2004) [78]	Cross sectional, $n=1$	LMS model with time modelled using cubic splines.	Not reported	Ethnicity
Thompson <i>et al.</i> (2009) [156]	Longitudinal, <i>mean</i> $n=12.2$	LMS model with time modelled using fractional polynomials.	Z-scores; visual comparison to empirical distributions	BMI
Bakker <i>et al.</i> (2010) [155]	Longitudinal, $n=1-3$	Multilevel model with time modelled using fractional polynomials. Unclear whether random effects were included.	Not reported	Smoking status
McDonald-Wallis <i>et al.</i> (2012) [84]	Longitudinal, <i>median</i> $n=10$ (<i>IQR</i> $9-11$)	Multilevel model with random effects. Time modelled using linear splines. Time to delivery included as outcome variable.	Not reported	Hypertensive disorders, BMI, maternal age, parity, smoking, maternal education
McDonald-Wallis <i>et al.</i> (2015) [79]	Longitudinal, <i>median</i> $n=10$ (<i>IQR</i> $9-11$)	Multilevel model with random effects. Time modelled using restricted cubic splines.	BIC [†] ; visual assessment of modelled and empirical values	BMI, smoking status, parity
[†] The Bayesian Information Criterion (BIC) is defined in Appendix B.3.6				

Firstly, a distributional assumption may be advantageous because it allows for estimation of extreme centiles when observations of extreme values are scarce. However, the available data must be carefully examined to ensure that the assumed distribution is reasonable. Assuming a distribution also enables defining a probability of an observation at a specific point in time. Central to the interpretation of centiles of a normal distribution is the Z score given by Equation 5.1. This is a standardised measure of the deviation of a value X from the mean μ , in units of standard deviations (SD) σ . For a normal distribution, Z scores can be converted to percentile values using standard conversion tables.

$$Z = \frac{X - \mu}{\sigma} \quad (5.1)$$

For BP, previous studies conflict with regards to the appropriateness of assuming a normal distribution of BP conditional on gestational age. While Macdonald-Wallis *et al.* [79, 84] assumed a normal distribution in their analysis, Ochsenein-Kolble *et al.* [78] and Thompson *et al.* [156] applied a transformation to account for skewness in the distributions of BP. The normality assumption must therefore be investigated and confirmed for all six vital signs. While values of SpO₂ across a population have a one-sided distribution (as values of SpO₂ cannot exceed 100%) [158], we will investigate whether the observations of mean SpO₂ in our sample of pregnant women are normally distributed.

Secondly, the smoothing function describes the trend across time. A lack of smoothing may result in irregularities that reflect features of a particular sample (such as sampling variability across ages, outliers and measurement variability) rather than the true underlying change. Over-smoothing, on the other hand, may flatten the true physiological trend so that peaks and troughs in the pattern of change are missed. The MGRS differentiates between methods that use a continuous formulation of time (e.g. polynomial expression) or methods based on a bin-and-smooth approach where measurements are grouped in time (e.g. a moving average) [90]. As defined in Chapter 2, we refer to the pattern of vital sign values as they evolve over time as the vital sign *trajectory*.

Finally, study design must be considered in order to correctly handle dependencies in the data. For longitudinal studies, observations from the same individual will be correlated across time. The effect of accounting for correlations between measurements is expected to be particularly significant for the estimation of extreme centiles [90]. The importance of accurate estimation of extreme centiles should not be underestimated, as these form the basis for clinical decision making and identification of abnormality.

Based on these considerations, we identified candidate models from the literature and from previous studies, with emphasis on choosing models to conform with different aspects of the MGRS recommendations. The models selected for our investigation are summarised in Table 5.2. Where several formulations of the models existed, we chose the simplest (most commonly implemented) formulation. Two methods, the multilevel model and kernel regression, were introduced in Chapter 2. The Gaussian process is a method from the engineering literature which is commonly applied to regression (and prediction) problems [159]. The LMS method has been recommended for estimation of centiles when a normal distribution of the response cannot be assumed [90, 160]. Gaussian process regression and the LMS method will be described in greater detail below.

Table 5.2: Comparison of different centile estimation methods

Model	Multilevel model	Kernel regression	LMS method	Gaussian process
Distribution assumption	Normal distribution	No distribution assumed	Flexible specification	Normal distribution
Smoothing method	Polynomial (conventional or fractional) or splines fitted to individual trajectories	Non-parametric smoothing parameter h	Polynomial (conventional or fractional) or splines fitted to population trajectory	Non-parametric smoothing using autocorrelation function length-scale parameter
Yields equation for prediction	Yes	No	Yes	No
Independence assumption for multiple measurements	Dependencies accounted for by random terms	Data assumed independent	Data assumed independent	Data assumed independent

5.2.1 Gaussian processes

From Rasmussen and Williams [159], a Gaussian process is a generalisation of the Gaussian probability distribution which describes a distribution over *functions*. It is defined by a mean function $m(\mathbf{x})$ and a covariance function $K(x_i, x_j)$ according to Equation 5.2.

$$f(\mathbf{x}) \sim \mathcal{GP}\left(m(\mathbf{x}), \mathbf{K}(x_i, x_j)\right) \quad (5.2)$$

The mean and covariance functions may take different forms and should be selected according to the characteristics of the data. Here, we consider a formulation of the Gaussian process where $m(\mathbf{x})$ does not take any specific form and so we set $m(\mathbf{x}) = 0$. The covariance function is a function of the times (x_i, x_j) of any two data points $(f(x_i), f(x_j))$ and encodes our prior knowledge about the behaviour we wish to model. It can take any form, provided that it leads to the production of a positive semidefinite covariance matrix. We will consider the squared-exponential covariance function¹, given by Equation 5.3, which is perhaps the most popular choice for many practical applications.

$$\text{Cov}(f(x_i), f(x_j)) = K(x_i, x_j) = \sigma_f^2 \exp\left(-\frac{1}{2\lambda}|x_i - x_j|^2\right) \quad (5.3)$$

Here, $\theta = \{\sigma_f^2, \lambda\}$ are hyperparameters. The function variance σ_f^2 defines the overall variance of the random function and thus the width of the resulting distribution (y -scale) of the function. The length scale λ controls for how long an observation will be correlated to future observations, which can be translated to the degree of smoothness (x -scale). By using the squared-exponential covariance function we make two key assumptions about the data. Firstly, we assume that the data are *autocorrelated*, i.e. data that appear close in time are highly correlated and as data samples become further apart in time correlation decreases. Secondly, we assume *stationarity*, i.e. that only the distance between two time points x_i and x_j is important, and not their absolute values.

¹This is also referred to as the Gaussian kernel function.

In our case, the underlying function $f(\mathbf{x})$, equivalent to the population mean, is unknown. The data are treated as noisy observations $\mathbf{y} = f(\mathbf{x}) + \varepsilon$. The covariance function of the observations $\text{Cov}(y_i, y_j)$ is given by Equation 5.4. We assume that the noise is additive, independent and identically distributed Gaussian noise ε with variance σ_n^2 .

$$\text{Cov}(y_i, y_j) = K(x_i, x_j) = \sigma_f^2 \exp\left(-\frac{1}{2\lambda}|x_i - x_j|^2\right) + \sigma_n^2 \quad (5.4)$$

The aim of Gaussian process regression is to reconstruct the underlying function $f(\mathbf{x})$ by removing the contaminating noise ε . To do this, we compute a weighted average $\bar{\mathbf{f}}_*$ of the observations $(\mathbf{x}, \mathbf{y})$ at a set of test inputs $\mathbf{x}_*$ using Equation 5.5. The resulting covariance of the underlying functions is given by Equation 5.6.

$$\mathbf{f}_* | X, \mathbf{y}, X_* \sim \mathcal{N}(\bar{\mathbf{f}}_*, \text{Cov}(\mathbf{f}_*)),$$

$$\text{where } \bar{\mathbf{f}}_* = \mathbf{K}_*^T (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y} \quad (5.5)$$

$$\text{Cov}(\mathbf{f}_*) = \mathbf{K}_{**} - \mathbf{K}_*^T (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{K}_* \quad (5.6)$$

Here, $\mathbf{K}_* = K(\mathbf{x}, \mathbf{x}_*)$ is the covariance between the training inputs $\mathbf{x}$ and the test inputs $\mathbf{x}_*$ and $\mathbf{K}_{**} = K(x_*, x_*)$ is the prior variance of x_* . As we are primarily interested in assessing the spread of the associated observations $\mathbf{y}_*$, we include the noise variance to obtain the expression for $\text{Cov}(\mathbf{y}_*)$ given by Equation 5.7.

$$\text{Cov}(\mathbf{y}_*) = \mathbf{K}_{**} - \mathbf{K}_*^T (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{K}_* + \sigma_n^2 \mathbf{I} \quad (5.7)$$

To optimise the values of the hyperparameters $\theta = \{\sigma_f^2, \lambda, \sigma_n^2\}$ we use a maximum likelihood approach, as described in Appendix B.5.

Gaussian processes have previously been used by Pimentel [26] and Wong [161] to investigate trajectories of vital signs in individual patients. When applied to repeated measurement data from an entire population, a major concern is that the model treats the data as independent, thereby ignoring the correlation between repeated measurements. This point will be discussed in more detail in the example application in Section 5.3.3.

5.2.2 The LMS method

The LMS method, originally proposed by Cole and Green [160], is used to model centiles of data that exhibits skewness. The observations y (vital sign measurements) are modelled in terms of three curves based on the same independent variable x (gestational age). These three curves are used to define an extended power transformation to correct for skewness, defined by Equation 5.8, and are called the L (controlling the amount of skewness λ), M (controlling the central value μ) and S (controlling the coefficient of variation $\sigma_L = \mu/\sigma$ where σ is the SD) curves.

$$z_{LMS} = \frac{1}{\sigma_L \lambda} \left\{ \left(\frac{y}{\mu} \right)^\lambda - 1 \right\} \quad \text{for } y, \mu, \sigma_L \text{ and } \lambda \neq 0 \quad (5.8)$$

The value z_{LMS} is the Z score of the power transformation. The mathematical relationship between the usual Z score from the normal distribution z_p (as previously defined by Equation 5.1) and LMS percentile p_{LMS} is given by Equation 5.9 [160, 162].

$$p_{LMS}^{th} \text{ percentile} = \mu(1 + \lambda \sigma_L z_p)^{1/\lambda} \quad (5.9)$$

To compute the trend with time, λ , μ , σ are computed at several time points and plotted against time x . The LMS curves are then smoothed with cubic splines using a method of penalized likelihood estimation proposed by Cole and Green [160]. Implementation of the LMS method is computationally complex and specialist software is normally used, such as the **GAMLSS** package in R [163]. Similar to the Gaussian process, the LMS method does not account for correlations between repeated measurements from the same subject.

More recently, Rigby and Stasinopoulos [164] have proposed alternative formulations of the LMS method which accounts for both skewness and kurtosis², however these formulations were not considered for the investigation presented here.

²Kurtosis is a measure of the peakedness of a distribution

5.3 Application of centile estimation models to 4P data

We initially chose three families of models for comparison: empirical centiles smoothed using kernel regression, Gaussian processes and three different formulations of a multilevel model. The Gaussian process and multilevel model are based on an assumption of normally distributed data conditional on gestational age. For vital signs where this assumption appeared less valid, we also considered the LMS method to account for skewness. The choices of model formulation, required for fitting each model to the data, are described below. We will then provide a practical example where we apply each model to compute centiles for heart rate and compare the results. Figure 5.1 shows an overview of the models considered.

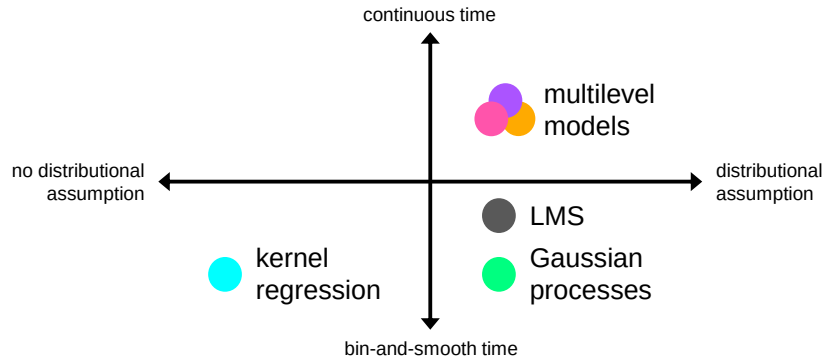


Figure 5.1: Overview of methods used for computation of centiles. The methods have been grouped according to whether they use a bin-and-smooth or continuous formulation of time, and according to whether a distribution is assumed.

5.3.1 Model specification

Model 1 – Empirical centiles smoothed using kernel regression

Following the methodology presented in [165, 166], we calculated empirical centiles for every completed week of gestation. As an example, at 28 weeks of gestation, we computed the median and centiles of all observations recorded within the interval 28^{+0} to 28^{+6} weeks^{+days} of gestation. The empirical centiles were smoothed using weighted kernel regression of order $p = 1$ (Appendices B.1 and B.2), where each data point i was assigned

a weighting w_i equal to the number of underlying observations. The degree of smoothness was controlled by the smoothing parameter h , which is set empirically to provide a good fit to the data, as described in Appendix B.1.

Model 2 – Gaussian process regression

We computed a Gaussian process using a standard formulation of the Gaussian process, where we assumed a mean function $m(\mathbf{x} = 0)$ and a squared-exponential covariance function (Appendix B.5). For the prior values of the hyperparameters $\theta = \{\sigma_f, \lambda, \sigma_n\}$, different combinations of the hyperparameters were tested by randomly selecting values from within the ranges provided below. However, the values of the priors were found not to have a large effect on the optimisation.

$$\sigma_f^2 = \{1 - 10\} \text{ bpm}^2; \quad \lambda = \{1 - 200\} \text{ days}; \quad \sigma_n^2 = \{5 - 150\} \text{ bpm}^2$$

Model 3 – Multilevel models

We defined three different formulations of the multilevel model to investigate the effect of different parameterisations of time with different numbers of fixed (β) and random ($\mathbf{b}_i$) effects. For all three models, we defined two levels; observations (level 1, subscript j) within subjects (level 2, subscript i). To facilitate interpretation of the model intercept, we used a transformation of the time variable $t_{ij} = x_{ij} - 10$, where x_{ij} is the time of the observation in weeks of gestation [167]. The advantage of this will become clearer in Section 5.5 but does not impact the resulting curves.

Model 3a – Random intercept model: Time was parameterised in the form of a conventional polynomial of degree P . P was selected by visual inspection of the shape of the individual within-subject trajectories and by comparison of error metrics such as the log-likelihood, AIC and BIC ³. The inclusion of a random intercept b_{0i} allowed us to explore the variation

³Akaike’s Information Criterion (AIC) and the Bayesian Information Criterion (BIC) were introduced in Chapter 2 and are explained in Appendix B.3.6.

in initial vital sign levels (intercepts) between women. However, when $P > 1$, the relative change in vital signs conditional on gestational age, which is described by the higher-order terms, will be the same for every woman in the population. The random intercept model is given by Equation 5.10.

$$y_{ij} = \beta_0 + b_{0i} + \sum_{p=1}^{p=P} \beta_p t_{ij}^p + e_{ij} \quad (5.10)$$

Model 3b – Random intercept and slope model: This model was similar to the random intercept model, with the inclusion of an additional random slope term $b_{1i}t_{ij}$. Consequently, both the initial level (intercept) and velocity of change (slope) in vital signs varied between women. The random intercept and slope model is given by Equation 5.11.

$$y_{ij} = \beta_0 + b_{0i} + \beta_1 t_{ij} + b_{1i} t_{ij} + \sum_{p=2}^{p=P} \beta_p t_{ij}^p + e_{ij} \quad (5.11)$$

Model 3c – Fractional polynomial random intercept model: Fractional polynomials offer greater flexibility than conventional polynomials for modelling the pattern of change with time, as they allow non-integer powers and repetition of powers. The powers are chosen from a set of eight powers $(-2, -1, -0.5, 0, 0.5, 1, 2, 3)$, where a power of zero is the log function $\log(t_{ij})$. Typically, we choose either a single power p_1 or a pair of powers (p_1, p_2) , which produce either a first or second-degree fractional polynomial expression, respectively. In the case of repeated powers $(p_1 = p_2)$, the second term is multiplied by $\log(t_{ij})$ [168]. For some applications, fractional polynomials have been found to offer better fits than conventional polynomials [153, 169]. Because the use of higher powers often results in very small coefficient values, a random component was only included for the intercept, as shown in Equation 5.12. As for Model 3a, this implied that only the initial vital sign level (intercepts) was assumed to vary between women.

$$y_{ij} = \beta_0 + b_{0i} + \beta_1 t_{ij}^{p_1} + \beta_2 t_{ij}^{p_2} + e_{ij} \quad (5.12)$$

For Models 3a-c, we initially assumed the residuals to be uncorrelated and normally distributed with a constant within-subject variance σ^2 so that $\mathbf{R} = \sigma^2 \mathbf{I}$. This assumption was subsequently relaxed by testing alternative formulations of $\mathbf{R}$ to account for autocorrelation and heteroscedasticity⁴, as described in Appendix B.3.3. When $\mathbf{R} = \sigma^2 \mathbf{I}$, the correlation between individuals is determined entirely by the between-subject covariance matrix $\mathbf{D}$, which defines the distribution of the random effects. For all models, $\mathbf{D}$ was unstructured (i.e. without any constraints apart from symmetry and positive definiteness). Model parameters were estimated using maximum likelihood as described in Appendix B.3.4.

5.3.2 Methods for assessment of model fit

We first assessed model fit *within* each model in order to establish the number and values of parameters, using the steps described in the previous section. Secondly, we compared model fit *across* the different models. Model fit was assessed using a combination of metrics and criteria as recommended by MGRS [90]. This included visual comparison of the fitted centiles to the raw data and empirical centiles, the use of quantile-quantile (Q-Q) plots to assess normality [170], and the inspection of the distribution of fitted Z scores (Equation 5.1) across gestational ages to assess for heteroscedasticity. To assess the fit of extreme centiles we examined the 3rd and 97th centiles, in line with the methodology previously developed for the INTERGROWTH-21st project [102, 153].

5.3.3 Comparison of models applied to heart rate data

Figures 5.2a-e show for each model of centile estimation the computed 3rd, 50th and 97th centiles superimposed on the raw HR data and corresponding empirical centiles. To facilitate comparison across the models, the centiles from each of the five models are plotted together in Figure 5.2f. In what follows, we will discuss some key characteristics of each model before we concluding on what the most appropriate model for the data is.

⁴The term heteroscedasticity is used to describe a variance (or standard deviation) that varies with respect to the independent variable (in this case, with gestational age)

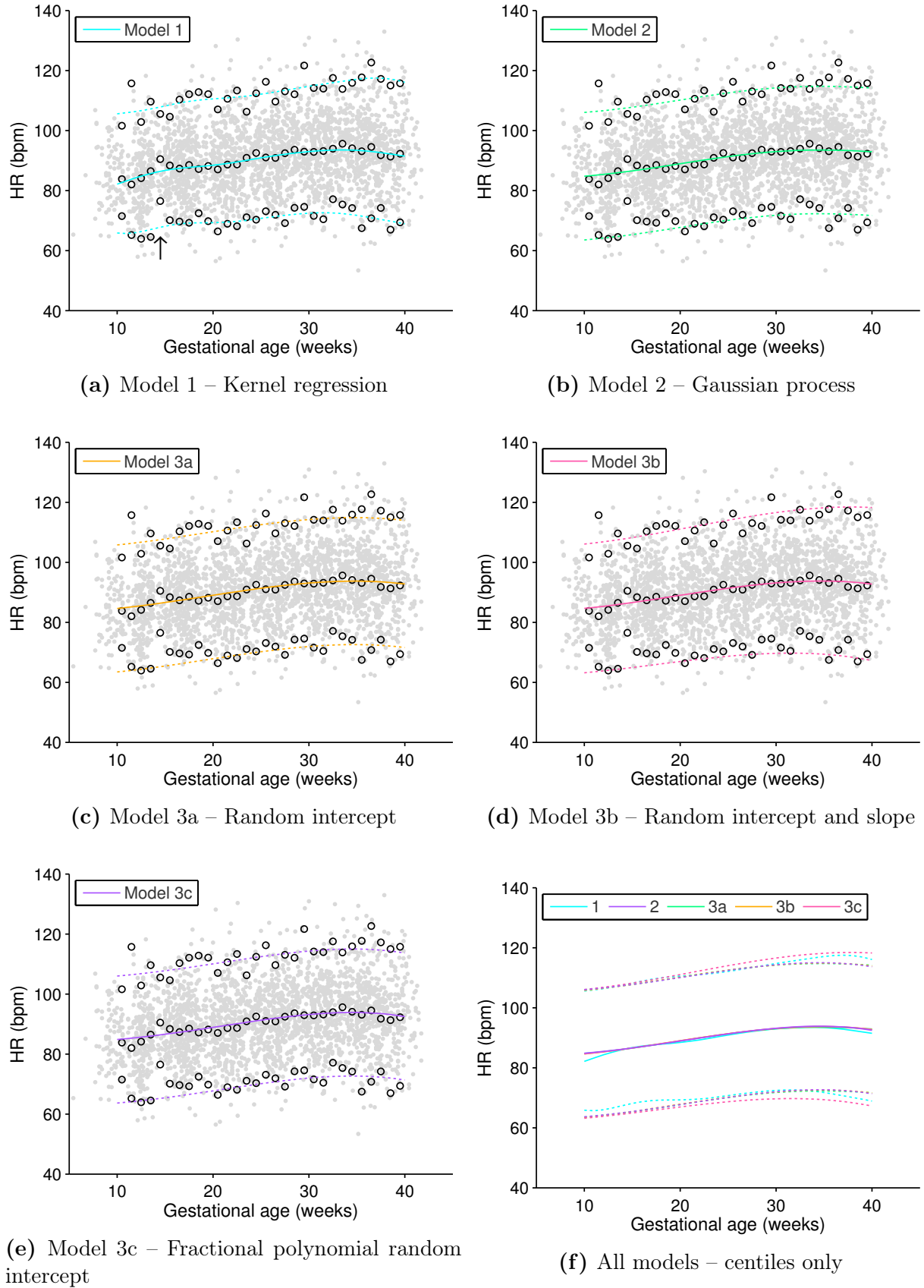


Figure 5.2: Comparison of centile estimation models for HR. Each plot shows the raw HR data (grey dots), empirical centiles and the output centiles from each model (solid line 50^{th} centile, dotted lines 3^{rd} and 97^{th} centiles)

Model 1 – Empirical centiles smoothed using kernel regression

Figure 5.2a show the centiles obtained from Model 1. A value of the smoothing parameter $h = 3$ was found to provide a good fit to the overall trend of the empirical centiles. However, we can see the effect of the variability in the underlying centiles, which occurred at gestational ages with a limited number of data points. An example of such variability can be seen in Figure 5.2a around gestational age 15 weeks, where there is a small bump in the 3rd centile (labelled with an arrow) resulting from an outlying corresponding empirical centile value. The coinciding 'bump' in the Model 1 centile is likely to be a consequence of the bin-and-smooth approach and the lack of a distributional assumption rather than a true feature of the population trend. While the kernel regression smoothing method is advantageous in that it does not involve a distributional assumption, the effect of few data points on the estimation of extreme centiles is a concern. The assessment of an appropriate value of h is also challenging. There are various methods to formally estimate h , however such techniques only provide an initial estimate, and it is advisable to use visual judgement to adjust this value to obtain a suitable balance between smoothness and accuracy for the given data set [171].

Model 2 – Gaussian process regression

Figure 5.2b show the centiles obtained from Model 2. Optimisation over the hyperparameters gave the following values: $\theta = \{\sigma_f^2, \lambda, \sigma_n^2\} = \{4.84 \text{ bpm}, 108.7 \text{ days}, 127.23 \text{ bpm}\}$. The long length scale λ and high value of the noise variance σ_n^2 result in a covariance of the observations $\text{Cov}(y_i, y_j)$ (Equation 5.7) that is nearly constant with gestational age. The Model 2 centiles match well with the empirical centiles from week 10 to 35, after which they appear slightly too narrow indicating that there might be some heteroscedasticity in the data that is not captured by the model. Some degree of heteroscedasticity in the data may be expected, as the effect of pregnancy complications on vital signs typically will increase progressively with gestational age.

Model 3 – Multilevel models

For Models 3a (Figure 5.2c) and 3b (Figure 5.2d), we found that a polynomial of degree $P = 3$ provided the best fit to the data. For Model 3c (Figure 5.2e), the best-fitting fractional polynomial was of order $\{p_1, p_2\} = \{3, 3\}$. For all three models, the assumption of uncorrelated and normally distributed residuals with a constant variance σ^2 proved reasonable⁵. To further investigate the validity of the normality assumptions, we constructed Q-Q plots of the standardised observations y_{ij} and estimated residuals e_{ij} using Equation 5.1. Figures 5.3a-b shows the Q-Q plots for Model 3a and 3b. As the points lie approximately along straight lines, the normality assumptions hold.

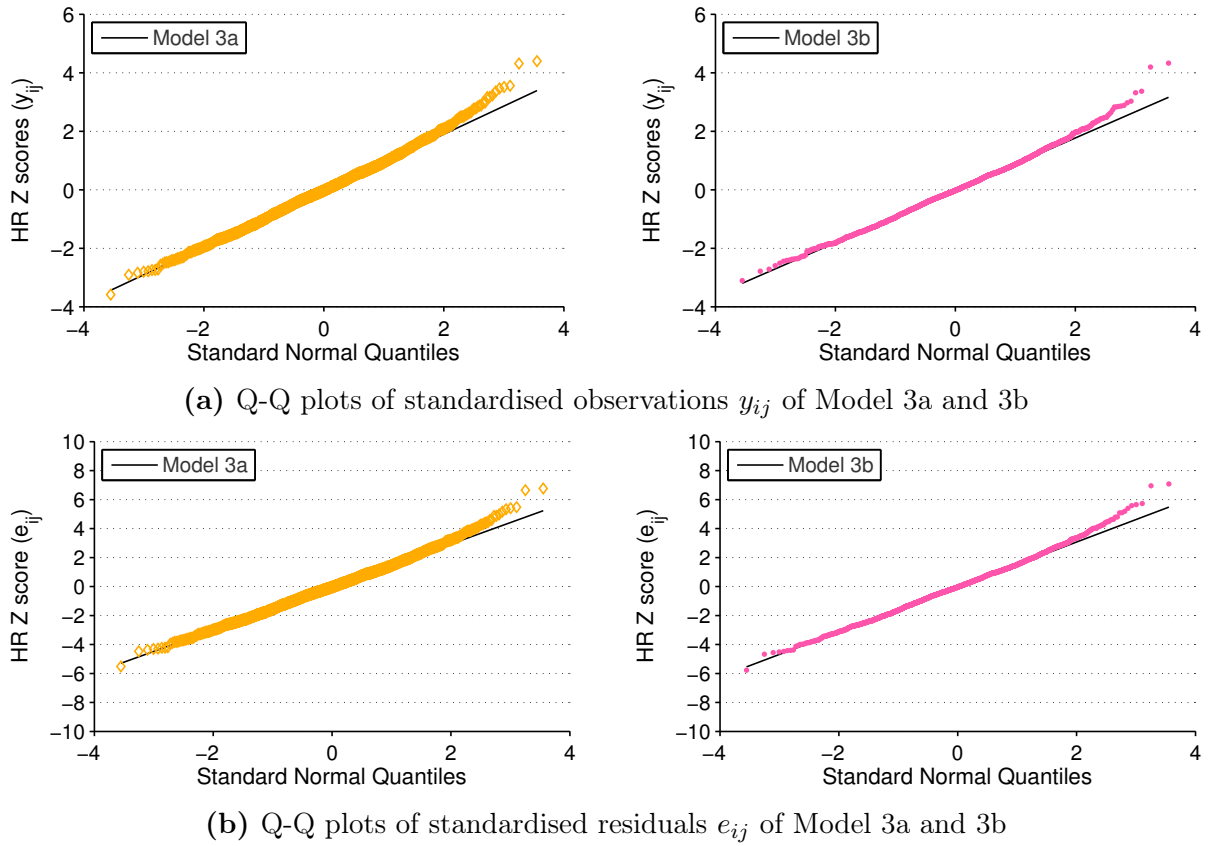


Figure 5.3: Q-Q plots of standardised observations (a) and residuals (b) of Models 3a and 3b

Although the three multilevel models were formulated differently, they compared reasonably well. Table 5.3 summarises the maximum absolute differences between the

⁵Plots of the standardised residuals against gestational age and tabulated values of the AIC, BIC and log likelihood ratio tests that support these results are included in Appendix D.1.

Table 5.3: Maximum absolute differences between centiles estimated from multilevel models

Comparison	Maximum absolute difference (bpm)		
	50 th centile	3 rd centile	97 th centile
{Model 3a, Model 3b}	0.03	4.35	4.35
{Model 3a, Model 3c}	0.26	0.0004	0.0004
{Model 3b, Model 3c}	0.24	4.34	4.34

mean and centiles of Models 3a-c. The differences between the means (50th centiles) of Models 3a-c, and between the 3rd and 97th centiles of Models 3a and 3c, were negligible. The difference between the models is only evident when examining the 3rd and 97th centiles in the last 3–4 weeks of pregnancy. During this period, the maximum difference between the centiles of Model 3a/c and 3b was 4.35 bpm. To understand why the separation between the 3rd and 97th centiles of Model 3b may increase with gestational age, we consider the random effects of the model which are used to model the centiles. For Models 3a and 3c, only a random intercept b_{0i} was included and the SD is given by Equation 5.13. For Model 3b, a random intercept b_{0i} and a random slope $b_{1i}t_{ij}$ were included and the corresponding SD is given by Equation 5.14. Here, D_{11} , D_{12} and D_{22} are elements of the matrix $\mathbf{D}$.

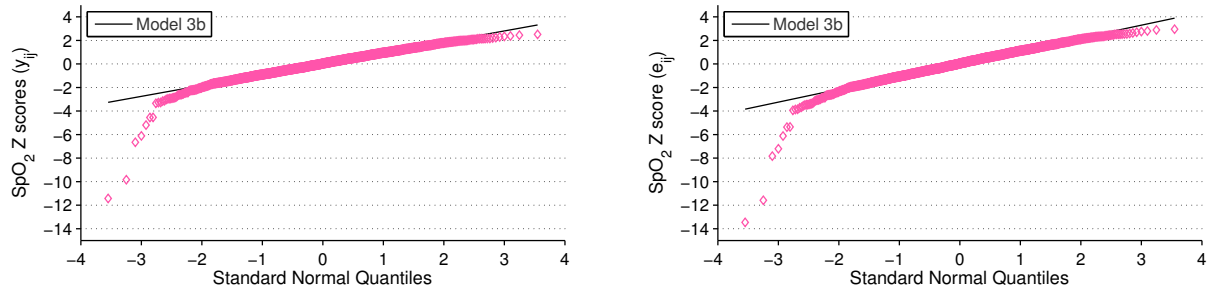
$$SD(y_{ij}) = \sqrt{\text{Var}(b_{0i} + e_{ij})} = \sqrt{D + \sigma^2} \quad (5.13)$$

$$SD(y_{ij}) = \sqrt{\text{Var}(b_{0i} + b_{1i}t_{ij} + e_{ij})} = \sqrt{D_{11} + 2D_{12}t_{ij} + D_{22}t_{ij}^2 + \sigma^2} \quad (5.14)$$

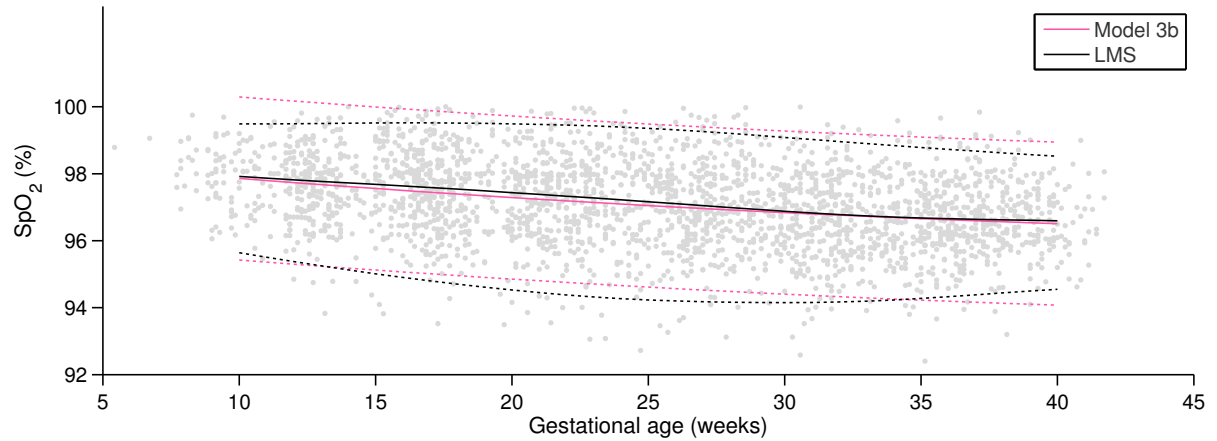
Equations 5.13 and 5.14 show that both a random intercept and slope are required to model heteroscedasticity; this explains why the separation between the centiles of Model 3b may increase with gestational age. Moreover, when $\mathbf{R} = \sigma^2\mathbf{I}$, correlations between individuals are only accounted for when $\mathbf{D}$ is a matrix with dimensions 2×2 or greater. If the within-subject correlation that arises from multiple measurements is ignored, centile precision may be overestimated, resulting in erroneously narrow centiles [172]. The observed difference of 4.35 bpm between the centiles of Model 3b and 3a/c could affect clinical decision-making for patients who are borderline abnormal.

5.3.4 Additional considerations for modelling SpO₂

Figure 5.4a shows the Q-Q plots, from application of Model 3b to the SpO₂ data, used to assess whether the mean measured values of SpO₂ in our sample could be assumed to follow a normal distribution conditional on gestational age. The points tend to a curve with a concave downward shape, indicating negative skewness in the data. We therefore computed an alternative estimate of centiles from the LMS method, which accounts for skewness, as shown in Figure 5.4b. The maximum absolute differences between the LMS-estimated centiles and the centiles computed from Model 3b are summarised in Table 5.4. As the differences between the 50th centiles and lower (3rd) centiles were small, we concluded that the measured mean values of SpO₂ were sufficiently normally distributed for the distributional assumption of the multilevel model to be valid.



(a) Q-Q plots of standardised observations and residuals of Model 3b



(b) Comparison of centiles for SpO₂ from the LMS model (black) and Model 3b (pink). A few outliers < 92% are not shown. While the 97th centile of Model 3b exceeds the limit of 100%, we chose to disregard this as upper centiles will not be used to score abnormal values of SpO₂.

Figure 5.4: Revised centiles for SpO₂

Table 5.4: Maximum differences between centiles of SpO₂ from the LMS and multilevel models

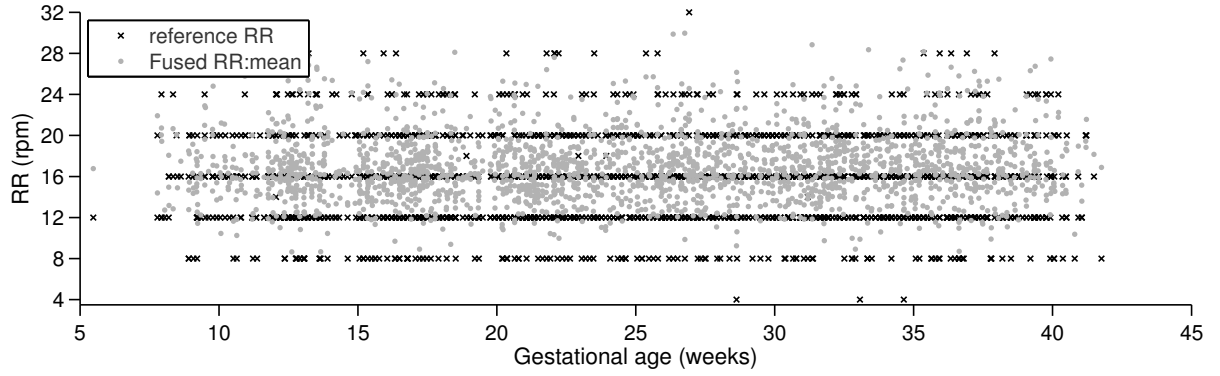
Comparison	Maximum difference (%)		
	50 th centile	3 rd centile	97 th centile
{LMS model, Model 3b}	0.14	0.47	0.80

A final point to make about SpO₂ is that although we have chosen to include estimates of both the upper and lower centiles here (to emphasise the distribution of the measured SpO₂ values), only a low value of SpO₂ (i.e. *below* the median) is a relevant indicator of physiological abnormality. This will be discussed in more detail in the next chapter.

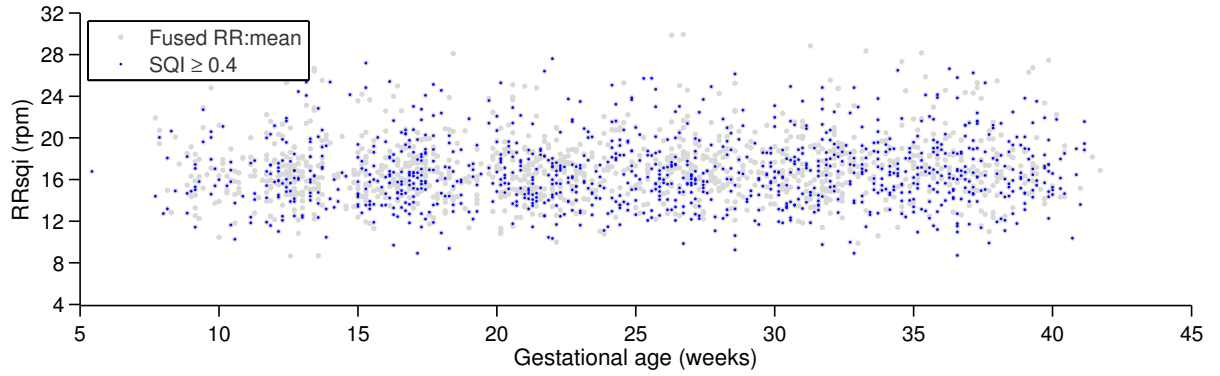
5.3.5 Additional considerations for modelling respiratory rate

The current reference standard for measurement of RR in clinical practice is a count of breaths for 15 seconds. However, as this measurement is quantised in steps of 4 rpm, we used the fused electronic RR estimate, described in Section 4.6, for the computation of centiles. Figure 5.5a shows the fused electronic measurement and the reference measurements plotted against gestational age. From the figure, the quantisation of the reference measurement, with respiratory rates that are multiples of four, is obvious. We further note that while there is a considerable number of reference measurements at 8 rpm, few of the electronically estimated RR are below 10 rpm. As discussed in Section 4.4.4, the relatively high number of low reference RR values is thought to be an artefact of the counting method rather than a representation of true low RR values.

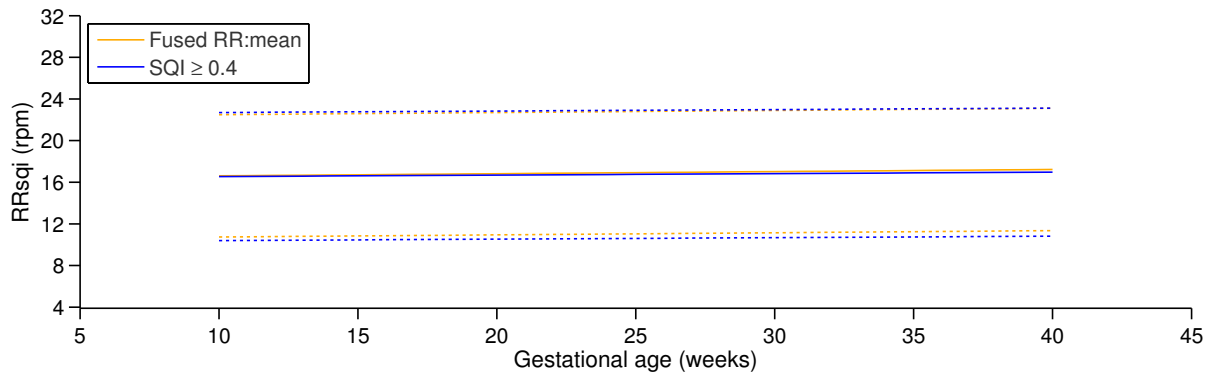
As discussed in Chapter 4, the RR:mean estimated will inevitably tend towards the mean value of RR across all of the recordings. This was seen in Section 4.4.4, where RR:mean values derived from the A_P and RIFV signals appeared accurate between 16 to 20 rpm, but tended to over- and under-estimate RR at low and high reference RR, respectively. However, more accurate estimates were obtained by only considering RR estimates from signals with a high $SQI \geq 0.4$. Based on this, we compared the centiles computed from all RR estimates ($n=2,065$) against the centiles computed from a subset of RR estimates derived from signals with $SQI \geq 0.4$ ($n=1,029$). The comparison of the



(a) Fused electronic RR:mean (grey) and reference (black) estimates of RR plotted against gestational age



(b) All fused electronic RR:mean estimates (grey) and estimates with $SQI \geq 0.4$ (blue) plotted against gestational age



(c) Centiles computed from all electronic RR estimates (orange) plotted together with centiles computed from estimates with $SQI \geq 0.4$ only (blue)

Figure 5.5: Data points and centiles of electronic RR estimates plotted against the reference RR and against a subset of good-quality RR, respectively.

Table 5.5: Maximum differences between centiles of RR when considering all recordings vs. only recordings with $SQI \geq 0.4$

Comparison	Maximum difference (rpm)		
	50 th centile	3 rd centile	97 th centile
{All recordings, $SQI \geq 0.4$ }	0.25	0.53	0.21

data and the centiles plotted against gestational age are shown in Figures 5.5b and 5.5c, respectively. The maximum absolute differences between the centiles are summarised in Table 5.5. The maximum difference between the extreme centiles is ~ 0.5 rpm. As this is less than a single breath we do not consider this difference clinically significant and we therefore assume that the centiles derived by including all estimates of RR, irrespective of quality, provide an accurate representation of the variation in RR with gestational age.

5.3.6 Final model selection

Based on the results and discussions in the previous section, we conclude that using 4P vital sign data, Model 3b was the most appropriate for the modelling of changes in vital signs conditional on gestational age. To summarise why this is the most appropriate model, we return our discussion to the three MGRS criteria introduced in Section 5.2.

The first point concerned the distributional assumption. The assumption of normality imposed by the Gaussian process (Model 2) and multilevel models (Models 3a-c) appears to be reasonable. Section 5.3.3 showed that this was the case for HR, but the normality assumption was also found to be valid for the other vital signs, including SpO_2 as shown in Section 5.3.4. The distributional assumption has two key advantages. Firstly, it enables a potentially more accurate prediction of extreme centiles, for which data may be limited (this was highlighted in the application of Model 1). Secondly, a distributional assumption is convenient because we can then compute the conditional likelihood of a new observation (x_*, y_*) conditional on gestational age x_* . For a normal distribution, the conditional likelihood $p(y_*|x_*, \hat{\mu}, \hat{\sigma}^2)$ is given by Equation 5.15 where $\hat{\mu}$ and $\hat{\sigma}$ are the estimated mean and SD conditional on gestational age x_* [159].

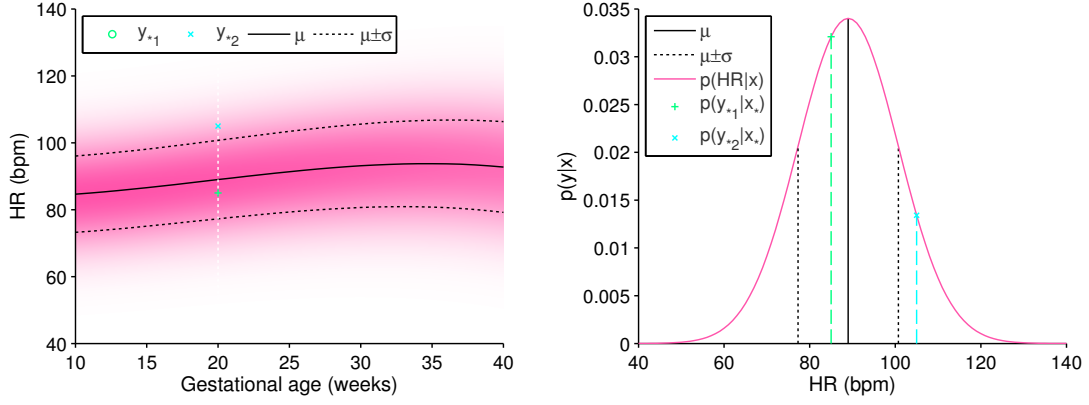


Figure 5.6: Two observations, y_{*1} and y_{*2} , of HR are collected at week 20 of pregnancy [left]. The likelihoods of the two points, $p(y_{*1}|x_*)$ and $p(y_{*2}|x_*)$, are then computed and compared [right].

$$p(y_*|x_*, \hat{\mu}, \hat{\sigma}^2) = (2\pi\hat{\sigma}^2)^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \frac{(y_* - \hat{\mu})^2}{\hat{\sigma}^2} \right\} \quad (5.15)$$

This applies to any model from which we can obtain an estimate of the mean and variance conditional on gestational age. Figure 5.6 shows an example of this, where the conditional likelihoods of two measurements of HR at gestational age $x_* = 20$ weeks are compared.

The second point concerns the choice of smoothing method. A conventional polynomial adequately describes the changes in vital signs with gestational age. The main concern of parametric smoothing is that it may not provide sufficient flexibility to model patterns of change over time that are not well represented by polynomial functions. However, the negligible difference between the trends predicted by the non-parametric models (Models 1 and 2) and the parametric models (Models 3a-c) shows that a polynomial formulation of time was adequate for our purpose. There is also little difference between the trends resulting from a conventional polynomial and the more flexible fractional polynomial formulation. This finding applies to the models for any vital sign. A main advantage of the polynomial formulation is that it yields equations for the estimated mean and SD at any value of gestational age. This is in contrast to non-parametric models, for which the likelihood at any gestational age must be computed and subsequently accessed using a look-up table or similar. For Model 3b, the expected SD for an observation at time x_*

would be found from Equation 5.14 with $t_* = x_* - 10$. The associated expected mean is given by Equation 5.16.

$$\hat{\mu} = E(y_*) = \beta_0 + \beta_1 t_* + \beta_2 t_*^2 + \beta_3 t_*^3 \quad (5.16)$$

The third and final point concerns the ability of the statistical models to account for correlations between repeated measurements from the same individual. In Model 3b, we accounted for within-subject correlations by allowing for random variation in the intercept and slope for each subject. For HR, the main effects of accounting for correlations was twofold; it resulted in slightly greater separation between the upper and lower centiles and enabled modelling of heteroscedasticity (i.e. an increase in the separation between centiles with gestational age). In Section 5.5, we will show how the random effects can be used to investigate individual rates of change in vital signs.

Overall, the differences between the centiles estimated by the different models were not large. The minimal differences may to some extent be explained by the measurements being made at pre-specified gestational age ranges following a strict protocol. Approximately 90% of the subjects were measured 4 to 7 times, and the frequency of the measurements was independent of previous measurements. A greater difference between the models that accounted for repeated measurements and the models that did not would perhaps have been found had there been greater variation in the number of measurements per individual.

5.4 Final centiles for all vital signs

Using Model 3b, applying the same principles for model fitting and assessment of fit as described in Section 5.3, we constructed centiles for all six vital signs. For Temp, we used the measurements from the Fora IR20b thermometer (Temp(BT)) for reasons explained in Section 3.4.3. For respiratory rate, we used the fused electronic measurements obtained from the accelerometer and PPG, as explained in Sections 4.6 and 5.3.5.

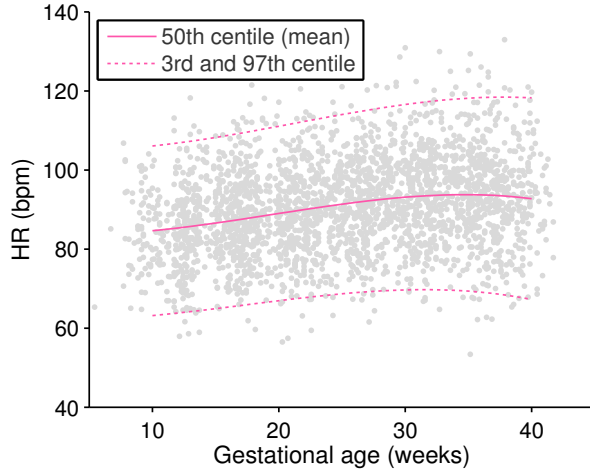
When fitting the models for SpO₂, Temp and RR, the model did not converge when a random slope term $b_{i1}x$ was included. This is likely to be because the ranges of values for SpO₂, Temp and RR were considerably smaller than the ranges of HR, SBP and DBP values. To help the model converge, Singer and Willett [167] recommended modelling the centiles on a log-scale. However, the effect of this was unlikely to be clinically significant and so was not attempted. Consequently, these models only include a random intercept term b_{i0} (this formulation is equivalent to that of Model 3a). The final centiles are shown in Figure 5.7 and the associated equations for the estimated mean and *SD* are provided in Appendix D.2. The changes in each vital sign are discussed below.

Mean HR was found to increase throughout pregnancy, from about 85 bpm (week 10) to 94 bpm (week 35). This is comparable to the increase reported elsewhere in the literature [40], although evidence was previously limited (as shown in Chapter 2). At around 35 to 40 weeks, the increase in HR was found to level out and the value of mean HR remained constant until term. The variance of HR was found to be heteroskedastic. Early in pregnancy at 10 weeks of gestation, the separation between the 3rd and 97th centiles was 43 bpm. By 40 weeks of gestation, the separation between these centiles increased to 51 bpm.

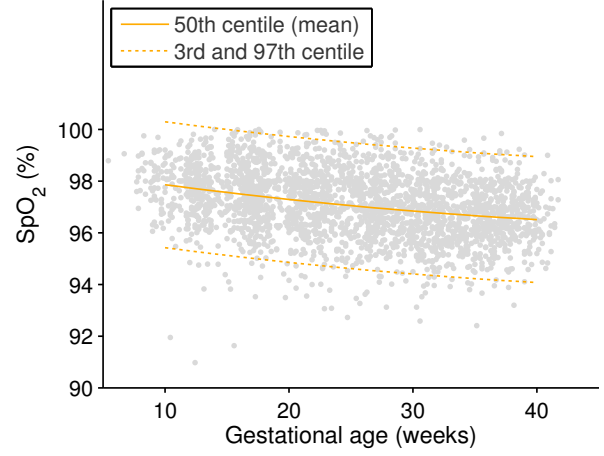
With a mean of 98% (week 10), measured mean SpO₂ was initially high in early pregnancy, as compared to measured values of SpO₂ in the non-pregnant state⁶. It subsequently decreased by approximately 1.5% to reach the non-pregnant level by week 40. The distribution of measured mean SpO₂ values was found to be close to that of a normal distribution. This is different to the distribution of SpO₂ across a population, which tends to form a one-sided distribution [158].

SBP and DBP initially decreased to minimum mean values of 114 and 70 mmHg (week 20), respectively, before increasing to maximum values of 121 and 80 mmHg (week 40). The *SD*, and hence the separation between the 3rd and 97th centiles, of both SBP and DBP increased with gestational age. This effect was greatest for SBP, for which the

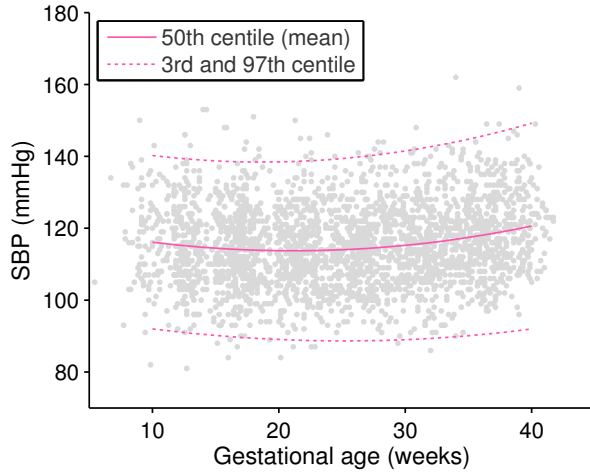
⁶As deduced from the postpartum values of SpO₂, using data from the 4P study data not included in this thesis



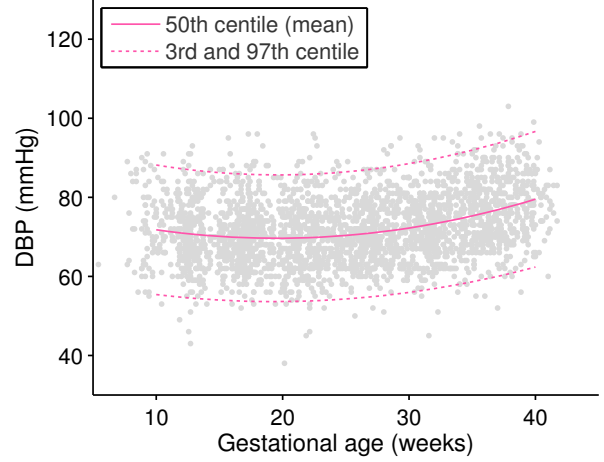
(a) Centiles of HR computed from Model 3b



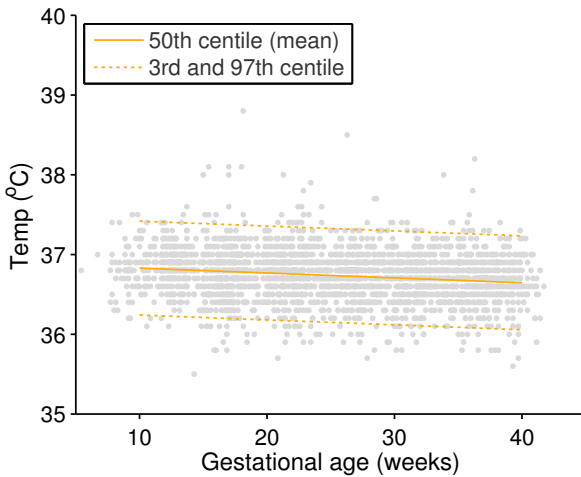
(b) Centiles of SpO₂ computed from Model 3a



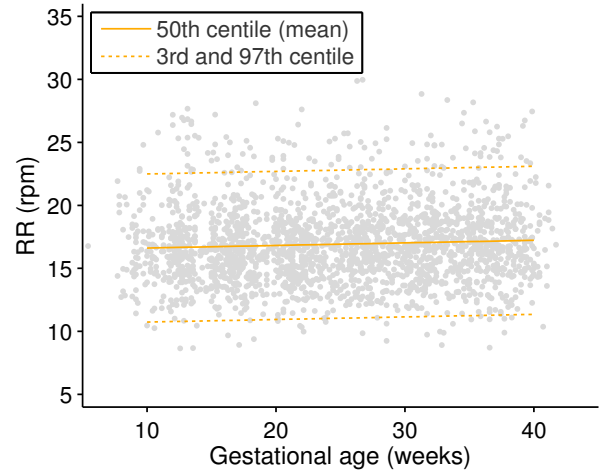
(c) Centiles of SBP computed from Model 3b



(d) Centiles of DBP computed from Model 3b



(e) Centiles of Temp computed from Model 3a



(f) Centiles of RR computed from Model 3a

Figure 5.7: Final centiles for each of the six vital signs superimposed on the raw data (grey dots). The centiles shown are the 50th centile (solid line) and the 3rd and 97th centiles (dotted lines). The colours indicate a random-intercept model (Model 3a, orange) and a random intercept and slope model (Model 3b, pink) – see Section 5.3.6 for justification of model choices.

separation between the upper and lower centiles increased from 49 mmHg (week 10) to 57 mmHg (week 40). The separation between centiles of DBP were 33 mmHg (week 10) and 35 mmHg (week 40). The finding of an initial decrease and subsequent increase in BP agreed well with findings from literature described in Chapter 2, Section 2.3.1.

Figure 5.8 shows a comparison of the mean BP trajectories from 4P and from the AVON study, as reported by Macdonald-Wallis *et al.* [79]. Data for the AVON study were collected during the period from 1991 to 1992. With data from $\sim 10,000$ women it is the biggest published study of maternal BP to date in the UK. While the overall trend is similar, the data from the 4P study show noticeably higher expected means of SBP and DBP. The difference in means could be an effect of different maternity populations and antenatal care schemes. However, as shown in Section 2.3.1, the increase could also reflect a population-level increase in maternal BP over the 20-year period separating the two studies. The difference is likely to be reduced by a difference in method of measurement⁷, hence it is not as big as the increase in SBP and DBP predicted by the systematic review of ~ 0.5 mmHg/year. We also note that the difference in DBP between the two datasets appear to be bigger than that of SBP. As maternal characteristics were poorly reported by Macdonald-Wallis *et al.* we are not able to attribute this difference to a difference in characteristics such as age, as discussed in Section 2.4. This difference is potentially also linked to the type and calibration of the measurement devices used in the studies.

The smallest changes were found for Temp and RR. Temp showed a slight linear decrease by approximately 0.2°C from 36.8°C (week 10) to 36.6°C (week 40). RR was found to increase slightly from 16.6 rpm (week 10) to 17.2 rpm (week 40), but this change is not clinically significant. As shown in Section 5.3.5, the same change was found independent of whether we considered all estimates of RR or only high-quality estimates. The finding of a nearly constant RR is unexpected for two reasons. Firstly, because the changes in SpO_2 in early pregnancy do not appear to be linked to changes in RR. Secondly, because pregnant women have frequently reported a feeling of breathlessness

⁷The AVON study used a manual measurement of BP, which based on the systematic review findings may have resulted in higher BP readings

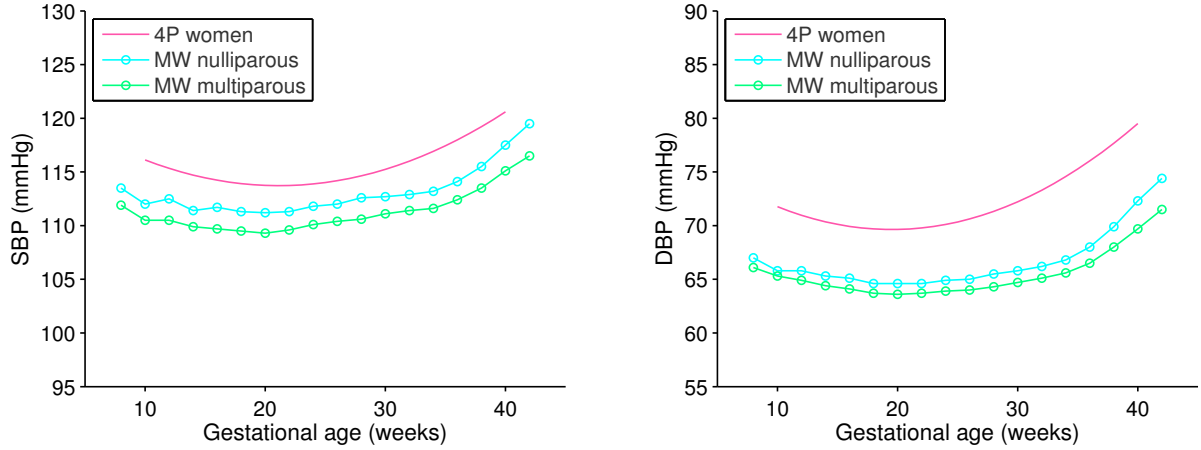


Figure 5.8: Computed mean for SBP [left] and DBP [right] from the 4P study (pink) compared to the mean BP for nulliparous (blue) and multiparous (green) women reported by Macdonald-Wallis *et al.* [79].

towards the end of pregnancy [173]. An alternative explanation could perhaps be that because the physiological adaptations to pregnancy occur over an extended period of time (rather than instantaneously), the body has time to adapt to the change so that a change in RR does not occur.

5.5 Modelling individual patterns of change

In routine antenatal care, individuals are often assessed at a single time point. However, using the multilevel modelling approach described here, it is also possible to assess an individual's trajectory or pattern of change in vital signs over time. In addition to the estimates of the population mean and variance, the multilevel model returns the random effect coefficient values $\mathbf{b}_i$ for each subject i . The vector $\mathbf{b}_i$ contains the random intercept b_{i0} and slope b_{i1} . Because we used a transformation of time $t_{ij} = x_{ij} - 10$ to fit the multilevel models (see Section 5.3.1), the intercept can be interpreted as the difference between each woman's vital sign value and the population mean at a gestational age $x_{ij} = 10$ weeks. The slope reflects the velocity of change. The random effects are referred to as *Empirical Bayes* (EB) predictions, because they are predicted using the prior distribution of random effects $\mathbf{b}_i \sim \mathcal{N}(0, \mathbf{D})$, our estimates of the model parameters

$\hat{\beta}$, $\hat{\mathbf{D}}$ and $\hat{\sigma}^2$ and the subject's data points $\mathbf{y}_i = y_{i1}, \dots, y_{in_i}$ (see Appendix B.3.5).

In this section, we will first provide the reader with a basic intuition for how the values of the estimated parameters affect the EB predictions. We will then use these estimates to assess how BP and HR differ between groups of women with different characteristics. In particular, we will investigate the patterns in BP and HR of the *very healthy* women (referred to as 'IG women' in Section 3.3) and women who developed hypertensive disorders of pregnancy (HDP).

5.5.1 Interpretation of the predicted random effects

For a basic random intercept model given by $y_{ij} = \beta + b_i + e_{ij}$, the EB prediction of a subject's mean $\tilde{\alpha}_i = \beta + b_{i0}$ represents a weighted average of the maximum likelihood (ML) estimate of an individual subject's observations and the population mean. By considering that $y_{ij} \sim \mathcal{N}(\alpha_i, \sigma^2)$ and $\alpha_i \sim \mathcal{N}(\beta, D)$, we can obtain Equation B.34, where $\bar{y}_i = n_i^{-1} \sum_{j=1}^{n_i} (y_{ij})$ is the ML estimate of the subject-specific mean^{8,9}.

$$\tilde{\alpha}_i = (1 - \omega_i)\hat{\beta} + \omega_i\bar{y}_i \quad (5.17)$$

where $\omega_i = \frac{D}{D + \frac{\sigma^2}{n_i}}$

The predicted subject mean $\tilde{\alpha}_i$ is thus a weighted sum of the estimated population mean $\hat{\beta}$ (if $\omega_i \rightarrow 0$, $\tilde{\alpha}_i \rightarrow \hat{\beta}$) and the ML estimate of the subject mean $\bar{y}_i$ (if $\omega_i \rightarrow 1$, $\tilde{\alpha}_i \rightarrow \bar{y}_i$).

The value ω_i is the *shrinkage factor*, which determines the amount of shrinkage towards the population mean [172]. In general, there will be more shrinkage ($\tilde{\alpha}_i \rightarrow \hat{\beta}$) if:

- the between-subject variance $\hat{D}$ is small (an informative prior)
- the within-subject variance $\hat{\sigma}^2$ is large (uninformative data)
- the cluster size n_i is small (an uninformative cluster)

This is illustrated in an example in Appendix B.3.5.

⁸The derivation of the expression for a general multilevel model is included in Appendix B.3.5

⁹The value n_i was previously defined as the number of measurements for subject i

5.5.2 Subject-specific trends of heart rate and blood pressure

This investigation aimed to qualitatively compare trajectories of SBP, DBP and HR between groups of women with different clinical outcomes. It was conducted in two parts. In the first part, we compared the distributions of random effects (intercepts b_{i0} and slopes b_{i1}) between *very healthy* women ('IG women'), to those of women who did not meet these criteria ('non-IG women'). In the second part, we compared the distributions of b_{i0} and b_{i1} between women with hypertensive disorders of pregnancy ('HDP'), including gestational hypertension, pre-eclampsia and severe pre-eclampsia, and women who were not diagnosed with any hypertensive disorder ('no HDP').

For each part, the joint and marginal posterior distributions of b_{i0} and b_{i1} were plotted and visually compared. The marginal distributions $p(b_{i0})$ and $p(b_{i1})$ for each group were obtained using the Gaussian kernel density estimator given by Equation 5.18 where h is a smoothing parameter [174]. The values of h were set empirically by visual comparison to histograms to ensure smooth distributions. A non-parametric form of density estimation was used because of the small number of subjects in each group, and because the distributions of the separate groups could not necessarily be assumed to be normal.

$$p(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^N \frac{1}{\sqrt{2\pi}h^2} \exp \left\{ -\frac{1}{2} \frac{\|\mathbf{x} - \mathbf{x}_n\|^2}{h^2} \right\} \quad (5.18)$$

In all cases, the prior joint distribution $p(\mathbf{b}_i)$ is given by the between-subject variance matrix $\mathbf{D}$. The correlation between b_{i0} and b_{i1} is given by Pearson's correlation coefficient ρ ¹⁰. As the amount of shrinkage for each subject's EB predictions is known to be affected by the number of measurements, subjects with a number of measurements $n_i < 5$ were excluded. Including these women would lead to a bias of the estimated densities towards the population mean (given by $b_{i0} = b_{i1} = 0$). A total of 408 and 409 women had five or more measurements of BP and HR, respectively.

¹⁰Pearson's correlation coefficient is defined in Appendix B.3.2

Very healthy women

The main differences in population characteristics between the mutually exclusive groups 'IG women' and 'non-IG women' were smoking status (all 'IG women' were non-smokers), BMI (all 'IG women' had a normal BMI between 20 and 25 kg/m²) and severity of pregnancy complications (no 'IG women' were hospitalised as a result of pregnancy complications) [95]. Figure 5.9 shows for SBP, DBP and HR the prior joint distributions and EB predictions of the random intercepts b_{i0} and slopes b_{i1} for each group and the marginal distributions $p(b_{i0})$ and $p(b_{i1})$ obtained using kernel density estimation. The prior joint distributions $p(\mathbf{b}_i)$ of HR, SBP and DBP showed negative correlation between b_{i0} and b_{i1} , with the random-effects of HR being the least negatively correlated ($\rho = -0.10$) and those of DBP being the most negatively correlated ($\rho = -0.36$).

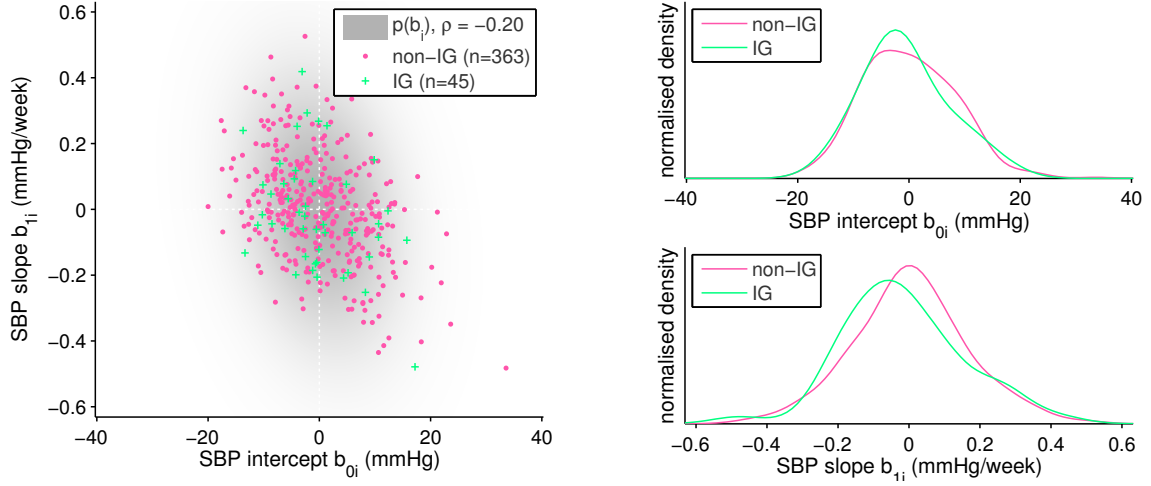
Overall, the differences between the two groups 'IG women' and 'non-IG women' were minimal. For SBP, the marginal distributions $p(b_{i0})$ of the two groups were similar. The marginal distribution $p(b_{i1})$ of 'IG women' was slightly left-shifted compared to that of 'non-IG women', indicating that the SBP trajectories of some 'IG women' were associated with a less steep slope. For DBP, the distribution $p(b_{i0})$ of 'IG women' was slightly left-shifted compared to that of 'non-IG women', indicating that DBP trajectories of women in this group tended towards a lower initial value of DBP. For HR, the marginal distributions $p(b_{i0})$ and $p(b_{i1})$ of 'IG women' were only slightly left-shifted, indicating HR trajectories with a slightly lower initial value and a less steep slope within this group.

Women with hypertensive disorders of pregnancy

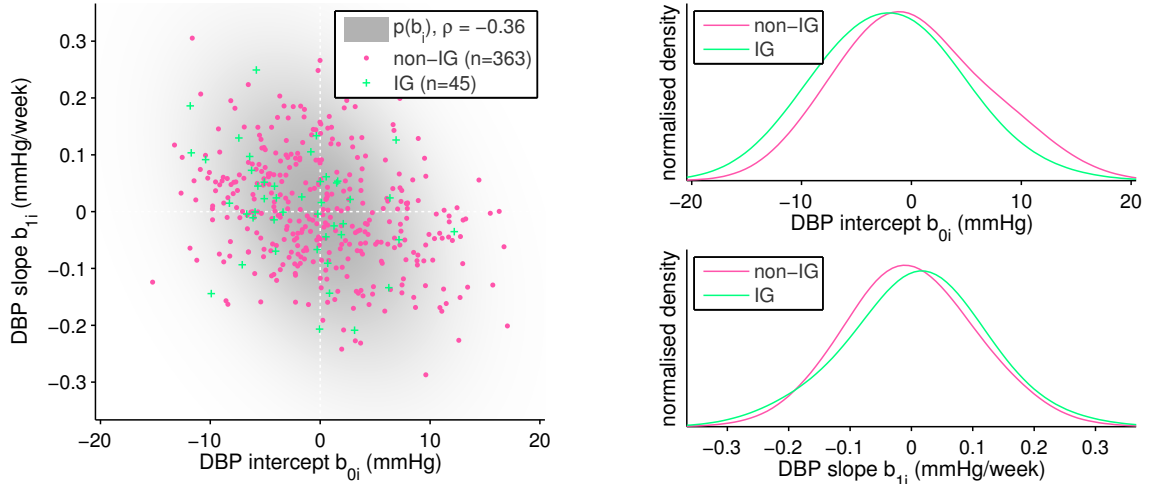
Of the 28 women in the 'HDP' group, 25 women were diagnosed with gestational hypertension, 5 women were diagnosed with pre-eclampsia¹¹ and 1 woman was diagnosed with severe pre-eclampsia (see Table 3.4 and note that these diagnoses were not mutually exclusive).

The results from the comparison of 'HDP' and 'no HDP' women are summarised in Figure 5.10. The prior distributions $p(\mathbf{b}_i)$ of SBP, DBP and HR are the same as before.

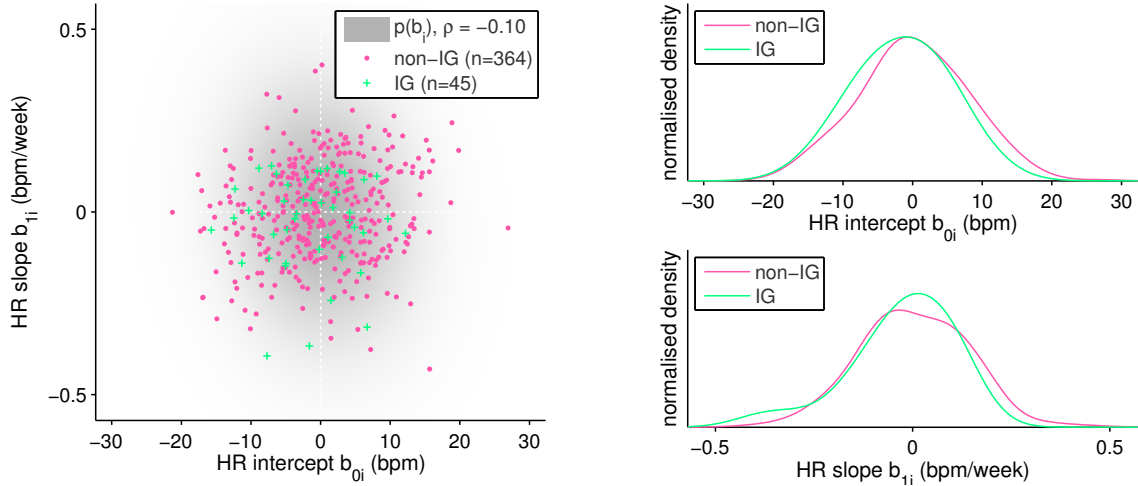
¹¹According to the ISSHP criteria for diagnosis of pre-eclampsia presented in Section 1.1



(a) Systolic blood pressure

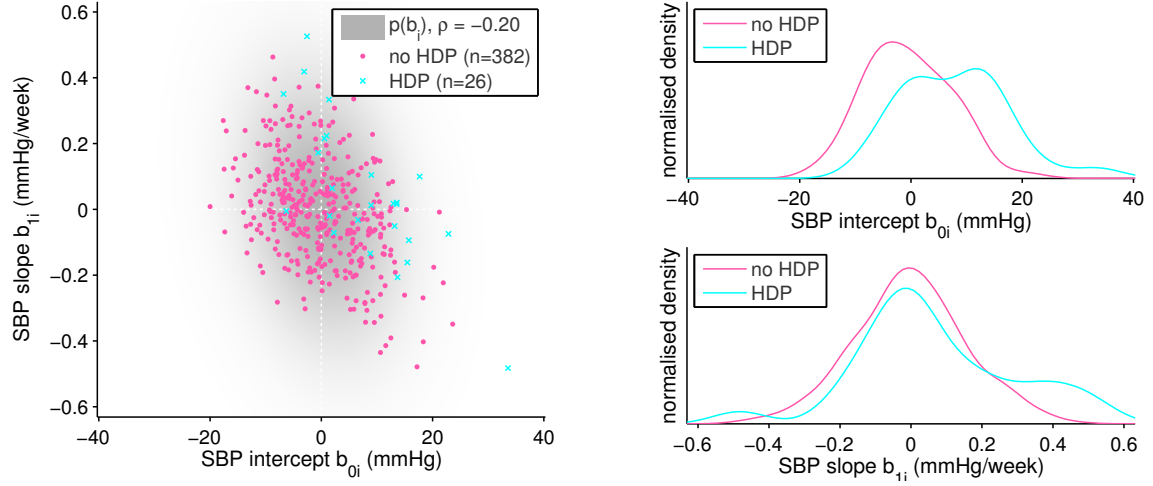


(b) Diastolic blood pressure

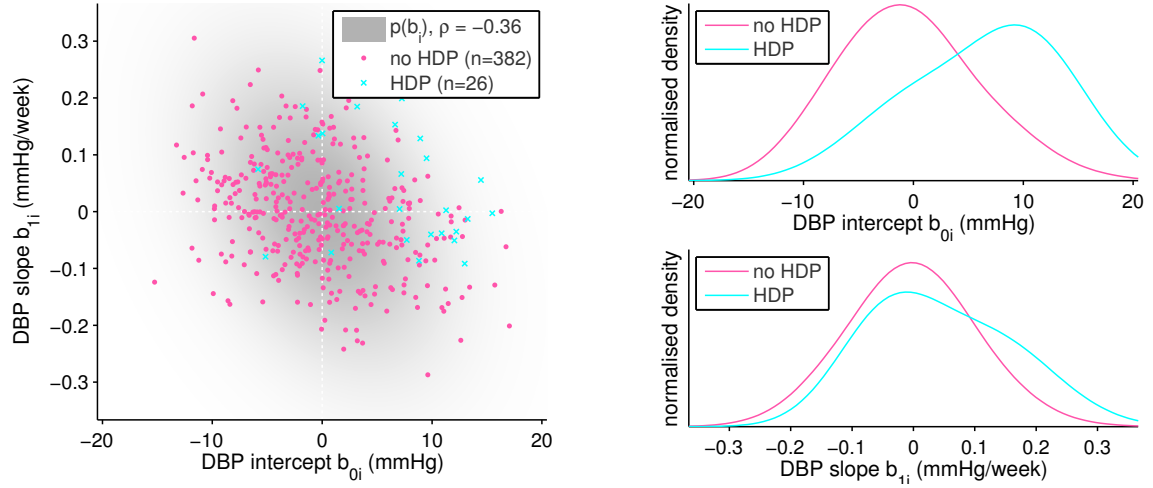


(c) Heart rate

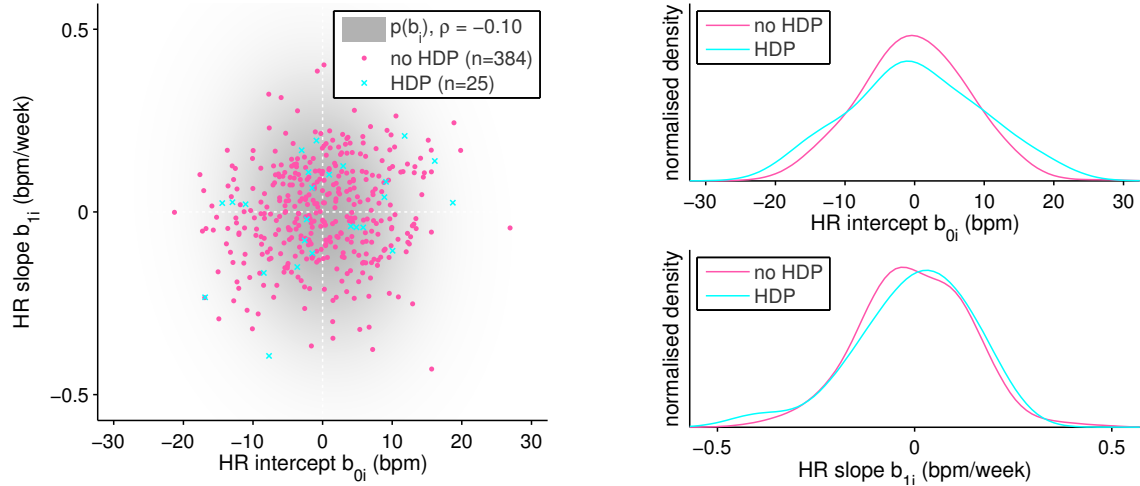
Figure 5.9: Patterns in random intercepts b_{i0} and slopes b_{i1} for 'IG women' (green circles/lines) and 'non-IG women' (grey dots/lines). The plots show the prior distribution $p(b_i)$ [left] and marginal distributions $p(b_{i0})$ and $p(b_{i1})$ [right]



(a) Systolic blood pressure



(b) Diastolic blood pressure



(c) Heart rate

Figure 5.10: Patterns in random intercepts b_{i0} and slopes b_{i1} for 'HDP' (blue crosses/lines) and 'no HDP' (grey dots/lines). The plots show the prior distribution $p(b_i)$ [left] and marginal distributions $p(b_{i0})$ and $p(b_{i1})$ [right]

For both SBP and DBP, the majority of 'HDP' women had values of $b_{i0} > 0$ and $b_{i1} > 0$. This can be seen from the joint distributions $p(b_{i0}, b_{i1})$ where most of the 'HDP' points lie in the upper-right quadrant. This implies that 'HDP' women tended to have trajectories of SBP and DBP with higher initial values and more steep slopes compared to the overall population mean. For the marginal distributions $p(b_{i0})$ and $p(b_{i1})$, this translates to a considerable right-shift in the distributions of 'HDP' women compared to 'non-HDP' women. However, a few 'HDP women' exhibited alternative patterns including SBP and DBP trajectories with either a lower initial value and steeper slope or a higher initial value and less steep slope, as compared to the overall population mean. For HR, no major differences were observed between 'HDP' and 'non-HDP' women.

Concluding remarks

While the SBP, DBP and HR trajectories of 'IG women' and 'non-IG women' did not differ much, a greater difference was found when comparing BP trajectories of 'HDP' and 'non-HDP' women – this would be expected given that HDP are diagnosed based on high values of BP. Figure 5.11 shows typical examples of vital sign trajectories from one woman in each of the 'IG women' and HDP' groups relative to the mean and SD of all individual trajectories within the population.

The findings we have described here agree well with previous studies, which have shown that women who go on to experience gestational hypertension and/or pre-eclampsia have both a higher initial BP and a steeper increase in BP, as compared to women who remain normotensive for the duration of pregnancy [84,175]. However, while these studies also showed that the extent of these patterns depends on factors such as disease severity, our investigation was limited by the number of women in each group, and by the fact that women with different diagnoses were grouped together. Some women in the 'HDP' group were diagnosed with additional medical conditions such as gestational diabetes, diabetes and cardiac disease which can also be expected to affect vital sign patterns [176]. As only a small proportion of women in our cohort developed such complications, we did not

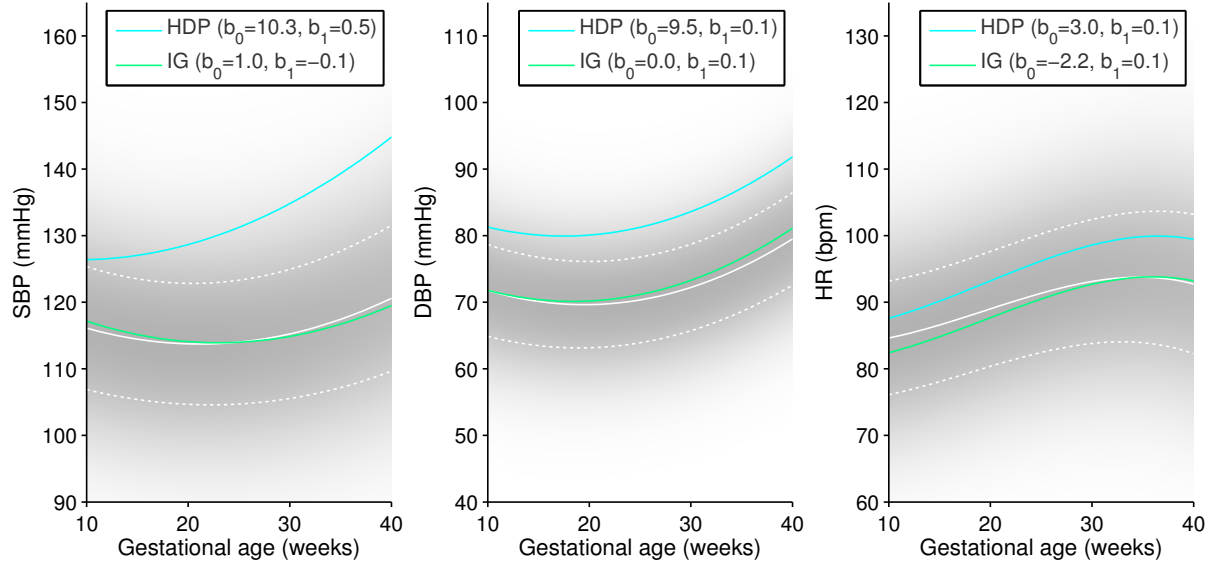


Figure 5.11: Examples of typical trajectories of [left] SBP, [middle] DBP and [right] HR of 'IG' and 'HDP' women. The distribution of trajectories is shown in grey, with the population mean ($b_{i0} = 0, b_{i1} = 0$) and SD superimposed in white. Typical 'IG' and 'HDP' trajectories are shown in green and blue, respectively. The associated values of the random intercepts b_{i0} and slopes b_{i1} are shown in the legends.

have enough data to stratify women based on different types of HDP and other medical conditions. For the study of individual patterns of disease, the PROGRESS (PROGnosis RESearch Strategy) group recommends the use of a dataset where at least 100 subjects are diagnosed with the disease of interest [177].

Additionally, we did not have information about the timing of diagnoses and treatment interventions (such as administration of antihypertensive medication) in any of the groups. Any treatment would inevitably affect the observed vital sign values. Nevertheless, the findings serve to illustrate some of the additional information that one can extract from the multilevel model. Because a minimum number of data points for each woman was required for this analysis, the method described here is mainly suited for retrospective analysis. In Chapter 7 we will also discuss alternative models to perform a prospective analysis for predicting abnormality based on individual vital sign trajectories. As the recruitment to the 4P study is ongoing, we will in the future have more data available to investigate subject-specific trajectories of vital signs in greater detail.

5.6 Discussion

The aim of this section was to assess different methods for design of gestation-specific reference ranges for vital sign data and to compute such reference ranges for all six vital signs for the antenatal period. The reference ranges presented serve to characterise the changes with gestational age in the trend and distribution of the vital sign measurements, and can be used as screening tools for potential problems in pregnancy.

We have shown that a change with gestational age is expected for all vital signs. The least change was observed for RR, which remained close to unchanged across pregnancy. Additionally, for HR, SBP and DBP we were able to show that the separation between the upper and lower centiles increases with gestational age. Part of this increase could potentially be attributed to the fact that some of the women in our analysis developed pregnancy complications. The effect of several common pregnancy complications that may affect vital signs, such as gestational hypertension and pre-eclampsia, tend to become more severe with increasing gestational age [11, 13].

Figure 5.12 shows the estimated reference ranges of SBP, DBP and HR compared to the thresholds currently used in antenatal care to diagnose hypertension and tachycardia (abnormally high HR) in pregnancy, as recommended by the NICE [10] and MOET [17] guidelines, respectively. The need to adjust the thresholds with gestational age in order to accurately assess when a patient's physiology is becoming abnormal is obvious. The use of reference ranges to characterise abnormal values of vital signs will be further explored in the next chapter.

The reference ranges presented provide an accurate picture of the pregnant population from the Oxford area, who attended antenatal clinics during the period from 2012 to 2015. When comparing the vital sign patterns of women who were defined as *very healthy* to other women in our population, we did not find significant differences. However, as expected, we did observe a difference between the BP trajectories of women who developed HDP and women who did not. To properly assess the extent to which the

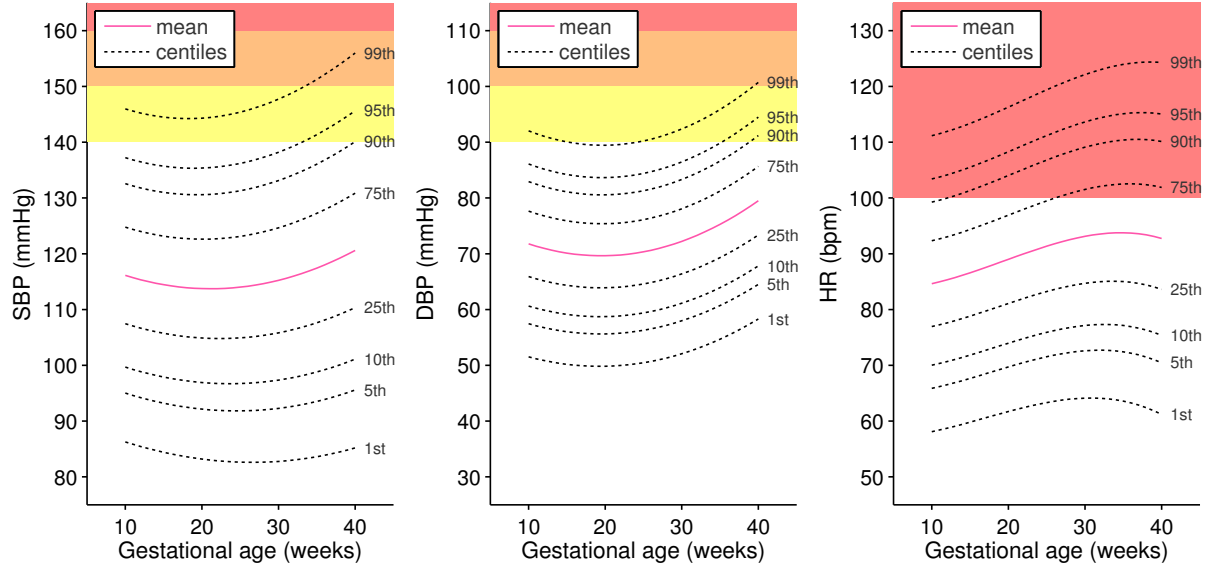


Figure 5.12: Computed centiles for SBP [left], DBP [middle] and HR [right] superimposed on current thresholds of normality. BP thresholds for diagnoses for mild hypertension, moderate hypertension requiring treatment and severe hypertension are shown in yellow, orange and red, respectively [10]. HR threshold for tachycardia is shown in red [17].

reference ranges are applicable to populations with different maternal risk profiles and antenatal care procedures, further research is needed. An extension of the 4P study to Newcastle and London is currently underway. Other populations from other geographic sites, both within the UK and internationally, should also be explored and the data acquired at various sites compared.

The reference ranges presented here did not account for covariates other than gestational age. Previous studies of pregnancy-induced changes of BP have shown that factors such as parity [79–81, 178], smoking status [155, 179], BMI/weight [80, 82] and ethnicity [78, 180] could affect the observed trends. Future work should investigate the effect (if any) of these covariates across our population for all six vital signs.

Finally, in Section 5.5 we began a discussion about subject-specific patterns in vital signs. The study of individual vital sign trajectories associated with medical diagnoses could be an exciting direction for future research, particularly in the context of prospectively predicting pregnancy complications linked to physiological deterioration. This topic will be discussed in more detail in Chapter 7.

Chapter 6

Design of a Modified Obstetric Early Warning Score

The need for a Modified Obstetric Early Warning Score (MEOWS) was presented in Chapter 1. Following the recommendation issued in the 2007 CEMD report regarding the introduction of a national MEOWS [1], several scores have been proposed. In this chapter, we will explain the fundamental design of a general early warning score. We will then examine a selection of MEOWS proposed in the literature in more detail. In particular, we will discuss the design and validation of these charts. Based on insights from this discussion, and the knowledge of normal ranges of vital signs during pregnancy established in the previous chapters, we will present the first version of a gestation-specific MEOWS score. The application of the score will be illustrated by means of patient data examples and future requirements for validation of the score will be discussed.

6.1 Design and evaluation of early warning systems

In order to assess the functionality of an Early Warning Score (EWS), some background knowledge on the purpose, design and evaluation of such systems is required. The principles described here are relevant for the design and performance evaluation of EWS in general, as well as MEOWS.

6.1.1 Purpose of early warning scores

EWS are often referred to as 'track-and-trigger' systems because they use physiological variables to 'track' an individual's condition, detect early deterioration and 'trigger' timely clinical intervention [34]. The goal of an EWS system is twofold: *(i)* it should correctly identify all events when patients are at risk of deterioration (defined as the positive case); and *(ii)* it should correctly recognise when a patient's condition does not give rise to concern (defined to be the negative case). A correct identification of the positive case is essential to enable timely referral and treatment of patients at risk of deterioration. A correct classification of the negative case is important to minimise the 'false alarm' rate and avoid any unnecessary workload for healthcare professionals [103,181].

6.1.2 Types of scoring algorithms used to trigger a response

Different EWS serve a common purpose as a tool for identification of patient deterioration. However, the designs of the charts may vary, both with respect to the variables included and the algorithms used to trigger a response. According to the definitions presented by Smith *et al.* [103,181], we separate between colour-coded trigger bands (single/multiple indicator trigger systems) and numerical scoring triggers (aggregate-weighted scoring systems). The fundamental functionality of both algorithms is however the same. As a first step, each physiological variable is measured and in turn assessed for normality by comparison to a set of pre-defined cut-off values. Secondly, each variable is assigned a score based on the deviation from normality. Finally, the scores for each variable are summed and the overall score is compared to an overall threshold to determine whether a response should be triggered.

For the colour-coded charts, the scoring bands which define the level of abnormality are typically indicated by the colours white (normal), yellow (borderline abnormal) and red (very abnormal). A response is triggered when a pre-defined combination of colours

is achieved, for example by two yellow values or a single red value. For the aggregate scoring systems, several scoring bands of different weightings are defined and each variable is assigned a weight depending on the degree of abnormality. Typically, seven scoring bands of weightings $T = \{3, 2, 1, 0, 1, 2, 3\}$, are used from abnormally low values (3) to normal values (0), to abnormally high values (3). The sum of scores for all variables is the aggregate (EWS) score. The final value is compared to a pre-determined trigger threshold (which varies for different systems) to determine whether a response should be triggered. Note that we use the word *cut-off* to describe the values of each vital sign which separate the different scoring bands and *threshold* (or *trigger threshold*) to describe the aggregate scoring criterion which determines whether a response should be triggered.

The 2007 CEMD report [1] advocated the use of a colour-coded chart because of its simplicity of use and interpretation. On the other hand, the National Early Warning System (NEWS), which is the UK standard EWS for use in the non-pregnant adult population, uses an aggregate scoring system, which enables the definition of multi-level response strategies for the escalation of care [27]. For the acutely-ill patient at risk of deterioration, NICE [24] has recommended that the EWS response strategy should be graded and consist of three levels for escalation of care: a low-score group that results in increased frequency of observations and alerting of the nurse in charge; a medium-score group, resulting in simultaneous urgent calls to the medical team caring for the patient and to personnel with competencies for acute illness; and a high-score group resulting in an emergency call to a team with critical care competencies and diagnostic skills. This response strategy can be translated to the levels of response described in Section 1.2.

6.1.3 Metrics for assessing the performance of early warning scores

The efficiency of EWS is commonly assessed using the metrics summarised by the confusion matrix in Table 6.1. In the case where a set of observations is assigned a score that triggers a response, it is identified as a positive case. Conversely, if the score does not trigger a response it is identified as a negative case. The model classifications are compared to

the true (actual) values, i.e. whether a patient who triggers a response truly is at risk of deterioration or morbidity (as we will discuss in Sections 6.2 and 6.3). We define the correct classifications of each set of observations as being either true positive (TP) or true negative (TN). Incorrect classifications are defined as either false positive (FP) or false negative (FN).

Table 6.1: A confusion matrix for the evaluation of EWS, showing the relationship between test outcomes and evaluation metrics. TP = true positive; TN = true negative; FP = false positive; FN = false negative.

		True diagnosis		
		Positive	Negative	
Model classification	Positive	TP	FP	Positive Predictive Value $\frac{TP}{TP+FP}$
	Negative	FN	TN	Negative Predictive Value $\frac{TN}{TN+FN}$
		Sensitivity $\frac{TP}{TP+FN}$	Specificity $\frac{TN}{TN+FP}$	Accuracy $\frac{TP+TN}{TP+TN+FP+FN}$

- **Positive predictive value (PPV):** the probability that a patient who triggers a response truly is at risk.
- **Negative predictive value (NPV):** the probability that a patient who does not trigger a response truly is not at risk.
- **Sensitivity (true positive rate):** a measure of the proportion of patients at risk who would trigger a response. Good sensitivity is desirable in order to ensure that all patients at risk of deterioration or morbidity are correctly identified as such.
- **Specificity (true negative rate):** a measure of the proportion of patients not at risk who did not trigger a response. Good specificity is desirable in a scoring system to minimise unnecessary cost and workload, as well as minimising the emotional burden placed on the patient by preventing unnecessary investigations. (1-specificity) is a measure of the *false alarm rate*, i.e. the proportion of patients that triggered an EWS response unnecessarily.
- **Accuracy:** the proportion of correctly identified cases within all examples.

An ideal scoring system would have a sensitivity and specificity close to 100%, such that most (or all) all triggered patients are correctly identified as having deteriorated, and very few have misleading triggers. In practice, it is necessary to establish a trade-off between

between sensitivity and specificity. This is commonly done using a Receiver-Operating-Characteristics (ROC) curve [182]. The ROC curve is computed by varying the threshold M of the scoring system and plotting all possible resulting values of sensitivity and specificity along the curve $sensitivity(M) = 1 - specificity(M)$. The area under the ROC curve (AUROC) is used to assess the *discriminatory ability* of the system, i.e. whether it correctly discriminates between true and false cases independently of a fixed threshold value. In practice the minimum AUROC is 0.5; this would be represented by a diagonal line from (0,0) to (1,1) – see Figure 6.1 below – and is equivalent to the discriminative power of assigning class labels by random guessing. An efficient scoring system is characterised by an AUROC close to 1. Figure 6.1 shows an example ROC curve adapted from Carle *et al.* [30] – this study will be discussed in more detail in Section 6.2.

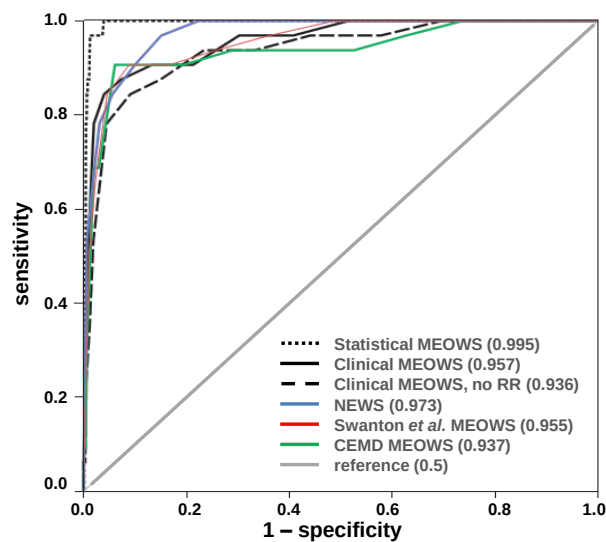


Figure 6.1: Example ROC curves adapted from Carle *et al.* [30]. The AUROC value associated with each curve is shown in brackets in the figure legend.

6.2 MEOWS reported in the literature

Having discussed the general purpose, fundamental design and key concepts for evaluation of EWS, we will now narrow our focus to examine EWS specifically designed for the obstetric population. Several different descriptions have been used to describe obstetric scoring systems but for simplicity, we will refer to all of these as Modified Early Obstetric Warning Scores (MEOWS), as initially proposed by the 2007 CEMD committee [1]. In

this section we will examine the design and validation of existing MEOWS in greater detail. Based on our observations, we will then define some requirements for the design of a novel evidence-based MEOWS.

6.2.1 Design of existing MEOWS

As a representative sample of the various MEOWS proposed to date, we identified 15 MEOWS scoring systems from clinical guideline publications and searches in the PubMed medical journal database. The majority of these were adapted either from the MEOWS published in the 2007 CEMD report [1] or from the NEWS score [27]. Consequently, they used either colour-coded trigger bands or numerical aggregate scoring triggers. The MEOWS are summarised in Table 6.2, where scores A to G are colour-coded charts and scores H to O are aggregate scoring systems.¹ The table summarises the vital signs included (HR, SpO₂, SBP, DBP, Temp, RR) and the cut-offs defined for each variable (i.e. whether for each variable an upper cut-off, a lower cut-off or both were defined). Figures 6.2a to 6.2f show the cut-off values and number of scoring bands for each MEOWS.

Of the 14 scoring systems considered here, we were only able to identify a description of the design process for three of the scores: the NEWS score (K) and the statistical (L) and clinical (M) scores developed by Carle *et al.*

The NEWS score [27] was derived from the Vital-PAC Early Warning Score (ViEWS) designed by Prytherch *et al.* [28]. ViEWS was derived from a large database of vital sign measurements, collected using the Vital-PAC patient monitoring system from a general adult population with acute medical admissions. To set the cut-off values and trigger thresholds, they used an approach described as *"an iterative, pragmatic, 'trial and error' process, where the thresholds were manually adjusted [guided by the AUROC metric] to maximize the ability of the score to discriminate in-hospital death within 24 hours of a vital signs dataset"*. The choice of seven bands with weightings of $T = \{3, 2, 1, 0, 1, 2, 3\}$ was

¹Note that scores K and N are in fact EWS for the non-obstetric population, but have been included here as they have been tested in the obstetric population

Table 6.2: Overview of MEOWS from the literature. The arrows $\uparrow$ and $\downarrow$ indicate use of an upper and lower parameter cut-off value, respectively; $-$ indicates that the parameter is not scored. Additional physiological parameters included in each score (if any) are not shown.

Track-and-trigger scores									
Escalation if score ≥ 2 ambers or ≥ 1 red									
No.	Year	Description	HR	SpO ₂	SBP	DBP	Temp	RR	
A	2007	CEMD [1]	$\uparrow\downarrow$	$\downarrow$	$\uparrow\downarrow$	$\uparrow$	$\uparrow\downarrow$	$\uparrow\downarrow$	
B	2009	Swanton <i>et al.</i> [33]	$\uparrow\downarrow$	$\downarrow$	$\uparrow\downarrow$	$\uparrow$	$\uparrow\downarrow$	$\uparrow\downarrow$	
C	2009	Sandwell and Birmingham NHS Trust [183]	$\uparrow\downarrow^\dagger$	$\downarrow$	$\uparrow\downarrow$	$-$	$\uparrow\downarrow$	$\uparrow\downarrow$	
D	2012	Singh <i>et al.</i> [35]	$\uparrow\downarrow$	$\downarrow$	$\uparrow\downarrow$	$\uparrow$	$\uparrow\downarrow$	$\uparrow\downarrow$	
E	2012	Winter <i>et al.</i> [66]	$\uparrow\downarrow$	$\downarrow$	$\uparrow\downarrow$	$\uparrow$	$\uparrow\downarrow$	$\uparrow\downarrow$	
F	2015	Maguire <i>et al.</i> [184]	$\uparrow\downarrow$	$\downarrow$	$\uparrow\downarrow$	$\uparrow$	$\uparrow\downarrow$	$\uparrow\downarrow$	
G	2015	Oxford MEOWS	$\uparrow\downarrow$	$\downarrow$	$\uparrow\downarrow$	$\uparrow\downarrow$	$\uparrow\downarrow$	$\uparrow\downarrow$	
† HR scored amber if HR > SBP									
Aggregate scores									
Medium risk aggregate score $\geq T1$.									
High risk aggregate score $\geq T2$ OR single parameter score $\geq T3$.									
No.	Year	Description	HR	SpO ₂	SBP	DBP	Temp	RR	T1 T2 T3
H	2009	St John's Hospital Livingston (SEWS) [183]	$\uparrow\downarrow$	$\downarrow$	$\uparrow\downarrow$	$-$	$\uparrow\downarrow$	$\uparrow\downarrow$	4 6 $-$
I	2011	Liverpool Women's NHS Foundation Trust [36]	$\uparrow\downarrow$	$-$	$\uparrow\downarrow$	$-$	$\uparrow\downarrow$	$\uparrow\downarrow$	4 6 $-$
J	2011	The Shrewsbury and Telford Hospital NHS Trust [36]	$\uparrow\downarrow$	$\downarrow$	$\uparrow\downarrow$	$-$	$\uparrow\downarrow$	$\uparrow\downarrow$	4 6 $-$
K*	2012	NEWS [27]	$\uparrow\downarrow$	$\downarrow$	$\uparrow\downarrow$	$-$	$\uparrow\downarrow$	$\uparrow\downarrow$	5 7 3
L	2013	Carle <i>et al.</i> (statistical) [30]	$\uparrow$	$-$	$\downarrow$	$-$	$\downarrow$	$\uparrow$	4 6 3
M	2013	Carle <i>et al.</i> (clinical) [30]	$\uparrow\downarrow$	$-$	$\uparrow\downarrow$	$\uparrow$	$\uparrow\downarrow$	$\uparrow\downarrow$	4 6 3
N*	2014	Austin <i>et al.</i> [185]	$\uparrow\downarrow$	$-$	$\uparrow\downarrow$	$-$	$\uparrow\downarrow$	$\uparrow\downarrow$	nr nr nr
O	2014	Paterson-Brown <i>et al.</i> [17]	$\uparrow\downarrow$	$-$	$\uparrow\downarrow$	$-$	$\uparrow\downarrow$	$\uparrow\downarrow$	nr nr nr
* K and N were EWS designed for the general population that were tested on pregnant women.									
nr = thresholds not reported									

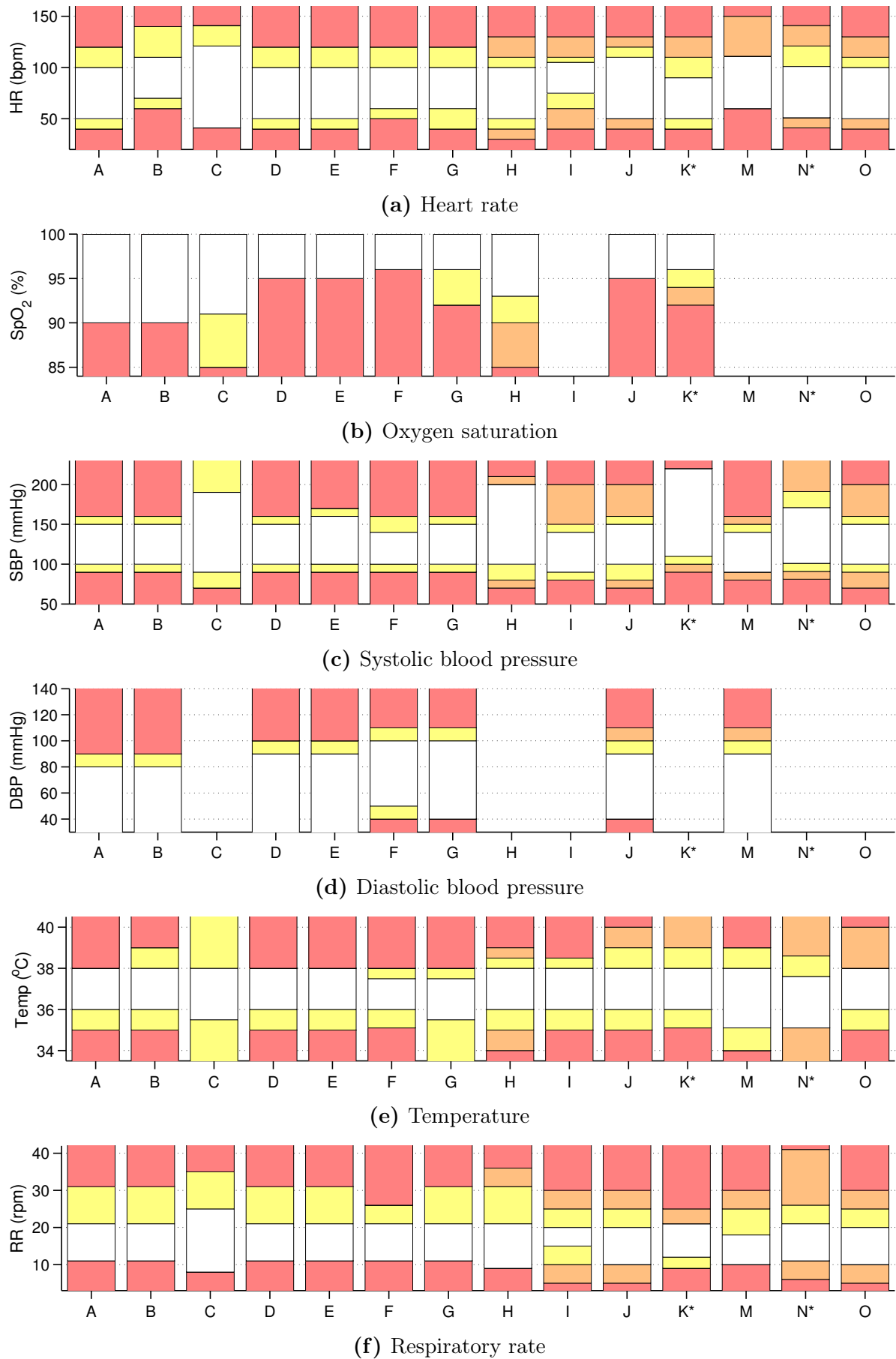


Figure 6.2: Comparison of MEOWS cut-off values from the literature for each of the six vital signs. The labels A-O refer to Table 6.2. The colours represent the scoring bands for normal (white) to very abnormal (red) vital sign values. Scores showing blank entries did not include the given vital sign(s). Score L is not shown because it did not fit a chart format.

made after observing that no additional benefit was gained by increasing the complexity of the scoring system [28]. To create NEWS, the NEWSDIG group² subsequently adjusted the ViEWS thresholds, based on the clinical opinion, to discriminate patients at risk of cardiac arrest, unanticipated ICU admission, or death within 24 hours [27].

The two scores proposed by Carle *et al.* [30] were designed specifically for the pregnant population using ICNARC data, a large dataset of obstetric admissions to the ICU. Their aim was to design a score to predict mortality from the most abnormal measurements from within the first 24 hours following critical care admissions. The two scores they proposed were developed in three steps: (i) statistically significant predictors were selected by univariate analysis; (ii) the statistically significant variables were entered into a multiple logistic regression model with mortality as outcome to decide the weightings of each variable; and, (iii) the impact of each variable was assessed using the AUROC metric. The statistical MEOWS (L) used the statistically optimal regression coefficients to weight each covariate to create an aggregate score. While this score was statistically optimal, it was complicated to compute and it did not include all of the vital signs recommended for monitoring [10, 24]. A clinical score (M) was therefore proposed, where non-significant variables were added and the ranges and weightings were adjusted empirically. To allow for ease of use, this score was presented on a chart which resembled the NEWS score [27].

Overall, as summarised in Table 6.2, we found that all of the scores included the four vital signs recommended by the 2007 CEMD report (HR, SBP, Temp, RR) [1]. In spite of the recommendations for the inclusion of SpO₂ and DBP made by NEWSDIG [27] and recent surveys of obstetric experts [32, 33], respectively, these vital signs were frequently omitted. Moreover, not all scores specified a lower cut-off for DBP. In addition to the vital sign variables shown in Table 6.2, each MEOWS included a number of additional clinical and physiological variables which we have not included in the comparison. The additional variables were either qualitative (subjectively assessed) features, such as assessment of level of consciousness, pain and amniotic fluid colour, or quantitative measurements, such

²National Early Warning Score Design and Implementation Group; a panel of experts established by the Royal College of Physicians

as proteinuria, urine output or blood glucose values.

Figures 6.2a to 6.2f show considerable variation in the cut-off values used to define normal and abnormal values for each vital sign. Importantly, none of the MEOWS accounted for changes in vital signs across gestational age by adjusting the cut-off values according to the stage of pregnancy.

6.2.2 Validation and reported performance of existing MEOWS

In 2012, Singh *et al.* [35] presented the first prospective validation of a MEOWS system. To the best of our knowledge, this is the only prospective validation of a MEOWS score to date. Using a MEOWS (D) which resembled the CEMD recommended system (A) with some minor modifications, they sought to identify *morbidity* within 30 days of triggering as a primary end point, rather than death or ICU admissions. Their population consisted of 676 ante- and postpartum maternity in-patients whose gestational ages ranged from 20 weeks to 6 weeks postpartum. Two hundred women (30%) triggered a response (here defined by a single red trigger or any combination of two yellow triggers). A total of 86 (13%) women had morbidity according to the criteria defined by the authors, which included haemorrhage, hypertension and suspected infection amongst others. The most frequent triggers were high BP (42%), high HR (28%) and low BP (18%). High BP (hypertension), high HR (tachycardia), high RR (tachypnoea) and high Temp (pyrexia) were found to be the best indicators of maternal morbidity. Low BP (hypotension) was a less sensitive indicator of morbidity than expected. According to the authors, this was due to the trigger threshold for low BP being set too low, indicating the need of further refinement of cut-off values. Overall, for the purpose of identifying morbidity within 30 days following patient assessment the MEOWS was 89% sensitive (95% CI 81 to 95%), 79% specific (76 to 82%), with a positive predictive value of 39% (32 to 46%) and a negative predictive value of 98% (96 to 99%).

Later attempts to validate MEOWS have been retrospective in nature. Carle *et al.* [30] validated the performance of both their statistical (L) and clinical (M) MEOWS using

a previously unseen proportion of the ICNARC dataset with mortality as an end-point. For the 2,200 women in their test-set, the mortality rate for all obstetric admissions was 2.8%. The gestational ages of the women were not specified, but both ante- and postnatal ICU admissions were included. The reported AUROC of the statistical and clinical MEOWS was 99.5% (95% CI 99.2 to 99.8%) and 95.7% (92.3 to 99.1%), respectively. For comparison, they also tested the discriminatory power of the MEOWS presented by CEMD 2007 (A), Swanton *et al.* (B) and the non-obstetric NEWS score (K) using the same data. The AUROC for these scores were 93.7% (88.4 to 99.1%), 95.5% (92.2 to 98.8%) and 97.3% (95.7 to 98.9%), respectively. Overall, the performance of the different scores was similar (Figure 6.1). While each of these scores used slightly different cut-off values and trigger thresholds, the finding illustrates that all of these scores are capable of identifying women at risk of mortality from within a high-risk population. However, it is questionable whether these results can be extrapolated to the general ward.

Alternative attempts to investigate the efficiency of MEOWS scores have been described by Austin *et al.* [185] and Maguire *et al.* [184]. With the aim of assessing whether the use of an EWS may have triggered earlier detection of morbidity and escalation of care, Austin *et al.* [185] retrospectively plotted the vital signs of 64 women admitted to ICU or high dependency unit on an EWS chart (N). They concluded (by clinical group consensus) that the EWS could have enabled earlier detection of deterioration and contributed to reducing the seriousness of maternal morbidity in five cases (7.6%). This included three admissions for sepsis and two admissions for post-partum haemorrhage. Maguire *et al.* [184] investigated whether the introduction of MEOWS (F) improved recording of vital signs among pregnant women with proven maternal bacteremia.³ In a before-and-after study they found that the introduction of MEOWS was associated with an improvement in the recording of vital signs, particularly RR. They also found that using a MEOWS to record vital signs led to a non-statistically significant decrease in the time between MEOWS trigger and administration of antibiotics. However, with only 60 and 17 included

³Bacterial growth on blood culture, a subset of sepsis

cases of bacteremia before and after the introduction of MEOWS, respectively, the study was too small to assess whether improved recordings led to improved clinical outcomes.

Finally, a recent review by Edwards *et al.* [38] sought to retrospectively compare the predictive power of published MEOWS for the development of severe sepsis in women with chorioamnionitis.⁴ They investigated six MEOWS; three colour-coded (B,D,E) and three more complex aggregate scores (I,J,M). From 913 cases, they were able to extract complete sets of physiological variables (HR, RR, BP, Temp and consciousness level) from 364 cases, of which five women developed severe sepsis. The AUROC of the systems ranged from 52% (D and E) to 72% (I). The PPV of all systems were very low (<2 to 15%) which meant that sepsis was overdetected. The authors hypothesised that modification of the cut-off values of MEOWS used to enhance the predictive power for other general obstetric conditions (e.g. hypertension) might have contributed to reducing the PPV for the specific condition investigated. We further note that as all of the women had proven chorioamnionitis, they were already defined to be at high risk of developing complications.

We were not able to identify reports of validation of the remaining MEOWS (C, G, H and O). This may be surprising, considering that three of these MEOWS (C, H and O) were actually referenced in clinical guidelines and audits [17,66,183]. Overall, the limited number of validation studies identified, in particular prospective studies, illustrate the urgent need for further development and validation of MEOWS.

6.2.3 Implications for future design of MEOWS

In a recent editorial, McGlennan and Sherratt [22] comment on the current state of MEOWS, stating that *"There has been no nationally unified chart, no convention as to the model used, no agreement on the parameters, no body of evidence exhibiting their efficacy, and no decisions made regarding an escalation pathway."* Our findings of significant heterogeneity in the design of MEOWS and a scarcity of validation studies support this

⁴Acute inflammation of the membranes and placenta diagnosed by a maternal temperature $\geq 38^{\circ}\text{C}$, uterine tenderness, maternal/fetal tachychardia and/or foul-smelling amniotic fluid

statement. With the exception of the two scores proposed by Carle *et al.* [30], the MEOWS identified from the literature were not based on statistical evidence, but rather designed based on empirical evidence and clinical consensus [32].

The work by Carle *et al.* represents a first attempt at designing an evidence-based MEOWS. Given that pre-admission data from the obstetric population are not widely available, they suggest that the ICNARC data currently represent the best and most comprehensive data from the sick obstetric population. However, critics of their approach have pointed out that neither the use of data from the ICU (where patients are likely to have had multiple interventions), nor the choice of mortality as an end-point, provide a suitable basis for a MEOWS whose aim is to identify women at risk of morbidity [22]. As emphasised in a recent editorial by Smith and Prytherch [186], the score should, as a minimum, be tested statistically against data from the general obstetric in-patient population before implementation in the general obstetric ward is considered. Moreover, gestational age was not accounted for. It is unclear whether the admissions included in the dataset occurred during the antepartum or postpartum period. As we have shown that pregnancy is associated with considerable physiological changes, the effects of further refinement of the score by accounting for gestational age should be investigated.

Both the score by Carle *et al.* [30] and the NEWS score [27] used binary outcomes (such as death/survival) to determine the vital sign cut-offs and trigger thresholds. Consequently, the range and cut-off values for each vital sign are set such that alerts predict a binary outcome at some later time. As observed by Tarassenko *et al.* [158], this approach may discriminate against patients with redeemable deterioration. This point is further emphasised by Pimentel [26], who states that, *"The true benefit of any EWS should be its ability to recognise patients who are deteriorating but who can have their outcome changed by a timely intervention, rather than identify those who are very likely to die independent of intervention"*. In the next section, we will therefore examine an alternative approach for design of MEOWS based on the use of the established normal ranges of vital signs to identify abnormality.

6.3 Centile-based early warning scores

To design a scoring system to alert for abnormal vital signs and general physiological deterioration, regardless of any specific outcome, Tarassenko *et al.* [158] proposed an alternative method based on the analysis of statistical distributions of vital signs. The key idea behind their approach was that clinicians would want to be alerted when their patients no longer conform to the limits of normal physiological behaviour, as defined by the extreme centiles of the statistical vital sign distributions.

Their centile-based early warning score (CEWS) was completed using 64,000 hours of data continuously collected from the bedside monitors of 863 acutely-ill hospitalised in-patients. Three levels of alerts were used to alert for abnormality. A score of $T = 1$ was assigned to vital signs on the borderlines of abnormal, defined to be between the 5th and 10th or between the 90th and 95th centiles. A score of $T = 2$ was assigned to vital signs between the 1st and 5th or between the 95th and 99th centiles. Finally, a score of $T = 3$ was assigned to vital signs below the 1st or above the 99th centiles.

6.3.1 Scoring bands of a centile-based MEOWS

In line with the principles of CEWS developed for the general adult population, we define a centile-based MEOWS (C-MEOWS) for the maternity population using the centiles derived from the 4P data. In total, we considered 2,645 observation sets collected from 439 women during the antepartum period, as described in Section 3.3. The centiles were modelled using a multilevel modelling approach, as described in Sections 5.3 and 5.4.

For HR, SBP, DBP, Temp and RR we define seven scoring bands, using the 1st, 5th, 10th, 90th, 95th and 99th centiles as described in the previous section. Note that these cut-offs could have been defined in terms of Z-scores as well, but the use of centiles is preferred to facilitate clinical interpretation. As SpO₂ has a one-sided distribution (as discussed in Section 5.3.4), we do not define an upper cut-off for SpO₂ and only use the 10th, 5th and 1st centiles as cut-offs to assign scores of $T = 1$, $T = 2$ and $T = 3$, respectively. The resulting C-MEOWS scoring bands are shown in Figure 6.3.

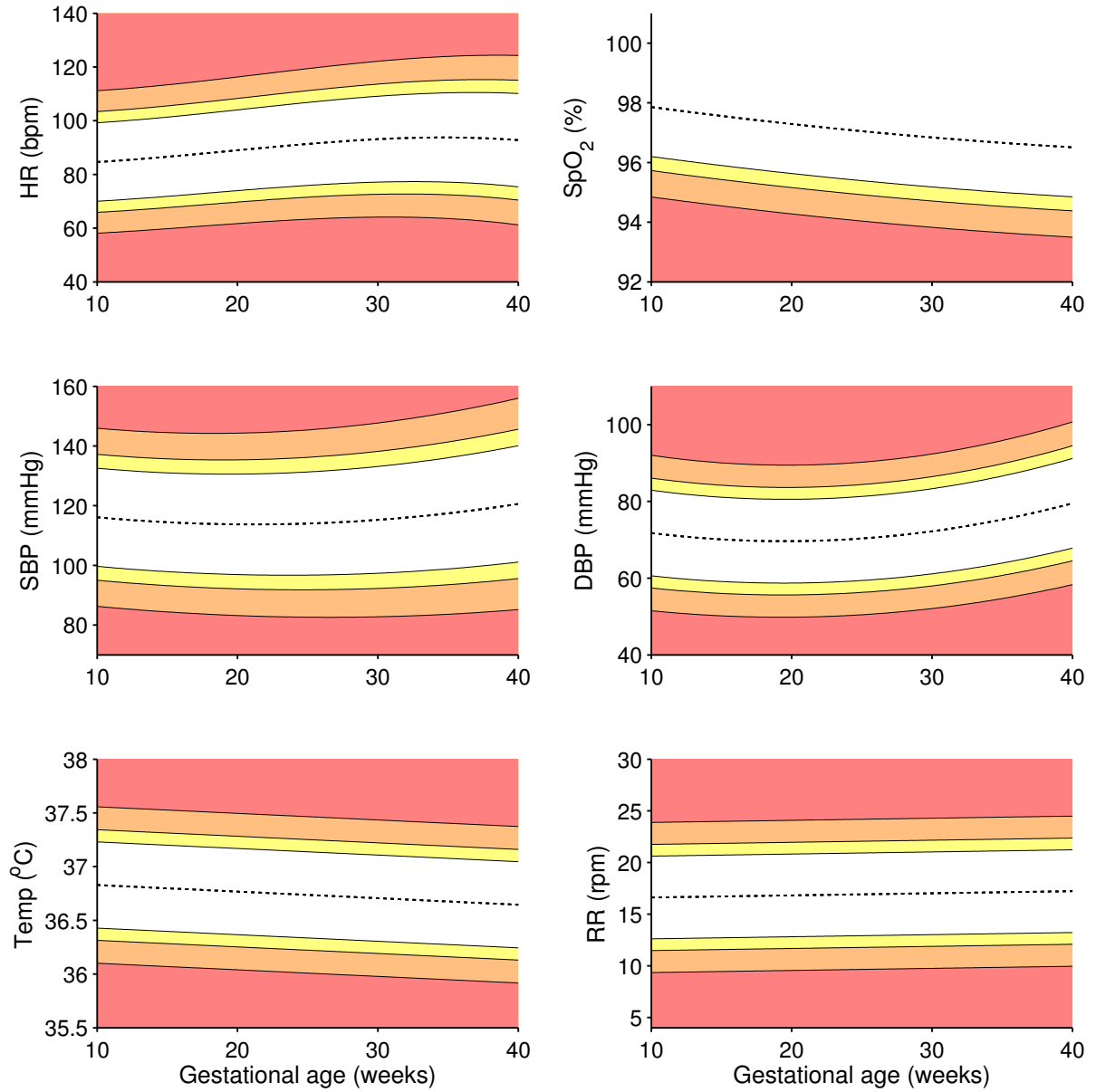


Figure 6.3: Gestation-specific C-MEOWS scoring bands for HR, SpO₂, SBP, DBP, Temp and RR. For all six vital signs, areas shaded in white, yellow, orange and red indicate scores of 0, 1, 2 and 3, respectively. The dotted black lines indicate the expected population mean values. For HR, SBP, DBP, Temp and RR, the solid black lines indicate the cut-off values at the 1st, 5th, 10th, 90th, 95th and 99th centiles. As SpO₂ has a one-sided distribution, cut-offs are only defined at the 1st, 5th and 10th centiles.

In what follows, we investigate the aggregate scores and effects of potential trigger thresholds resulting from this system. Further, we illustrate the use of the C-MEOWS scoring bands by means of patient examples, before concluding with a discussion on the requirements for a more thorough, future validation of C-MEOWS in Section 6.4.

6.3.2 Aggregate scores from the centile-based MEOWS

Following application of the centile-based cut-offs, aggregate C-MEOWS scores were computed for each set of vital sign observations by summing the scores generated across the six vital signs HR, SpO₂, SBP, DBP, Temp and RR. As this summation requires complete observation sets of all six vital signs, we imputed the missing values in observation sets with one or two missing variables. This was done by replacing the missing variable(s) by the population mean value for the corresponding variable(s). In doing so, we assume that any missing variable does not contribute to the final score (equivalent to assigning a score of $T = 0$ to the missing variable). Observation sets with more than two missing variables were excluded from the analysis.

Of the initial 2,645 observation sets included in our analysis (Section 3.3), we found that 2,068 (78.2%) observation sets contained measurements of all six vital signs. An additional 499 (18.8%) and 55 (2.1%) observation sets contained only one and two missing variables, respectively, resulting in a total of 2,622 (99.1%) observation sets with two or fewer missing variables. Additionally, as centiles are only defined for the period 10 to 40 weeks of gestation, we only included the 2,549 (96.4%) observation sets made within this period. Following the application of these exclusion criteria, 2,527 (95.5%) observation sets from 438 women remained for analysis. Note that this investigation is based on in-sample data. Validation of C-MEOWS using out-of-sample (external) data will be discussed in Section 6.4.

Figure 6.4 shows the distribution of aggregate C-MEOWS scores computed across these observation sets. A total of 1,103 (43.7%) of the observation sets had a score of $M = 0$. The maximum aggregate score was $M = 10$ (assigned to 1 observation set).

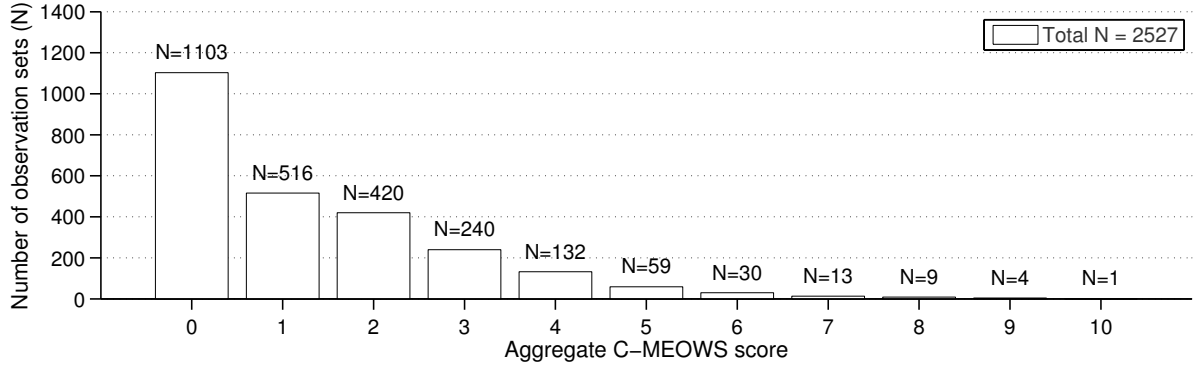


Figure 6.4: Number of appointments per C-MEOWS aggregate score

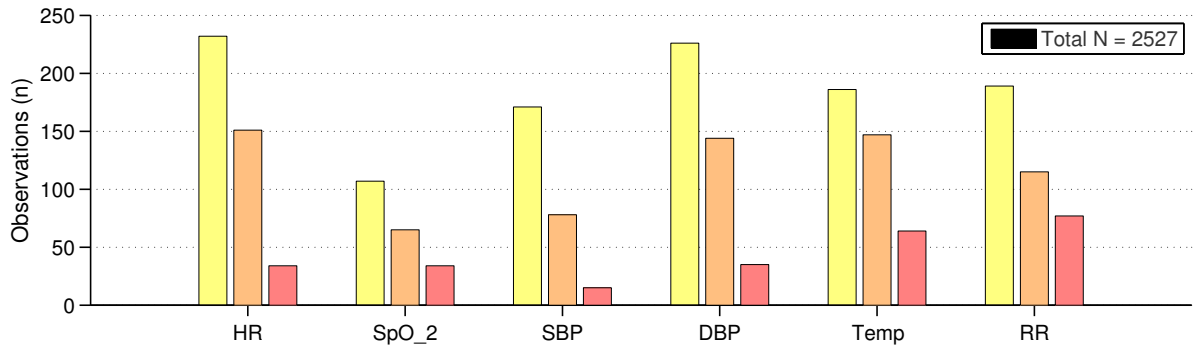


Figure 6.5: Number of observations of each vital sign assigned scores of $T = 1$ (yellow), $T = 2$ (orange) and $T = 3$ (red), respectively, across all 2,527 sets of vital sign observations.

Figure 6.5 shows the distribution of single parameter scores $T = \{1, 2, 3\}$ for each of the six vital signs. HR and DBP were most frequently assigned scores of $T = 1$, in 232 (9.2%) and 226 (8.9%) of the observation sets, respectively. A total of 237 (9.3%) of the observation sets contained at least one vital sign observation with the maximum single parameter score of $T = 3$. RR was most frequently assigned a score of $T = 3$, in 77 (3.0%) of the observation sets. The fact that RR was frequently assigned the maximum score of $T = 3$, as compared to the other vital signs, could reflect artefacts in the methods of electronic estimation of RR. As shown in Section 4.5, the accuracy of RR estimates is strongly dependent on the quality of the associated respiratory waveform and extreme estimates of RR may be associated with artefacts in the respiratory waveform.

6.3.3 Trigger frequency for different C-MEOWS thresholds

The frequency at which women trigger an alert using C-MEOWS is determined by (i) the cut-off values for each variable's scoring bands; (ii) the associated weightings T ; and (iii) the aggregate C-MEOWS trigger threshold M . Given that the cut-off values and weights are set to the values described in Section 6.3.1, we investigated the frequency at which an alert would be triggered with different trigger thresholds of $M \geq 4$ and $M \geq 6$, respectively (as previously shown in Table 6.2, thresholds of $M \geq 4$ and $M \geq 6$ were commonly used by the MEOWS identified in the literature to trigger an intervention).

From Figure 6.4, we found that 248 (9.8%) and 57 (2.3%) observation sets were assigned aggregate scores of $M \geq 4$ and $M \geq 6$, respectively. As each woman had a median (IQR) number of 6(1) vital sign observation sets, some women triggered an alert on more than one occasion during their pregnancy. In total, 162 (37.0%) women achieved an aggregate score of $M \geq 4$ on at least one occasion. If the threshold was increased to $M \geq 6$, 51 (11.6%) women triggered a response on at least one occasion. By comparison, in the MEOWS validation study by Singh *et al.* [35], 30% of women were found to trigger an alert based on observation of vital signs at some point during pregnancy.

Figure 6.6 shows the distribution of single parameter scores $T = \{1, 2, 3\}$ for each of the six vital signs in the 248 observation sets with an aggregate score $M \geq 4$. From this figure, we see that SBP, DBP, Temp and RR, which we collected as discrete measurements, were more likely to be assigned high scores of $T = 2$ and $T = 3$ compared to HR and SpO₂, which we obtained as average measurements over an extended period of time (see Chapter 3 for description of how the measurements were obtained). Guidelines for measurement of BP commonly advise using the mean of multiple measurements to obtain an average BP measurement, rather than just a single measurement, in order to minimise the occurrence of potentially false alerts associated with single measurements [187]. Future work and MEOWS guidelines should therefore consider the effect of repeated measurements of SBP, DBP, Temp and RR, where a repeated measurement should be advised to confirm any abnormal measurement.

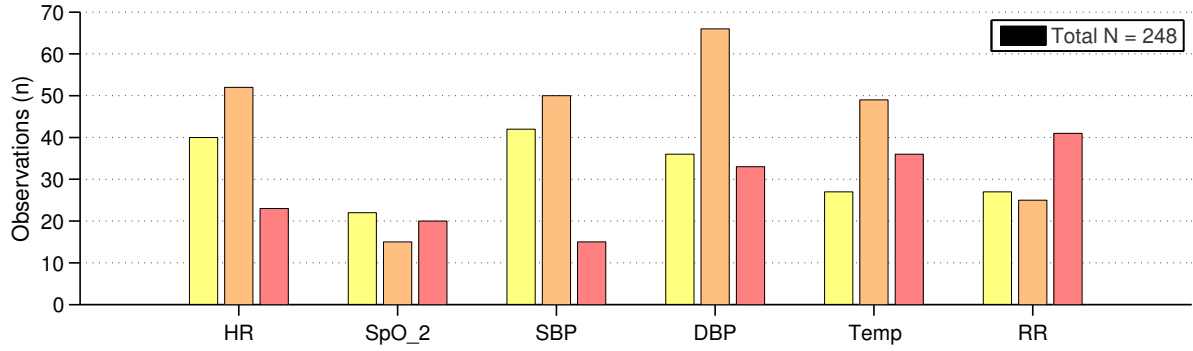


Figure 6.6: Number of observations of each vital sign assigned scores of $T = 1$ (yellow), $T = 2$ (orange) and $T = 3$ (red), respectively, across the 248 sets of vital sign observations with an aggregate score $M \geq 4$.

In Section 6.2 we found that many of the MEOWS currently presented in the literature only include one BP variable (usually SBP). The values of SBP and DBP are correlated, thus including both as variables in the aggregate score may make the score overly sensitive to abnormal values of BP [27,30]. In our dataset, we found that within the 248 observation sets assigned an aggregate score of $M \geq 4$, SBP and DBP on their own contributed a score $T_{SBP} + T_{DBP} \geq 4$ in 54 (21.8%) of the observation sets. As DBP has previously been shown to be a strong indicator of hypertensive disorders of pregnancy and of haemorrhage [34,35], omitting DBP from a MEOWS score altogether seems ill-advised. However, future work should consider whether it is necessary to adjust the relative weightings of SBP and DBP to avoid designing a MEOWS that is overly sensitive to abnormal values of BP.

6.3.4 Example applications of C-MEOWS

To illustrate the application of the C-MEOWS in comparison to the Oxford MEOWS scoring bands currently used within the OUH NHS Trust, we extracted vital sign data retrospectively from the hospital charts of three randomly selected women who were not part of the 4P study. All three women had been diagnosed with hypertensive disorders of pregnancy. Their HR, SBP and DBP observations have been plotted in Figures 6.7 to 6.9. The different colours indicate whether the vital sign measurements were collected during routine antenatal appointments (blue dots) or from MEOWS charts (red crosses),

as well as the timings of any treatment intervention (black dotted lines) and the time of delivery (green solid line). At present, MEOWS charts are typically only used when the a patient is admitted to the hospital. A brief description of each case is also provided below.

Patient 1 (Figure 6.7): Woman diagnosed with gestational hypertension and early signs of pre-eclampsia

- Between 20 to 30 weeks of gestation: prescribed anti-hypertensive medication on several occasions following episodes of raised BP;
- 31⁺⁵ weeks of gestation: referred to the hospital by GP and admitted with raised BP and proteinuria (an indicator of pre-eclampsia). Subsequently discharged as she did not show any other symptoms but frequency of visits increased;
- 37⁺⁰ weeks of gestation: episodes of raised BP and induction of labour for risk of pre-eclampsia.

Patient 2 (Figure 6.8): Woman with chronic rheumatoid arthritis, diagnosed with pre-eclampsia

- 29⁺³ weeks of gestation: prescribed anti-hypertensive medication following episodes of raised BP;
- 31⁺⁰ weeks of gestation: admitted to hospital with headache and disturbed vision (symptoms of pre-eclampsia), no proteinuria;
- 32⁺⁶ weeks of gestation: dizzy and unwell, low BP;
- 35⁺⁷ weeks of gestation: admitted to hospital with headache and symptoms of pre-eclampsia, prescribed additional anti-hypertensive medication;
- 37⁺⁴ weeks of gestation: labour (scheduled cesarean section).

Patient 3 (Figure 6.9): Woman diagnosed with essential hypertension

- BP regulated with anti-hypertensive medication throughout pregnancy;
- 38⁺⁴ weeks of gestation: spontaneous delivery.

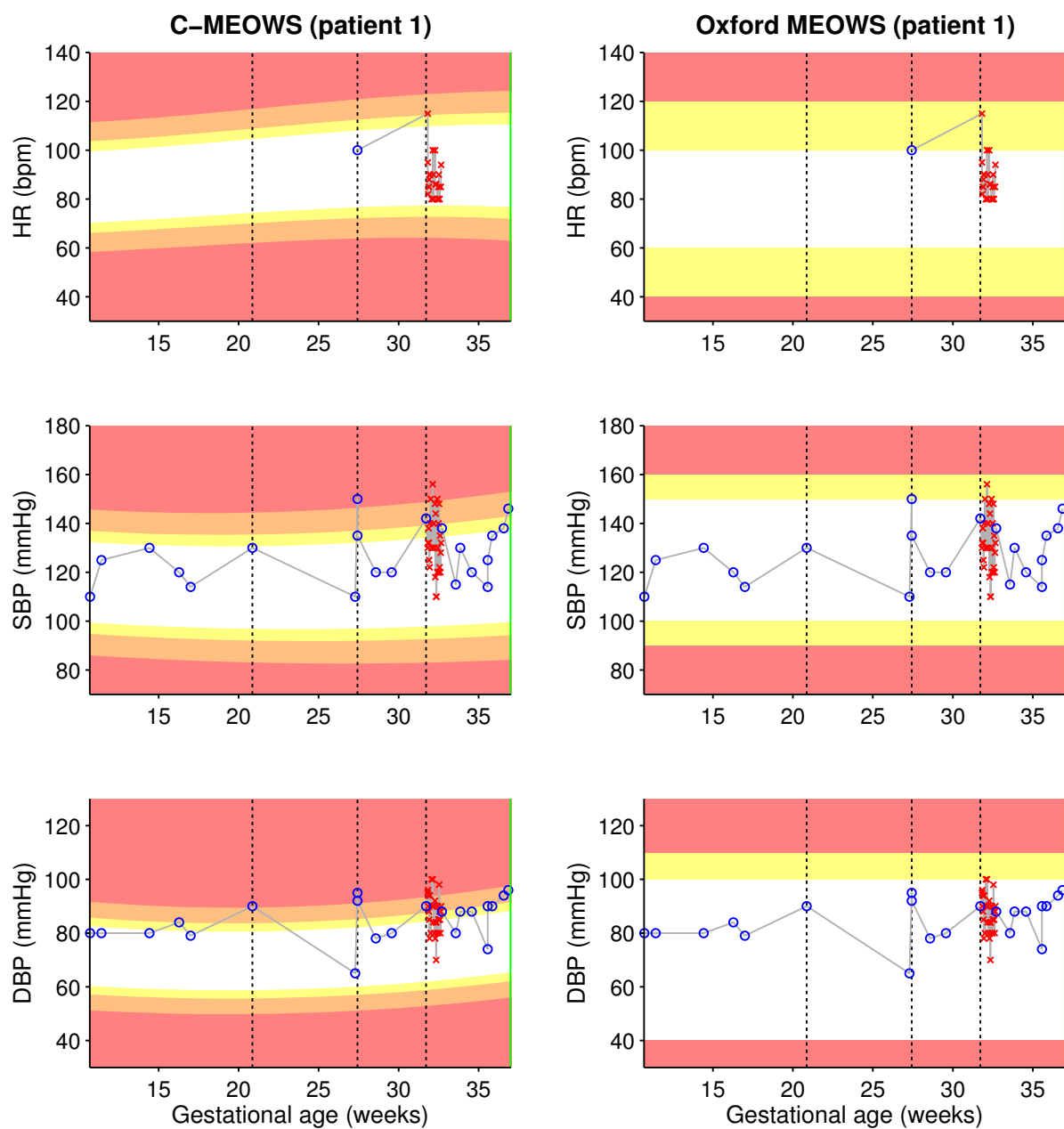


Figure 6.7: Example patient 1: Woman diagnosed with gestational hypertension and early signs of pre-eclampsia

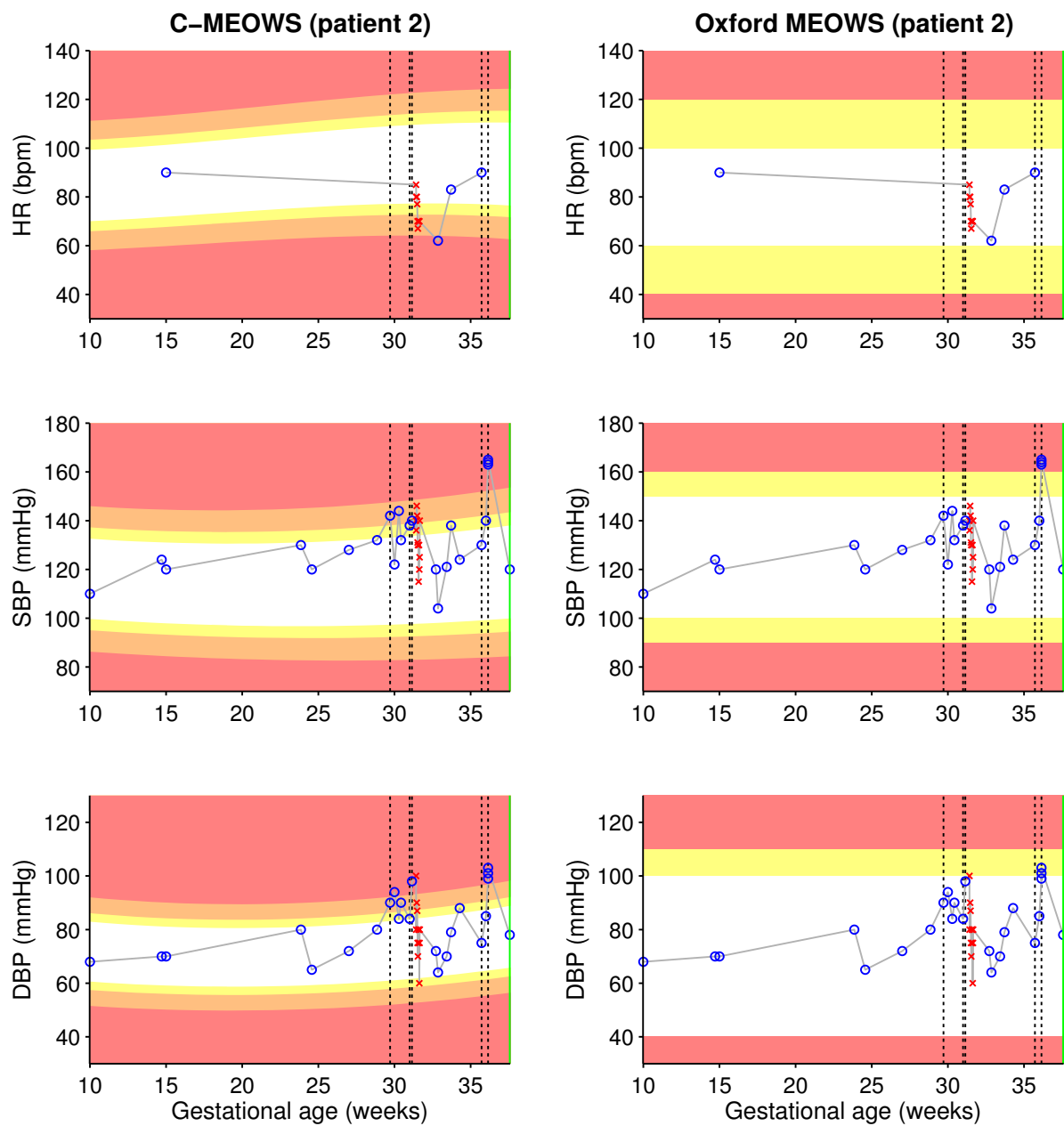


Figure 6.8: Example patient 2: Woman with chronic rheumatoid arthritis, diagnosed with pre-eclampsia

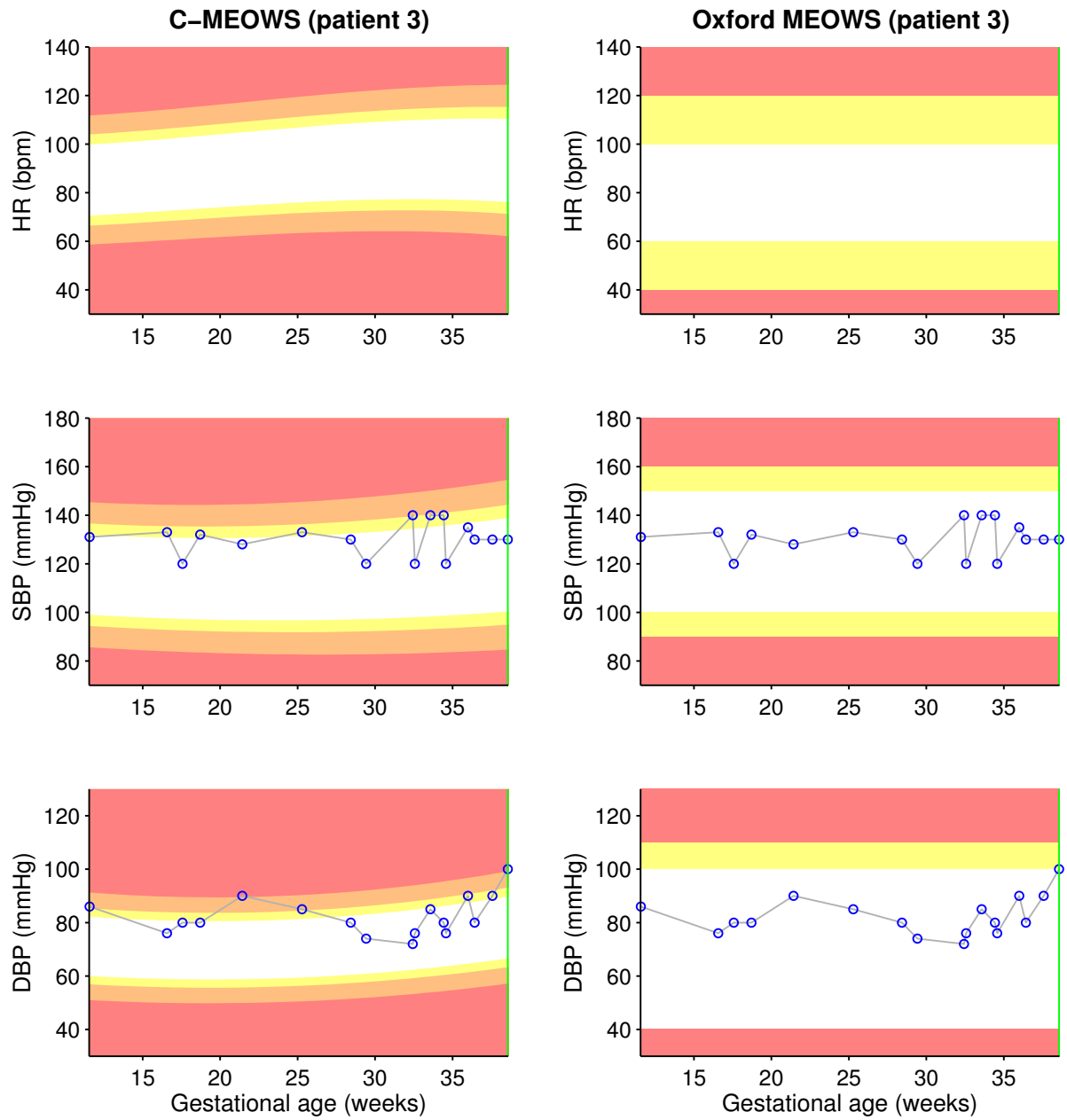


Figure 6.9: Example patient 3: Woman diagnosed with essential hypertension

From their hospital records, we found that patients 1 and 2 both showed early signs of pre-eclampsia⁵. For both women, the time-points of the alerts generated using the C-MEOWS cut-off values appear to align very well with the time points of interventions delivered by the healthcare professionals. In the period prior to the women being admitted to the hospital (shortly after week 30 of gestation, as indicated by the red crosses), both women alerted multiple times using the C-MEOWS cut-offs. In contrast, if we apply the Oxford MEOWS cut-off values over the same time period, neither woman generated any alert for high BP prior to hospital admission.

Patient 3 was diagnosed with essential hypertension prior to pregnancy and was using anti-hypertensive medication for the duration of her pregnancy. Her BP observations throughout pregnancy are marginally outside the C-MEOWS cut-offs used to define normality. She did not receive any additional intervention, but was followed up through frequent hospital appointments. Towards the end of her pregnancy, her recorded values of SBP and DBP increase. As the C-MEOWS cut-off values increase as well, her BP readings remain mostly in the yellow scoring band ($T = 1$) until delivery. This corresponds well with the assumption that the lowest alert level should trigger no intervention other than a repeated measurement of vital signs to confirm the abnormal value, or more frequent observations. This clearly helps to illustrate the potential advantage of, and need for, gestation-specific thresholds.

Overall, compared to the Oxford MEOWS (and indeed compared to the other scoring systems discussed in Section 6.2), the ranges of normal values ($T = 0$) defined by the C-MEOWS cut-offs are considerably narrower. This does imply that C-MEOWS, with the cut-off values proposed here, will generate more frequent alerts than the current Oxford MEOWS score. However, if the aim is to alert for high risk of morbidity and need for more frequent observations, these examples serve to demonstrate that the cut-offs proposed here may be reasonable. We have only provided examples using observations of HR, SBP and DBP (the only vital signs that are currently monitored during routine

⁵According to the ISSHP criteria [12], pre-eclampsia is diagnosed as raised BP $\geq 140/90$ mmHg and proteinuria on at least two occasions

appointments), for three women. Considerably more work is required to thoroughly validate MEOWS cut-off values and trigger thresholds.

6.4 Directions for future validation of C-MEOWS

The C-MEOWS presented in this chapter represents a first attempt at constructing an evidence-based MEOWS with gestation-specific cut-off values. The proposed cut-off values are based on the knowledge of normal ranges of vital signs, as described in Chapter 5, and the principles of the centile-based early warning score, as proposed by Tarassenko *et al.* [158]. We have not made any definite recommendations for thresholds to alert for referral to a higher level of care, or what a suitable response strategy should be. To validate the exact threshold values and relate the frequency of alerts to the frequency of complications, a more comprehensive investigation is required.

A major question which remains unanswered is what exactly MEOWS should warn of. McGlennan and Sherratt [22] propose several options: *"What should a system warn of? Mortality? – too rare. The need for ICU admission? – possibly, but still quite rare at 260 per 100 000 maternities. Serious morbidity? – maybe, but the definitions are woolly and it will be harder to validate the system."* The anecdotal patient examples presented here indicate that the centile-based approach, using the thresholds proposed by Tarassenko *et al.* [158], may be suitable for designing a score that alerts for the more common (and potentially less severe) pregnancy complications such as mild hypertension. The primary use of a MEOWS system should be the ability to spot the deteriorating patient. By focussing on prevention of morbidity and reversible deterioration, it may well be possible to prevent serious morbidity, avoid admission to ICU and reduce obstetric death rates.

As emphasised by Singh *et al.* [35], it is important to consider that MEOWS should be a screening tool to identify individuals *likely* to have morbidity. It is *not* a diagnostic test that seeks to confirm its presence definitively. This implies that when striking a balance between the goals *(i)* correct identification of the positive case; and, *(ii)* correct

identification of the negative case, it is more important to correctly identify all positive cases, rather than to correctly reject all negative cases.

For a screening tool to be of value, irrespective of the chosen outcome, it must be cost-effective, safe and validated [35]. The proposed C-MEOWS is based on a strong assumption that vital sign values which deviate from the statistically-derived normal indicate risk of subsequent deterioration. The current model, where an ROC curve is used to optimise alerting thresholds to discriminate between binary outcomes, may lead to misclassification of patients with reversible abnormality. Therefore, as previously emphasised by Subbe [188], a prospective clinical trial will be crucial in judging the clinical usefulness of the centile-based approach.

Nevertheless, as such a trial is likely to be very resource demanding, we would recommend a retrospective validation to be carried out first, using both in-sample and out-of-sample data. Such a validation would serve two purposes: firstly, it would be useful in order to help guide the decision of what weightings and scoring bands should be used and what the appropriate initial trigger thresholds should be. Secondly, a retrospective validation using out-of-sample data would provide an opportunity to investigate to what extent the centiles and cut-off values derived from the 4P dataset may be generalised to other study populations.

A major limitation of the work presented here is the lack of validation in a second (external) validation data set. The topic of further retrospective and prospective validation of MEOWS will be discussed in more detail under future work in Section 7.2.

Chapter 7

Conclusions and future work

The early detection of severe illness in mothers remains a challenge to all involved in their care. The relative rarity of such events, combined with the normal changes in physiology associated with pregnancy, compounds the problem [1]. The overall aims of the investigations presented in thesis were *(i)* to establish what the expected changes in vital signs (HR, SpO₂, SBP, DBP, Temp and RR) associated with normal pregnancy are; and, *(ii)* to use this knowledge to design the first evidence-based, gestation-specific MEOWS. In what follows, we will summarise the main findings of the work presented in this thesis and present directions for future work.

7.1 Contributions of this thesis

In Chapter 2, we presented a systematic review of studies describing changes in HR and BP with gestational age in *low-risk* pregnancies. To the best of our knowledge, this is the first attempt at a combined qualitative and quantitative synthesis of the current evidence base to describe changes in vital signs during the antepartum period. Importantly, we showed that pregnancy-related changes in BP and HR occur with gestational age. Further, we showed that the reported values of BP are affected by the choice of measurement device and parity of the subject. We also found a tendency towards an annual increase

in BP per year from the mid-1960s until today across the population of pregnant women. This tendency was quantified to 0.5mmHg/year, but this figure must be interpreted with care due to the limited number of available studies. Our findings were further limited by the quality of the studies, as the majority of the studies were assigned a low-quality score for criteria including population selection, use of appropriate measurement method and accurate determination of gestational age. We were not able to identify sufficient numbers of studies which described pregnancy-related changes in the other vital signs (SpO₂, Temp and RR), or which described changes during the intrapartum and/or immediate postpartum period, to include these in the systematic review.

In Chapter 3, we introduced the 4P study. The 4P study was set up to collect measurements of SBP, DBP, HR, SpO₂, RR and Temp from pregnant women during the antepartum, intrapartum and postpartum period. It is the first multi-centre study to collect measurements of all six vital signs longitudinally across the three stages of pregnancy. To ensure a high-quality data collection process, we designed the 4P digital health system. The system uses standard medical devices interfaced to a tablet computer to electronically collect measurements of vital signs, which are transmitted to the 4P database in real-time. The system has been used successfully to collect vital sign measurements by the study midwives and women monitoring at home. Steps for validation of vital sign measurements, required in order to obtain a single, robust value from each vital sign measurement, were presented. High-quality measurement data are an essential pre-requisite for the computation of normal ranges.

For the investigation presented in this thesis, we used a dataset consisting of 2,645 observation sets collected during the antepartum period from 439 women who had delivered by the 23rd November 2015. By the 11th February 2016, when this thesis was submitted, a total of 4,309, 524 and 6,099 observation sets had been collected during the antepartum, intrapartum and postpartum periods, respectively. This represents vital sign data collected from 719 women across three hospital sites, of whom 580 had delivered. With a projected recruitment rate of 20 women per month across the three sites, we aim for the collection

of vital sign measurements from 1,000 pregnant women to be completed by April 2017.

While HR, SpO₂, SBP, DBP and Temp are measured electronically, RR is the only vital sign that is currently measured manually in the antenatal clinic and hospital wards. In Chapter 4, we therefore investigated two novel methods for electronic measurement of respiratory rate: measurement of the chest-wall movement from breathing using the in-built accelerometer of a smartphone and measurement of the respiratory-induced HR variability using the PPG signal from a pulse oximeter. We assessed five signal processing algorithms for estimation of RR from the measured respiratory waveforms. To compare the algorithms, we used a reference RR computed from a nurse count of breaths over a time period of 15 seconds. At high RR, the most accurate estimates of RR were obtained from the mean interval between peaks of consecutive breaths (PK:mean, mean absolute errors of 3.2 and 4.6 rpm for estimates derived from the accelerometer and pulse oximeter, respectively). At low RR, the most accurate estimates were obtained by selecting the highest amplitude peak in the AR spectrum (AR:peak, mean absolute errors of 3.1 and 3.6 rpm for estimates derived from the accelerometer and pulse oximeter, respectively).

Using a signal quality index (SQI) we showed that it is possible to improve the estimation accuracy by only considering high-quality respiratory waveforms. The SQI was subsequently used to 'fuse' RR estimates from the PPG and accelerometer waveforms by selecting the estimate from the waveform of the highest quality as indicated by the SQI. The final 'fused' electronic measurement had a mean absolute error of 2.9 rpm relative to the reference RR. This estimate was used in the computation of normal ranges of RR in Chapter 5. The chapter concluded with a discussion about requirements for further refinement of the RR estimation signal processing algorithms, which include a more accurate reference RR of a count of breaths for 60 seconds. We also highlighted how the SQI should be used to assess the quality of the signal at the point of measurement.

In Chapter 5 we presented reference ranges for changes in HR, SpO₂, SBP, DBP, Temp and RR with gestational age, derived using data from the 4P study. To derive these ranges, we compared different statistical methods for computation of reference

ranges including empirical centiles, Gaussian processes and three different formulations of a multilevel model. Using a combination of criteria for assessment of centile fit, the best-fitting model was selected to be a conventional polynomial formulation of a multilevel model. This model includes a random intercept and slope to account for correlations in repeated measurements from the same individual.

Chapter 5 concluded with a comparison of subject-specific trajectories of HR, SBP and DBP for groups of women in our data set with different pregnancy-related conditions. These were obtained from the random-effect estimates, which describe the deviation in intercept and slope of an individual's trajectory as compared to the population mean trajectory (as shown in Figure 5.11). While the present investigation was limited by the number of women in our data set with different medical and pregnancy-related conditions, it highlighted how the multilevel model outputs can be used to estimate subject-specific vital sign trajectories. In Section 7.2.2, we will discuss potential extensions of this work including strategies for prospective prediction of physiological abnormality based on vital sign trajectories.

Finally, in Chapter 6 we presented a first attempt at designing a gestation-specific MEOWS chart. The proposed vital sign cut-offs are based on centiles of measured normal values of vital signs at each gestational age. This is in contrast to current MEOWS, which are mainly based on empirical evidence and clinical opinion. The alerting frequency associated with different aggregate score thresholds was investigated using in-sample data. By defining aggregate score thresholds of $M \geq 4$ and $M \geq 6$, 162 (37.0%) and 51 (11.6%) women, respectively, triggered a response at least once during their pregnancies. We did not propose any definite alerting threshold or strategy for responding to alerts generated by the proposed cut-offs. The use of the score was illustrated by means of vital sign observations from three randomly selected women. These examples, which used out-of-sample data, indicated that the proposed cut-offs could be reasonable starting points in order to predict the time points at which these women had received interventions or treatment solely based on vital sign data.

A major limitation of the work is the lack of validation and assessment of generalisability performed which should ideally be performed using out-of-sample data. Directions for a thorough retrospective and prospective validation of the MEOWS score will be discussed in Section 7.2.3.

7.2 Directions for future work

At the end of each chapter, we have suggested improvements to the analysis methods presented. Here, we discuss more general directions of future work.

7.2.1 Normal changes in vital signs during the intrapartum and postpartum period

Using data collected in the 4P study, reference ranges for normal changes in vital signs will be derived for the intrapartum and postpartum periods, using the same methods as those proposed for the analysis of antepartum data in this thesis. These data will provide insights into the dynamic vital sign changes associated with normal physiological recovery following pregnancy and childbirth. Additionally, the effect of covariates on the reference ranges should be investigated, including maternal weight/BMI, parity, maternal age and smoking status. Individual-specific covariates could be incorporated in the multilevel model using the methodology presented in Chapter 2.

7.2.2 Patient-specific vital sign trajectories

In Chapter 5 we introduced the concept of subject-specific trajectories of vital sign data. Following application of the multilevel model to data from the entire 4P study population, we illustrated how the model outputs could be used to assess the differences between different individuals' vital sign trajectories. The study of individuals' vital sign trajectories are interesting from at least two points of view.

Firstly, retrospective clustering of subject-specific vital sign trajectories could provide additional information about normal and abnormal vital sign patterns, which could subsequently be linked to patterns of disease development. This analysis could either be performed in a univariate framework or time-series of multiple variables could be combined in a multivariate framework. Such a framework could incorporate different vital sign data, as well as other clinical data. Examples of other clinical data include the use of proteinuria and BMI for prediction of pre-eclampsia [189] or white blood cell count and lactic acid tests for prediction of sepsis [190]. Previous studies have for example analysed vital sign trajectories in patients post-surgery [26]. This particular work by Pimentel [26] used Gaussian process regression to model vital sign trajectories, which were subsequently clustered using a kernel similarity measurement for comparing the underlying latent functions to provide a measure of patient-to-patient similarity. The use of Gaussian processes is attractive for this purpose because it enables modelling of change at different time scales without restricting the shape of the trajectories to a polynomial form.

In the field of genetics, clustering trajectories of biomarker data has been used to reveal information about different sub-types or phenotypes of diseases. Schulam *et al.* [191] used an approach based on multilevel modelling to remove nuisance measurement variability in time-series data and reveal underlying latent group-trajectories. In the context of pregnancy, such investigations could be particularly relevant to further our understanding of complex, multisystem disorders such as pre-eclampsia, for which the exact pathophysiology of the disease is still not fully understood, and to subsequently inform the design of more personalised approaches for prevention and treatment.

Secondly, the knowledge of individual trajectories of clinical measurements associated with different disease sub-types has been used in models for *prediction* of disease. Schulam *et al.* [192] have proposed an algorithm for learning population and latent disease-subgroup trajectories offline, combined with an online procedure for dynamically learning individual-specific parameters, which they used to predict the course of interstitial lung disease.

Another prognostic model framework has been proposed by Proust-Lima [193]. Both used models of underlying sub-group disease trajectories to predict the most probable future trajectories conditional on sets of observed measurement data. The use of a restricted number of latent group trajectories enables prediction when a limited number of observations is available. The main challenge of predictive models based on longitudinal data is that the predictive ability of the model depends on the number and time-span of the observed data. Predictive ability typically increases as more data points become available. This was emphasised by Macdonald-Wallis *et al.* [189], who in an attempt to predict the occurrence of pre-eclampsia showed that the addition of routinely collected BP measurements from 28 weeks onwards was required for accurate prediction of disease.

7.2.3 Plans for design, validation and implementation of MEOWS

With respect to the final design of MEOWS, several questions remain to be answered: What outcomes should MEOWS alert for? Which variables should be included? Which alerting thresholds should be used and what should the response strategies be?

To inform the initial design of MEOWS, a first step should be to complete a thorough retrospective validation. This will require defining what outcomes the score should alert for. The list of maternal morbidities proposed by Singh *et al.* [35] could be considered a reasonable starting point for this. To select which variables should be included, we discussed in Chapter 6 an approach based on optimisation of AUROC. This approach measures the *discriminatory power* of the scoring system, i.e. the ability to correctly distinguish between positive and negative cases. Cook [194] has argued that in order to select variables for inclusion in a risk assessment score, the estimated risk or predicted values, and how well these predict actual risk, may be an equally important model aspect as sensitivity and specificity, on which the ROC curve is based. This is known as *calibration* of the risk assessment tool, and resembles the procedure employed by Carle *et al.* [30] who used multiple logistic regression to select variables and associated weightings for inclusion in MEOWS.

As a second step, any proposed MEOWS should be assessed prospectively in a clinical intervention trial. Previously, we have seen examples of EWS for the non-obstetric population being validated in *before-and-after* studies, such as the validation study presented by Pullinger *et al.* [195] of the use of electronic EWS in the Emergency Department. A similar trial design could also be used to validate MEOWS. However, as obstetric complications (fortunately) are quite rare, such a study will likely have to span multiple centres and will therefore involve a considerable amount of time and resources to complete.

With respect to the implementation of MEOWS in clinical practice, this thesis has highlighted some of the benefits of electronic collection of vital sign data. The introduction of systems for electronic collection of vital sign data will also permit more flexibility in the design of early warning tools, for example with respect to the use of gestation-specific cut-off values, cut-off values set to account for potential patient-specific covariates such as BMI and parity or the use of non-integer variable weightings. Moreover, electronic implementation would eliminate some of the obstacles currently observed for effective implementation of MEOWS [196], including human error in recording of vital signs and in the computation of aggregate scores [197]. The design of an electronic system for collection and recording of vital signs in the obstetric setting could be based on the system developed as part of this thesis, which could be further adapted to incorporate the MEOWS cut-off values and scoring bands.

7.2.4 Final remarks

There is a long way to go before a nationally agreed, evidence-based and validated MEOWS system is ready for implementation in clinical practice, an aim towards which the work presented in this thesis can only be considered an initial step. Yet it seems intuitive that a better understanding of the physiological changes associated with normal pregnancy can enable earlier detection of severe pathology, lead to a better chance of improving the current state of maternal care and ultimately reduce the occurrence of serious maternal morbidity and mortality.

Appendix A

Supporting material for systematic review

A.1 Addition to methods

A.1.1 Search strategy

The search terms used to identify papers of interest in the MEDLINE, EMBASE, and CINAHL databases are listed in Table A.1. The search terms were specified in three different categories: period of pregnancy, measurement and reference value. To be included, a paper had to match at least one term from each category. The search terms were developed by Tatjana Petrinic (Bodleian Health Care Libraries, University of Oxford). A preliminary search was initially conducted, after which the search terms were adjusted before the final search. The following limits were applied to the search:

- human studies
- abstract available in English (but full text article can be in a language different to English)
- journal articles that describe original research only (exclude reviews, letters, comments, editorials, conference papers)

Table A.1: Search terms

Period of pregnancy	Measurement	Reference values
pregnan*	“vital sign*”	trend*
maternal	“early warning”	pattern*
obstetric*	EWS*	range*
“expectant mother*”	“modified early obstetric warning”	change*
“expecting mother*”	score*	measur*
peripartum	MEOWS	monitor*
“peri partum”	chart*	record*
antepartum	normogram*	assess*
“ante partum”	physiolog*	evaluat*
postpartum	haemodynam*	observ*
“post partum”	hemodynam*	guidance
intrapartum	“heart rate*”	guideline*
“intra partum”	“pulse rate*”	technique*
puerperium	“pulse oximetry*”	method*
trimester*	“oxygen saturation*”	systematic*
perinatal*	SpO2	chart*
antenatal*	Sp02	
postnatal*	oximetry	
	“blood pressure*”	
	(temperature AND body)	
	temperature*	
	“respirat* rate”	
	“breath* rate”	
	respirat*	
	breath*	
	“cardiac rate*”	

A.1.2 Definition of normal / healthy / low-risk women

Under the definition of *normal*, *healthy* or *low-risk*, women with the following conditions would be excluded:

- Known hypertension (defined as SBP > 140 mmHg and DBP > 90 mmHg), ever requiring treatment
- Known congenital cardiac disease
- Known ischaemic cardiac disease
- Known bronchiectasis
- Known fibrotic lung disease
- Known malignant disease

The following are *not* reasons for exclusion:

- Obesity
- Smoking
- Diabetes
- Renal disease without hypertension
- High BP at recruitment (defined as SBP > 140 mmHg and DBP > 90 mmHg, not previously diagnosed with hypertension)
- Pregnancy induced hypertension or pre-eclampsia diagnosed in a previous pregnancy (even if treated, as long as not developed essential hypertension)

A.1.3 Rules for extraction of measurements

The rules listed in Table A.2 define methods of measurement that were deemed to be representative of those methods used in standard clinical practice. Because some studies presented measurements in graphical form, it was also necessary to define a minimum acceptable resolution for accurate extraction of data from graphs. Data from graphs were extracted using an open-source software package (<http://digitizer.sourceforge.net/>).

Table A.2: Accepted methods of measurements.

Vital sign	Acceptable methods	Not acceptable methods	Minimum accuracy
Blood pressure	Non-invasive cuff Korotkoff sounds	Intraarterial Echocardiography Impedance cardiography	2 mmHg
Heart rate	Pulse oximetry Electrocardiography Blood pressure monitor	–	2 bpm
Oxygen saturation	Pulse oximetry	Blood gas analysis	1 %
Temperature	Rectal Oral Tympanic Auxillary Forehead	Intrauterine Other invasive methods	0.1 °C
Respiratory rate	Manual count (by viewing or by stethoscope)	Impedance pneumography	1 rpm

To avoid overrepresentation, rules for selection of a single data point were agreed. These rules were developed to ensure that data are relevant to normal clinical practice for intermittent observations undertaken by clinicians.

1. Where measurements are recorded in different positions, measurements made in the seated position (or position closest to it) will be recorded, according to the hierarchy of positions by the European Hypertension Society [187]. For blood pressure, only measurements made in the sitting or left lateral positions (or close to it) will be recorded.
2. When more than one baseline measurement is reported, the first reported measurement will be recorded.
3. Where measurements are recorded at different times of day, the measurements closest to midday will be recorded.
4. Where measurement methods are compared to a non-invasive gold standard the gold standard measurement will be used.
5. Where Korotkoff stage 4 and 5 diastolic blood pressures are reported, stage 5 will be recorded in line with the European Hypertension Society guidelines [187].
6. Where vital signs are reported both in the form of empirical and modeled values, empirical values will be used.
7. Where data are reported for several variations of subgroups, longitudinally measured data will be prioritised over single data points, even if this will involve excluding measurements from certain subgroups.

A.2 Addition to results

A.2.1 Reasons for rejection of studies

Table A.3: Reasons for rejection of studies.

Reason for rejection	Studies (n)
Wrong study type	15
Women not low-risk OR on medication	28
Not antenatal period OR no relevant baseline measurements	26
Inappropriate technique OR position	29
Data not recorded by healthcare professional	2
No relevant measurement data reported OR data not presented in a format suitable for extraction	49
Sample size < 50 women	51
Range of gestations too wide OR not adequately specified	14
Total number of studies rejected	215

A.2.2 Details of studies included

The details of the studies included in the systematic review are summarised in Table A.4 on the following pages.

Table A.4: Quality assessment criteria

Author	Year	Country ^a	Design		Method BP	Method HR	Position	Women (n)	Groups (n)	Occasions (n)	Quality score (%)
MacGillivray <i>et al.</i>	1969	UK	LT ^b	P ^b	manual ^c	—	sitting	226	1	7	31.6%
Okonofua <i>et al.</i>	1992	Nigeria	LT	P ^b	manual	—	sitting	189	1	6	36.8%
Peterson <i>et al.</i>	1992	USA	LT ^b	P ^b	nr	—	nr	668	2	6	10.5%
Halligan <i>et al.</i>	1993	Ireland	LT	P	manual	—	sitting	98	1	4	63.2%
Churchill <i>et al.</i>	1996	UK	LT	P	manual ^c	—	sitting	209	1	3	78.9%
vanOppen <i>et al.</i>	1996	NL	LT	P ^b	—	ECG	sitting	50	1	4	50.0%
Matkin <i>et al.</i>	1999	USA	LT ^b	P	nr	—	nr	2853	2	35	10.5%
Strevens <i>et al.</i>	2001	Sweden	LT	R ^b	manual	—	sitting	600	1	10	44.4%
Iwasaki <i>et al.</i>	2002	Japan	LT	R ^b	automated	—	sitting	1570	1	3	16.7%
Lo <i>et al.</i>	2002	NZ	LT	P ^b	manual ^c	—	sitting	101	1	4	52.6%
Strevens <i>et al.</i>	2002	Sweden	LT	R ^b	manual	—	sitting	498	3	9	33.3%
Gu <i>et al.</i>	2006	China	LT	R ^b	—	other	lateral	182	2	8	41.2%
Troisi <i>et al.</i>	2006	USA	LT	R ^b	nr	—	recumbent	86	1	3	5.6%
Lauszus <i>et al.</i>	2007	Denmark	LT ^b	P	manual	—	nr	59	1	2	31.6%
Miller <i>et al.</i>	2007	US	LT	P	manual	—	sitting	1733	1	3	31.6%
Hilmert <i>et al.</i>	2008	USA	LT	P	automated	—	sitting	164	1	3	47.4%
Bakker <i>et al.</i>	2010	NL	LT	P	automated	—	sitting	7106	3	3	57.9%
Jasovic-Siveska <i>et al.</i>	2011	Macedonia	LT ^b	P	manual	—	nr	400	3	4	15.8%
Timmermans <i>et al.</i>	2011	NL	LT ^b	P	automated	—	nr	3187	3	3	21.1%
Grindheim <i>et al.</i>	2012	Norway	LT	P	automated ^c	other	lateral	57	1	4	57.9%
Lee <i>et al.</i>	2012	USA	LT	P	nr	—	sitting	1684	1	2	21.1%
Ishikuro <i>et al.</i>	2013	Japan	LT	P ^b	automated	—	sitting	587	3	8	42.1%
Khan <i>et al.</i>	2013	India	LT ^b	P	manual	—	sitting	200	3	3	26.3%
Farias <i>et al.</i>	2014	Brazil	LT	P	automated	—	sitting	199	1	3	42.1%
Mahendru <i>et al.</i>	2014	UK	LT	P	automated ^c	automated	lateral	54	1	3	84.2%

continued on next page ...

... Table A.4 continued from previous page

Author	Year	Country ^a	Design		Method BP	Method HR	Position	Women (n)	Groups (n)	Occasions (n)	Quality score (%)
Macdonald-Wallis <i>et al.</i>	2015	UK	LT	R ^b	manual	—	sitting	10327	2	16	33.3%
Demidov <i>et al.</i>	1978	Russia	CS	P	nr	nr	laterallateral	50	1	1	22.2%
Tuffnell <i>et al.</i>	1990	UK	CS ^b	P ^b	nr	nr	sitting	70	2	1	5.6%
Ekholm <i>et al.</i>	1994	Finland	CS	P	manual	ECG	sitting (BP); supine (HR)	86	1	1	33.3%
Tape <i>et al.</i>	1994	USA	CS	P ^b	manual ^c	—	sitting	59	1	1	61.1%
Elvan-Taspinar <i>et al.</i>	2003	NL	CS	P	manual	—	recumbent	54	1	1	55.6%
Ochsenbein-Kolble <i>et al.</i>	2004	Switzerland	CS	P	automated	—	sitting	3234	1	37	66.7%
Ngan <i>et al.</i>	2007	China	CS	P	nr	—	lateral	58	1	1	38.9%
Turan <i>et al.</i>	2008	UK	CS	P ^b	—	echo	lateral	4689	2	1	76.5%
vanDerGraaf <i>et al.</i>	2013	Australia	CS	P	manual ^c	—	nr	225	2	1	83.3%
Carpenter <i>et al.</i>	2015	UK	CS	P ^b	—	ECG	supine	99	1	1	52.9%

Design: LT = longitudinal, CS = cross-sectional, P = prospective, R = retrospective

^a Country of study (if stated), otherwise country of study group

^b Study design (quality criteria 1 and/or 8) not explicitly stated or description unclear

^c BP measurement device suitable for use in pregnancy (quality criterion 13)

Appendix B

Mathematical methods

B.1 Classical kernel regression

In kernel regression, as described by Wand and Jones [171], a kernel $K(x)$ is centered on each data point $x = X_i$. Our kernel of choice is the normal distribution $\mathcal{N}(0, 1)$ (with zero mean and unit variance), as specified by Equation B.1. This kernel is unimodal, symmetric about zero and satisfies $\int K(x)dx = 1$, which is the criterion for a valid probability density function.

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \quad (\text{B.1})$$

Data points close to the centre of the kernel, at $x = X_i$, will have a larger contribution towards the regression curve because the amplitude of the kernel is at its maximum at this point. Similarly, data points further away from the kernel centre, where the amplitude of the kernel is smaller, will have a smaller contribution. The smoothness of the regression curve is controlled by incorporating a smoothing-parameter, which we call the bandwidth h , into the definition of the kernel by a change of variables as shown in Equation B.2.

$$K_h(u) = \frac{1}{h} K\left(\frac{u}{h}\right) \quad (\text{B.2})$$

Kernel regression can be thought of as a spatially weighted average of the observed values (X_i, Y_i) [166]. A local kernel estimator $\hat{m}(x; p, h)$ is fitted at each point x of the regression curve, using weighted least squares. The weights are determined by the kernel weights $K_h(X_i - x)$. The parameter p is the order of a local polynomial kernel estimator. In the literature, the local kernel estimator has been reported to take a variety of forms, such as linear, cubic and quintic polynomial functions. [165, 166] In general, the polynomial kernel estimator takes the form given by Equation B.3, where $\beta = [\beta_0, \dots, \beta_p]^T$ are the polynomial coefficients.

$$Y_i = \beta_0 + \beta_1(X_i - x) + \dots + \beta_p(X_i - x)^p \quad (\text{B.3})$$

The value of $\hat{m}(x; p, h)$ is the height (intercept) of the fit $\hat{\beta}_0 = Y_i(X_i = x)$ where $\hat{\beta}$ minimises Equation B.4.

$$\sum_{i=1}^n \{Y_i - \beta_0 - \dots + \beta_p(X_i - x)^p\}^2 K_h(X_i - x) \quad (\text{B.4})$$

Here, we only consider polynomial functions of order $p = 0$ or $p = 1$ for which there exist simple, closed-form solutions. In the case of $p = 0$, the least squares solution for $\hat{m}(x; p = 0, h)$ is the Nadaraya-Watson estimator given by Equation B.5.

$$\hat{m}(x; p = 0, h) = \frac{\sum_{i=1}^n K_h(X_i - x) Y_i}{\sum_{i=1}^n K_h(X_i - x)} \quad (\text{B.5})$$

In the case of $p = 1$, $\hat{m}(x; p = 1, h)$, the local linear estimator is given by Equations B.6 and B.7.

$$\hat{m}(x; p = 1, h) = n^{-1} \sum_{i=1}^n \frac{\{\hat{s}_2(x; h) - \hat{s}_1(x; h)(X_i - x)\} K_h(X_i - x) Y_i}{\hat{s}_2(x; h) \hat{s}_0(x; h) - \hat{s}_1(x; h)^2} \quad (\text{B.6})$$

$$\hat{s}_r(x; h) = n^{-1} \sum_{i=1}^n (X_i - x)^r K_h(X_i - x) \quad (\text{B.7})$$

When choosing the appropriate order p of the local polynomial kernel estimator, the

main factors to consider are (i) the ability of the estimator to follow large changes in gradient; and (ii) and the behaviour of the estimator near the boundaries. A higher order polynomial will increase the ability of the estimator to follow large changes in gradient, but also increase the computational complexity of the estimation. At the boundaries, the regression curve will tend to the form of the underlying polynomial.

B.2 Weighted variable bandwidth kernel regression

Classical kernel regression, as described in Section B.1, is suitable for the analysis of raw measurements, such as individual patient data. When the data points represent aggregate data (as opposed to individual patient data), each data point is associated with a *gestational age range* and *sample size*. These are accounted for by implementation of a *variable bandwidth* and data point *weightings*, respectively. These adaptations to the classical formulation of kernel regression were first proposed by Fleming [75].

The variable bandwidth is incorporated by replacing the constant kernel width h by a variable bandwidth h_i , which enables individual specification of the width of the kernel centered on each data point. The value of h_i is calculated using Equation B.8, where ΔX_i is the gestational age range associated with each data point and h_c is a common multiplier that controls the overall smoothness of the curve.

$$h_i = h_c \Delta X_i \quad (\text{B.8})$$

To reflect the differences in sample sizes, a weighting term w_i is added to the kernel function. For each data point, the weighting w_i is equal to the sample size associated with that point. The resulting kernel function is shown in Equation B.9.

$$K_h^i(u) = \frac{w_i}{h_i} K\left(\frac{u}{h_i}\right) \quad (\text{B.9})$$

B.3 General multilevel model

Multilevel models arise from the need to accurately model nested data structures. Repeated measurements, where individual patients have multiple observations, is an example of a nested data structure. In these cases, the standard assumption of independence is likely to be violated because of dependence among observations within the same cluster. Multilevel models extend conventional linear regression to handle such dependence.

Multilevel models are also referred to as random-effect models [198], hierarchical linear models [76], or multilevel linear mixed effects models [199], each with a specific notation that varies depending on the exact formulation of the model and the field of application. We choose to use the term *multilevel linear mixed effects model* or simply *multilevel model* to emphasise the hierarchical structure of the model, as well as the fact that the model is linear in the parameters.

To illustrate its hierarchical structure, the multilevel model is commonly presented as a two-stage linear model. Equations B.10 and B.11 describe the level-1 and level-2 models, respectively.

$$\mathbf{y}_i = \mathbf{Z}_i\beta_i + \mathbf{e}_i \quad (\text{B.10})$$

$$\beta_i = \mathbf{A}_i\beta + \mathbf{B}_i\mathbf{b}_i \quad (\text{B.11})$$

The level-1 model (Equation B.10) describes how, for an individual $i = 1, \dots, n$, the vector of observed responses $\mathbf{y}_i$ vary with the predictors included in the covariate matrix $\mathbf{Z}_i$. In our case of repeated measurements, the measured response is a physiological variable (e.g. blood pressure) and the covariate matrix is a parametrisation of a single predictor of time x_{ij} .¹ Both the total number of measurements $j = 1, \dots, n_i$ and the associated time points $x_{ij} = x_{i1}, \dots, x_{in_i}$ may vary between the individuals; this is a special case of the multilevel model which is often referred to as a growth curve model [198]. The residual $\mathbf{e}_i$ is the occasion-specific (level 1) residual, equivalent to measurement error or within-subject variation.

¹In general, $\mathbf{Z}_i$ can contain a wide range of predictors, not just time

The level-2 model (Equation B.11) separate the level-1 regression coefficients β_i into the *fixed* and *random* effect coefficients β and $\mathbf{b}_i$, respectively (hence the name *mixed-effect* models). The fixed effect coefficients β represent the underlying (unknown) population mean. The random effect coefficients $\mathbf{b}_i$ are also referred to as subject-specific (level 2) residuals or between-subject variation and represent individual variation that can not be explained by characteristics of the population. The matrices $\mathbf{A}_i$ and $\mathbf{B}_i$ are design matrices; $\mathbf{A}_i$ contains the between-subject (level 2) covariates and $\mathbf{B}_i$ is a matrix of zeros and ones that describes which components of β that have associated random effects $\mathbf{b}_i$. The design matrix $\mathbf{A}_i$ is discussed in more detail in Section B.4.1.

B.3.1 Variance components

We assume the level-1 residuals to be normally distributed, according to,

$$\mathbf{e}_i \sim \mathcal{N}(0, \mathbf{R}_i). \quad (\text{B.12})$$

For simplicity, it is common to assume that the level-1 residuals are identically distributed across individuals and uncorrelated between both individuals and measurement occasions, i.e. $\mathbf{R}_i = \sigma^2 \mathbf{I}_{n_i}$ where σ^2 is a constant within-subject variance and $\mathbf{I}_{n_i}$ is an identity matrix of size $(n_i \times n_i)$ so that $\text{Cov}(e_{ij}, e_{ij}) = 0$ for all i and j where $i \neq j$.

The random effects $\mathbf{b}_i$ (level-2 residuals) are treated as random variables drawn from an assumed multivariate normal distribution with zero mean and covariate matrix $\mathbf{D}$.

$$\mathbf{b}_i \sim \mathcal{N}(0, \mathbf{D}) \quad (\text{B.13})$$

We further assume that the level-2 residuals are uncorrelated across individuals and that the level-1 and level-2 residuals are uncorrelated across both individuals and occasions, i.e. that $\text{Cov}(\mathbf{b}_i, \mathbf{b}_j) = 0$ for all $i \neq j$ and $\text{Cov}(e_{ij}, \mathbf{b}_i) = 0$ for all i and j .

Longitudinal data, where the measurements are ordered by time, are likely to display heteroscedasticity and autocorrelation. To account for both of these, it is possible to use a more complex specification of the level-1 covariance matrix.

B.3.2 The conditional and marginal models

Inference is made based on the composite model (Equation B.14), which is obtained by combining the level-1 and level-2 models and so that $\mathbf{X}_i = \mathbf{Z}_i \mathbf{A}_i$ and $\mathbf{Z}_i^* = \mathbf{Z}_i \mathbf{B}_i$. In Equation B.14, p and k are the number of fixed and random effect coefficients, respectively.

$$\underset{(n_i \times 1)}{\mathbf{y}_i} = \underset{(n_i \times p)(p \times 1)}{\mathbf{X}_i \beta} + \underset{(n_i \times k)(k \times 1)}{\mathbf{Z}_i \mathbf{b}_i} + \underset{(n_i \times 1)}{\mathbf{e}_i} \quad (\text{B.14})$$

Depending on whether population or individual-specific trends are of interest, inference is made using either the marginal or conditional distribution of $\mathbf{y}_i$. The conditional distribution of $\mathbf{y}_i | \mathbf{b}_i$ is given by Equation B.15.

$$\mathbf{y}_i | \mathbf{b}_i \sim \mathcal{N}(\mathbf{X}_i \beta + \mathbf{Z}_i \mathbf{b}_i, \mathbf{R}_i) \quad (\text{B.15})$$

The marginal distribution is obtained by integrating over the random effects $\mathbf{b}_i$ and is described by Equation B.16.

$$\mathbf{y}_i \sim \mathcal{N}(\mathbf{X}_i \beta, \mathbf{V}_i) \quad (\text{B.16})$$

The covariance matrix of the marginal distribution is $\mathbf{V}_i = \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T + \mathbf{R}_i$. From this we make two interesting observations. Firstly, we note that $\mathbf{V}_i$ is a function of the level-1 predictor (which in this case is time). This enables modelling of heteroscedasticity in the variance of the observed population response $\mathbf{y}_i$. Secondly, the correlations between observations are induced through the between-subject variation term $\mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T$, even when the within-subject errors are assumed to be independent (i.e. $\mathbf{R}_i = \sigma^2 I_{n_i}$). In other words, matrix $\mathbf{D}$ describes the unexplained individual-specific time-invariant effects which arise from a correlation between multiple measurements from the same individual.

In the case of a random slope model, where $\mathbf{b}_i = [b_{i0}, b_{i1}]$, the correlation between the random intercept and slope is given by Pearson's correlation coefficient $\rho = D_{12} / \sqrt{D_{11} D_{22}}$ where D_{11} , D_{12} and D_{22} are elements of $\mathbf{D}$.

B.3.3 Specification of covariance matrices for longitudinal data

Up to this point, we have assumed that an individual's pattern of responses over time is adequately captured by the individual-specific trajectory $\mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i$ so that the level-1 residuals $\mathbf{e}_i = [e_{i1}, \dots, e_{in_i}]$ can be assumed independent. For longitudinal data, where repeated measures from an individual are ordered by time, it is necessary to test for heteroscedasticity and autocorrelation in the level-1 residuals. This is achieved by a change in the specification of the covariance matrix $\mathbf{R}_i$. Equation B.17 shows how $\mathbf{R}_i$ can be decomposed into a variance structure component $\mathbf{W}_i$ and a correlation structure $\mathbf{C}_i$.

$$\mathbf{R}_i = \mathbf{W}_i\mathbf{C}_i\mathbf{W}_i \quad (\text{B.17})$$

To model level-1 heteroscedasticity (i.e. increasing or decreasing variance with time), we parameterise the $\mathbf{W}_i$ component of the matrix. A multitude of formulations have been proposed, and the reader is referred to Pinheiro and Bates [199] for an overview.

The covariance structure $\mathbf{C}_i$ can be defined to model autocorrelation. For datasets where time can take any value, correlation can be defined in an exponential form (AR(exp), Equation B.18) that is a function of the time x_{ij} of each observation. In this case, θ is the correlation coefficient that is estimated from the data. Specification of a different covariance structure has minimal effect on the computation of fixed effects but may affect the total model variance [167].

$$\mathbf{C}_i = \begin{bmatrix} 1 & & & & \\ \exp(-\theta|x_{2i} - x_{1i}|) & 1 & & & \\ \exp(-\theta|x_{3i} - x_{1i}|) & \exp(-\theta|x_{3i} - x_{2i}|) & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ \exp(-\theta|x_{n_i} - x_{1j}|) & \exp(-\theta|x_{n_i} - x_{2i}|) & \exp(-\theta|x_{n_i} - x_{3i}|) & \cdots & 1 \end{bmatrix} \quad (\text{B.18})$$

B.3.4 Parameter estimation for the multilevel model

Several methods have been proposed to make inference about the values of the model parameters β , σ^2 and $\mathbf{D}$ and the random effects $\mathbf{b}_i$. In this thesis, we use maximum likelihood (ML) estimation, as proposed by Harville [200], to solve for the parameters. The description below has been adapted from Wu and Zhang [201].

The multilevel model is based on strong assumptions of normality, as described by Equation B.15. From this, we establish the generalised likelihood function in Equation B.19.

$$p(\beta, \mathbf{b}, \mathbf{D}, \mathbf{R}|\mathbf{y}) \propto p(\mathbf{y}|\beta, \mathbf{b}, \mathbf{D}, \mathbf{R})p(\beta)p(\mathbf{b})p(\mathbf{D})p(\mathbf{R}) \quad (\text{B.19})$$

If we consider that only $\mathbf{b}$ is a random variable this reduces to Equation B.20.

$$p(\beta, \mathbf{b}, \mathbf{D}, \mathbf{R}|\mathbf{y}) \propto p(\mathbf{y}|\beta, \mathbf{b}, \mathbf{D}, \mathbf{R})p(\mathbf{b}). \quad (\text{B.20})$$

The likelihood function of the parameters given $\mathbf{y}$, is thus given by Equation B.21.

$$\begin{aligned} \mathcal{L}(\beta, \mathbf{b}, \mathbf{D}, \mathbf{R}|\mathbf{y}) &= (2\pi)^{-N/2}|\mathbf{R}|^{-1/2} \exp \left\{ -\frac{1}{2}[\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{b}]^T \mathbf{R}^{-1}[\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{b}] \right\} \\ &\times (2\pi)^{-Jq/2}|\mathbf{D}|^{-1/2} \exp \left\{ -\frac{1}{2}\mathbf{b}^T \mathbf{D}^{-1}\mathbf{b} \right\} \end{aligned} \quad (\text{B.21})$$

Here $N = \sum_{i=1}^n n_i$ and nk is the dimension of $\mathbf{b}$. Since $\mathbf{b}$ is a random variable and not a model parameter, we integrate over $\mathbf{b}$ to obtain the conventional likelihood function shown in Equation B.22.

$$\begin{aligned} \mathcal{L}(\beta, \mathbf{D}, \mathbf{R}|\mathbf{y}) &= \int L(\beta, \mathbf{b}, \mathbf{D}, \mathbf{R}|\mathbf{y})d\mathbf{b} \\ &= (2\pi)^{-N/2}|\mathbf{V}|^{-1/2} \exp \left\{ -\frac{1}{2}[\mathbf{y} - \mathbf{X}\beta]^T \mathbf{V}^{-1}[\mathbf{y} - \mathbf{X}\beta] \right\} \end{aligned} \quad (\text{B.22})$$

If we take the log of both sides, this becomes the log likelihood (dropping the constant) expressed in Equation B.23.

$$\log \mathcal{L}(\beta, \mathbf{D}, \mathbf{R}|\mathbf{y}) = -\frac{1}{2}N \log(2\pi) - \log |\mathbf{V}| - \frac{1}{2}[\mathbf{y} - \mathbf{X}\beta]^T \mathbf{V}^{-1}[\mathbf{y} - \mathbf{X}\beta]. \quad (\text{B.23})$$

Harville (1977) showed that minimising the log likelihood function yields the generalised least squares (GLS) solutions for β and $\mathbf{b}_j$. The GLS solutions are

$$\hat{\beta} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y}, \quad (\text{B.24})$$

$$\hat{\mathbf{b}}_i = \mathbf{D} \mathbf{Z}_i^T \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i \hat{\beta}). \quad (\text{B.25})$$

We note that if $\mathbf{V}_i = \sigma^2 \mathbf{I}_n$, i.e. if for all individuals j the residuals can be assumed to be homogenous (homoscedastic) and independent, Equation B.24 equals the ordinary least squares (OLS) estimate of β . Since $\mathbf{V}$ is a function of $\mathbf{Z}$, this would also apply if the number ($n_i = n$) and values of the level 1 predictors were the same across individuals, i.e. the dataset is balanced. If on the other hand, the dataset is unbalanced and the residuals are heteroscedastic and/or correlated, the GLS solution must be used. The GLS solution accounts for differences in variance between individuals by weighting the observations $\mathbf{y}$ by the inverse of the variance matrix $\mathbf{V}_i = \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T + \mathbf{R}_i$. Effectively, this implies that individuals with lower variance contribute more weight to the estimation of $\hat{\beta}$.

B.3.4.1 Parameters estimates for unknown variance components

Since the variance components cannot be directly observed, they must be estimated using an iterative method. Several algorithms have been proposed, including expectation maximisation (EM) and the Newton-Raphson gradient algorithms. Below, the EM method for ML estimation is presented, based on the description from [201].

For this estimation, we assume that $\mathbf{R}_i$ takes the form given by Equation B.26.

$$\mathbf{R}_i = \sigma^2 \mathbf{I}_{n_i} \quad (\text{B.26})$$

Under the assumptions of normality and when $\mathbf{b}_i$ and $\mathbf{e}_i$ are known, the ML estimators of σ^2 and $\mathbf{D}$ would be given by Equation B.27

$$\hat{\sigma}^2 = N^{-1} \sum_{i=1}^n \mathbf{e}_i^T \mathbf{e}_i, \quad \hat{\mathbf{D}} = n^{-1} \sum_{i=1}^n \mathbf{b}_i^T \mathbf{b}_i \quad (\text{B.27})$$

However, as $\mathbf{b}_i$ and $\mathbf{e}_i$ are not known we replaced their conditional expectations given the observations $\mathbf{y}$ and $\beta = \hat{\beta}$, which equals the ML estimates $\hat{\sigma}^2$ and $\hat{\mathbf{D}}$,²

$$E(\hat{\sigma}^2 | \mathbf{y}, \beta = \hat{\beta}) = N^{-1} \sum_{i=1}^n \{ \hat{\mathbf{e}}_i^T \hat{\mathbf{e}}_i + \sigma^2 [n_i - \sigma^2 \times \text{tr}(\mathbf{V}_i^{-1})] \} \quad (\text{B.28})$$

$$E(\hat{\mathbf{D}} | \mathbf{y}, \beta = \hat{\beta}) = n^{-1} \sum_{i=1}^n \{ \hat{\mathbf{b}}_i^T \hat{\mathbf{b}}_i + [\mathbf{D} - \mathbf{D} \mathbf{Z}_i^T \mathbf{V}_i^{-1} \mathbf{Z}_i \mathbf{D}] \}. \quad (\text{B.29})$$

where

$$\hat{\mathbf{e}}_i = \mathbf{y}_i - \mathbf{X}_i \hat{\beta} - \mathbf{Z}_i \hat{\mathbf{b}}_i \quad (\text{B.30})$$

and $\hat{\mathbf{b}}_i$ is given by Equation B.25.

B.3.4.2 Steps of the EM algorithm

Based on this, the EM algorithm proceeds as follows (iterations $r = 0, 1, 2, \dots$):

Step 0 (set $r = 0$): Initialise the variance component parameters by setting $\sigma_{(r)}^2 = 1$ and $\hat{\mathbf{D}} = I_{k \times k}$ where k is the number of random effects.

Step 1 (update $r = r+1$): Update $\hat{\beta}$ and $\hat{\mathbf{b}}_j$ using Equation B.24 and B.25, respectively:

$$\begin{aligned} \hat{\beta}_{(r)} &= \left(\sum_{i=1}^n \mathbf{X}_i^T \hat{\mathbf{V}}_{i(r-1)}^{-1} \mathbf{X}_i \right)^{-1} \left(\sum_{i=1}^n \mathbf{X}_i^T \hat{\mathbf{V}}_{i(r-1)}^{-1} \mathbf{y}_i \right) \\ \hat{\mathbf{b}}_{i(r)} &= \hat{\mathbf{D}}_{(r-1)} \mathbf{Z}_i^T \hat{\mathbf{V}}_{i(r-1)}^{-1} \left(\mathbf{y}_i - \mathbf{X}_i \hat{\beta}_{(r)} \right) \end{aligned}$$

$$\text{where } \hat{\mathbf{V}}_{i(r-1)}^{-1} = \mathbf{Z}_i \hat{\mathbf{D}}_{(r-1)} \mathbf{Z}_i^T + \hat{\sigma}_{(r-1)}^2 \mathbf{I}_{n_i}.$$

²For the full derivations of these, please refer to [201]

Step 2: Update $\hat{\sigma}_{(r)}^2$ and $\hat{\mathbf{D}}_{(r)}$ using Equation B.28 and B.29, respectively:

$$\begin{aligned}\hat{\sigma}_{(r)}^2 &= N^{-1} \sum_{i=1}^n \left\{ \hat{\mathbf{e}}_{i(r)}^T \hat{\mathbf{e}}_{i(r)} + \hat{\sigma}_{(r-1)}^2 [n_i - \hat{\sigma}_{(r-1)}^2 \text{trace}(\hat{\mathbf{V}}_{i(r-1)}^{-1})] \right\} \\ \hat{\mathbf{D}}_{(r)} &= n^{-1} \sum_{i=1}^n \left\{ \hat{\mathbf{b}}_{i(r)} \hat{\mathbf{b}}_{i(r)}^T + [\hat{\mathbf{D}}_{(r-1)} - \hat{\mathbf{D}}_{(r-1)} \mathbf{Z}_i^T \hat{\mathbf{V}}_{i(r-1)}^{-1} \mathbf{Z}_i \hat{\mathbf{D}}_{(r-1)}] \right\}\end{aligned}$$

where $N = \sum_{i=1}^n n_i$, i.e. the total number of measurements across all individuals, and

$$\hat{\mathbf{e}}_{i(r)} = \mathbf{y}_i - \mathbf{X}_i \hat{\beta}_{(r)} - \mathbf{Z}_i \hat{\mathbf{b}}_{i(r)}. \quad (\text{B.31})$$

Step 3: Compute the incremental increase in likelihood (from Equation B.23). Repeat Steps 1 and 2 until convergence.

The ML method has some important limitations. Firstly, ML estimators of the variance $\mathbf{V}$ do not account for losses of degrees of freedom resulting from estimation of the model's fixed effects β . Secondly, the ML estimators are derived under strong assumptions of normality, both of the observations $\mathbf{y}$ and of the residuals $\mathbf{b}_i$ and $\mathbf{e}_i$.

B.3.5 Empirical Bayes prediction of random effects and shrinkage

The Empirical Bayes predictions of $\mathbf{b}_i$ are obtained as the mean of the posterior distribution of $\mathbf{b}_i$, given the observations $\mathbf{y}_i = y_{i1}, \dots, y_{in_i}$ ³. This is obtained from Bayes theorem in Equation B.32.

$$\text{Posterior}(\mathbf{b}_i | y_{i1}, \dots, y_{in_i}) \propto \text{Prior}(\mathbf{b}_i) \times \text{Likelihood}(y_{i1}, \dots, y_{in_i} | \mathbf{b}_i) \quad (\text{B.32})$$

The $\text{Prior}(\mathbf{b}_i)$ is the normal distribution of the random effects, $\mathbf{b}_i \sim \mathcal{N}(0, \mathbf{D})$. The $\text{Likelihood}(y_{i1}, \dots, y_{in_i} | \mathbf{b}_i)$ is the ML estimate of the total residual $\mathbf{X}_i \hat{\beta}$ of the observations $\mathbf{y}_i$. This gave us the expression of $\mathbf{b}_i$ from Equation B.25, which was given as

³Recall that n_i is the number of observations for subject i

$$\mathbf{b}_i = \widehat{\mathbf{D}}\mathbf{Z}_i^T \mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}})$$

From this expression, we can show that the subject-specific mean $\mathbf{X}_i\alpha_i = \mathbf{X}_i\widehat{\boldsymbol{\beta}} + \mathbf{Z}_i\tilde{\mathbf{b}}_i$ is of the form of a weighted average of the population mean $\mathbf{X}_i\widehat{\boldsymbol{\beta}}$ and the observed data $\mathbf{y}_i$,

$$\begin{aligned} \mathbf{X}_i\alpha_i &= \mathbf{X}_i\widehat{\boldsymbol{\beta}} + \mathbf{Z}_i\tilde{\mathbf{b}}_i \\ &= \mathbf{X}_i\widehat{\boldsymbol{\beta}} + \mathbf{Z}_i\widehat{\mathbf{D}}\mathbf{Z}_i^T \mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}) \\ &= (\mathbf{I}_{n_i} - \mathbf{Z}_i\widehat{\mathbf{D}}\mathbf{Z}_i^T \mathbf{V}_i^{-1})\mathbf{X}_i\widehat{\boldsymbol{\beta}} + \mathbf{Z}_i\widehat{\mathbf{D}}\mathbf{Z}_i^T \mathbf{V}_i^{-1}\mathbf{y}_i \\ &= (\mathbf{I}_{n_i} - \Omega_i)\mathbf{X}_i\widehat{\boldsymbol{\beta}} + \Omega_i\mathbf{y}_i \end{aligned} \tag{B.33}$$

The matrix $\mathbf{V}_i = \widehat{\mathbf{R}}_i + \mathbf{Z}_i\widehat{\mathbf{D}}\mathbf{Z}_i^T$ is the total variance of $\mathbf{y}_i$. The shrinkage factor is in this case given by $\Omega_i = \mathbf{Z}_i\widehat{\mathbf{D}}\mathbf{Z}_i^T \mathbf{V}_i^{-1}$.

B.3.5.1 Example of the effect of shrinkage

For the basic random intercept model given by $y_{ij} = \beta + b_i + e_{ij}$, Equation B.33 reduces to the expression in Equation B.34, where $\bar{y}_i$ is the ML estimate of the subject mean⁴.

$$\begin{aligned} \tilde{\alpha}_i &= (1 - \omega_i)\hat{\beta} + \omega_i\bar{y}_i \\ \text{where } \omega_i &= \frac{D}{D + \frac{\sigma^2}{n_i}} \end{aligned} \tag{B.34}$$

From this, we see that the predicted subject mean $\tilde{\alpha}_i$ is a compromise between the estimated population mean $\hat{\beta}$ (if $\omega_i \rightarrow 0$) and the ML estimate of the subject-specific mean $\bar{y}_i$ (if $\omega_i \rightarrow 1$).

To illustrate this by means of an example, we consider a subject with $n_i = 3$ observations $\mathbf{y}_i = \{3, 4, 5\}$. The ML estimate of the total observations is therefore $\bar{y} = 4$. The

⁴The derivation of the expression for a general multilevel model was included earlier in Appendix B.3.5

population mean has already been estimated to be $\beta = 0$. Depending on the values of the within-subject variance σ^2 and the between-subject variance D , the EB estimate of b_i will vary. We consider two scenarios:

1. $D = 1$ and $\sigma^2 = 4$. The shrinkage factor becomes $\omega_i = 0.43$.
2. $D = 4$ and $\sigma^2 = 1$. The shrinkage factor becomes $\omega_i = 0.92$.

The two scenarios are illustrated in Figure B.1. This example was adapted from [172].

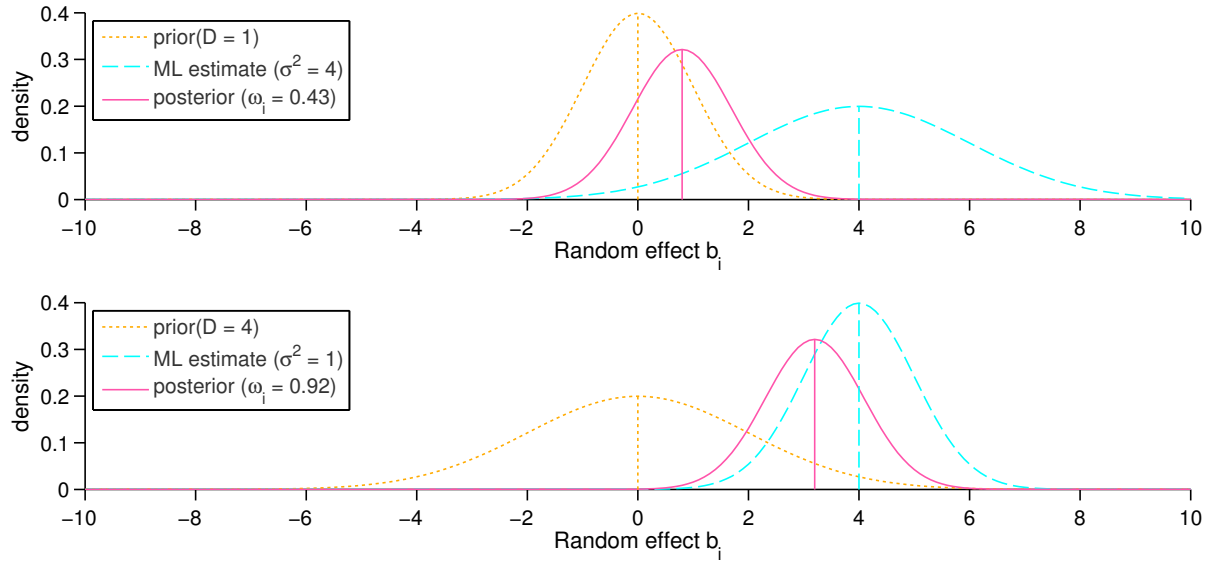


Figure B.1: Effect of shrinkage on the posterior random effect estimates.

B.3.6 Selection of model parameters

Akaike's Information Criterion (AIC) [202] and the Bayesian Information Criterion (BIC) [203] are useful for the comparison of the fit of the mean trajectory for non-nested⁵ multilevel models [199]. The AIC and BIC criteria are given by Equations B.35 and B.36, respectively.

$$AIC = -2\log\mathcal{L} + 2n_{par} \quad (\text{B.35})$$

$$BIC = -2\log\mathcal{L} + n_{par} \log(N) \quad (\text{B.36})$$

⁵Two statistical models are nested if the components of the first model represent a complete subset of the components of the second model.

Here, $\log \mathcal{L}$ is the log likelihood, n_{par} denotes the number of parameters in the model and N denotes the total number of observations used to fit the model. Under these definitions, the appropriate model is selected based on smaller values of both AIC and BIC. Because of the additional term $\log(N)$, BIC tends to favour more parsimonious models.

B.4 Multilevel model for aggregate data

The general multilevel model for analysis of raw measurements (frequently referred to as *individual patient data* or *IPD*) was described in Section B.3. In the general multilevel model the parameters are entirely estimated from the data. In the multilevel model for aggregate data (MLAD), we fix the diagonal elements of the within-study (level 1) covariance matrix to the variance of the mean (SE^2) of the observed estimates at the time points reported in each study. The occasion-specific variance of the mean SE_{tij}^2 is given by Equation B.37, where $\sigma_{tij}^2 = \sigma_{ti1}^2, \dots, \sigma_{tin_i}^2$ are the study reported variances and w_{ij} is the associated sample size.

$$SE^2 = \frac{\sigma_{tij}^2}{w_{ij}} \quad (\text{B.37})$$

Equation B.38 shows the adapted formulation of the covariance matrix $\mathbf{R}_i$.

$$\mathbf{R}_i = \begin{bmatrix} SE_{ti1}^2 & & & & \\ 0 & SE_{ti2}^2 & & & \\ 0 & 0 & SE_{ti3}^2 & & \\ \vdots & \vdots & \vdots & \ddots & \\ 0 & 0 & 0 & \dots & SE_{tin_i}^2 \end{bmatrix} \quad (\text{B.38})$$

This formulation of the multilevel model is similar to the *random time-effects model* proposed by Ishak *et al.* [74]. However, in order to account for differences in sample size we use SE^2 instead of variance σ^2 . This has the effect of weighting each observation by

the reciprocal of the associated SE^2 , so that more precise estimates with a larger sample size are more influential. Ideally, if the covariance between estimates were to be reported we could also fix the off-diagonal elements. Where covariances are not reported, these can be estimated by assuming a correlation structure, as discussed in Section B.3.3. However, as this introduces additional parameters the ability to model covariances may be limited by the availability of data.

B.4.1 Inclusion of level 2 covariates

As for the general multilevel model, discussed in Section B.3, correlations may be modelled by including time-dependent random effects (e.g. random intercepts and slopes). By allowing for random variation between studies we relax the assumption that between-study heterogeneity affects observations at each measurement occasion in a given study the same way. Such heterogeneity may arise for example due to different rates of loss to follow up.

Study-specific (level-2) covariates (e.g. method of measurement and study year) may be included in the specification of the design matrix $\mathbf{A}_i$. Study-level covariates were used in the assessment of subgroups in Section 2.2.3 using the specification of matrix $\mathbf{A}_i$ shown in Equation B.39. Here, m_{1i} and m_{2i} denotes method of measurement and study year, respectively.

$$\mathbf{A}_i = \begin{bmatrix} 1 & 0 & 0 & m_{1i} & m_{2i} \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} \quad (\text{B.39})$$

B.5 Gaussian processes

A Gaussian Process is defined as a stochastic process for which a finite number of observations y_i at time points x_i in the dataset have a joint Gaussian distribution. The dataset of N pairs of input-output values is expressed as $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$. Here, we provide a brief summary of Gaussian Processes. For a more thorough introduction the reader is referred to Rasmussen and Williams [159].

From Rasmussen and Williams [159], a Gaussian Process is defined entirely by its mean function $m(\mathbf{x})$ and covariance function $K(x_i, x_j)$ according to Equation B.40.

$$f(\mathbf{x}) \sim \mathcal{GP}\left(m(\mathbf{x}), \mathbf{K}(x_i, x_j)\right) \quad (\text{B.40})$$

Here, we consider the covariance function to be the squared-exponential covariance function given by Equation B.41.

$$\text{Cov}(y_i, y_j) = K(x_i, x_j) = \sigma_0^2 \exp\left(-\frac{1}{2\lambda}|x_i - x_j|^2\right) + \sigma_n^2 \quad (\text{B.41})$$

Here, $\theta = \{\sigma_0^2, \lambda, \sigma_n^2\}$ are hyperparameters. We wish to estimate the values of these hyperparameters.

B.5.1 Hyperparameter estimation for Gaussian processes

The predictive distribution for the Gaussian process is given by Equation B.42.

$$\mathbf{f}_* | X, \mathbf{y}, X_* \sim \mathcal{N}(\bar{\mathbf{f}}_*, \text{Cov}(\mathbf{f}_*)) \quad (\text{B.42})$$

$$\bar{\mathbf{f}}_* = \mathbf{K}_*^T (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y}$$

$$\text{Cov}(\mathbf{f}_*) = \mathbf{K}_{**} - \mathbf{K}_*^T (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{K}_*$$

Here, $\mathbf{K}_* = K(\mathbf{x}, \mathbf{x}_*)$ is the covariance between the training inputs $\mathbf{x}$ and the test inputs $\mathbf{x}_*$ and $\mathbf{K}_{**} = K(x_*, x_*)$ is the prior variance of x_* .

To optimise over the hyperparameters, we introduce the marginal likelihood $p(\mathbf{y}|X)$ as the integral of the likelihood $p(\mathbf{y}|\mathbf{f}, X)$ times the prior $p(\mathbf{f}|X)$ shown in Equation B.43.

$$p(\mathbf{y}|X) = \int p(\mathbf{y}|\mathbf{f}, X)p(\mathbf{f}|X)d\mathbf{f} \quad (\text{B.43})$$

Both the likelihood and prior are Gaussian, expressed as $\mathbf{y}|\mathbf{f} \sim \mathcal{N}(\mathbf{f}, \sigma_n^2 \mathbf{I})$ and $\mathbf{f}|X \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$, respectively. Integration of these yields the negative log marginal likelihood in Equation B.44.

$$\begin{aligned} \mathcal{L} &= -\log p(\mathbf{y}|\mathbf{x}) \\ &= \frac{1}{2} \log |\mathbf{K} + \sigma_n^2 \mathbf{I}| + \frac{1}{2} \mathbf{y}^T (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y} + \frac{N}{2} \log(2\pi) \end{aligned} \quad (\text{B.44})$$

This is equivalent to observing that $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{K} + \sigma_n^2 \mathbf{I})$. The most likely values of the hyperparameters are then obtained by setting the derivative of the log likelihood with respect to θ to zero and solving for each of the hyperparameters using gradient descent. In MATLAB, the estimation of hyperparameters for Gaussian processes is implemented as part of the `gpm1` package [204].

Appendix C

Supporting material: Vital sign measurement

C.1 Measurement of blood pressure

C.1.1 Oscillometric principle

Automated monitors estimate systolic (SBP) and diastolic (DBP) blood pressure in three main steps as summarised below using the oscillometric principle. The steps are illustrated in Figure C.1.

1. The cuff is inflated (either automatically or manually) to a high pressure. The pressure is subsequently released at a controlled *deflation rate*. The cuff pressure is recorded during the deflation process.
2. The recorded cuff pressure is detrended to obtain the oscillometric signal. The pressure in the cuff at the point where the oscillations reach maximum amplitude is taken to be the mean arterial blood pressure (MAP).
3. SBP and DBP are estimated as fractions of the envelope curve of the oscillometric signal with respect to the maximum value. Specifically, SBP is defined at a point before (to the left of) the MAP and DBP is defined at a point after (to the right of) the MAP.

Monitors from different manufacturers typically use different percentage values or *algorithms*.

The exact algorithm is typically determined empirically by optimisation relative to ausc-

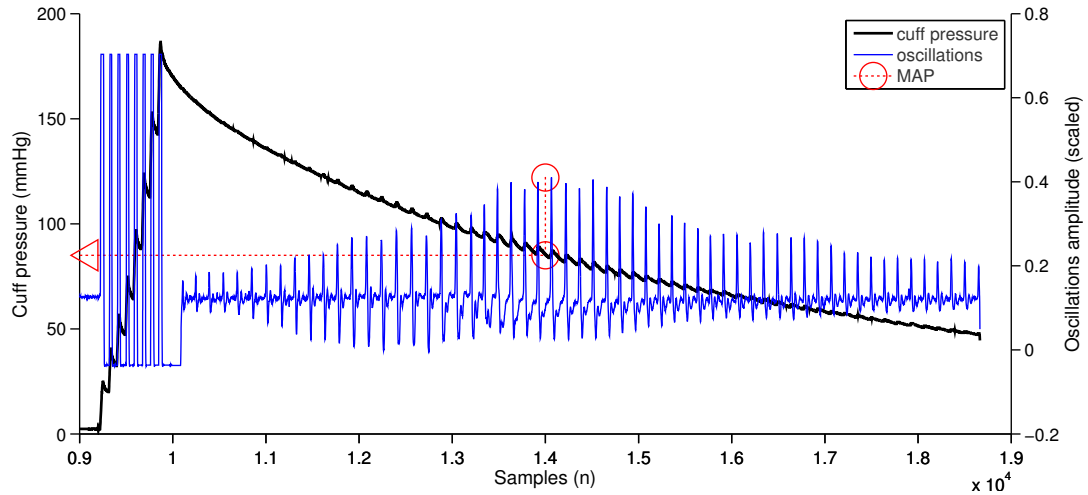


Figure C.1: Cuff pressure signal obtained from a manual BP device (bold black line, left axis). The oscillations (blue line, right axis) are obtained by detrending the pressure signal. The mean arterial BP is estimated at the point where the oscillations reach maximum amplitude (red circle and lines). This example BP signal was plotted from data kindly provided by the authors of [205].

ultatory BP measurements [187]. It likely to differ for certain populations, such as pregnant women (hence the need for a pregnancy approved device). Details of the algorithm and deflation rate are generally treated as proprietary information.

C.2 Measurement of heart rate and oxygen saturation

C.2.1 Signal quality index SQI(flags)

The pulse oximeter used in this study (Nonin WristOx₂ Model 3150, Nonin Medical Inc) output a series of pulse quality indices in addition to the PPG waveform and values of HR and SpO₂. An overview of the outputs from the pulse oximeter is presented in Table C.1. Table C.1 summarises the different flags output by the pulse oximeter to indicate artefactual data. As these flags are output for transient artefacts, sections that are deemed as 'valid' may still contain artefacts. These artefacts are typically flagged once per pulse. A simple SQI was therefore devised according to the following rules:

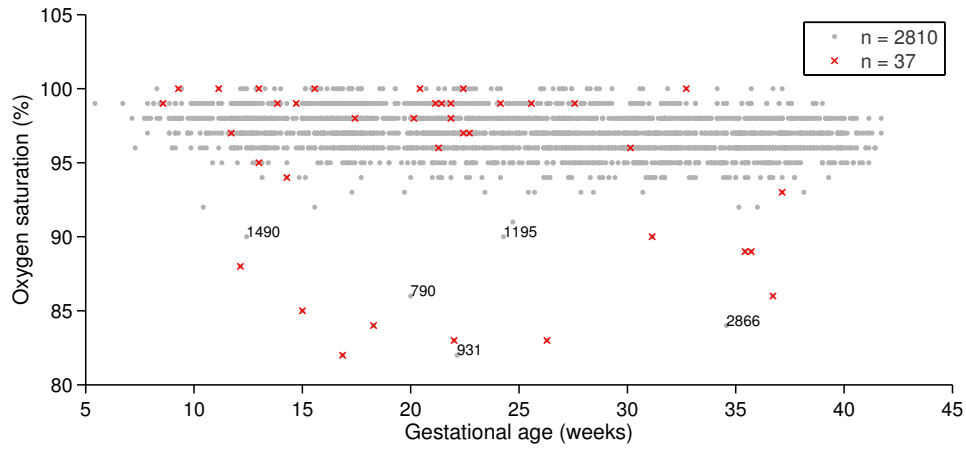
Table C.1: Characteristics of the pulse oximeter outputs

Output values	Update frequency	Description
PPG waveform	75 Hz	Infrared signal
Heart rate	3 Hz	Mean over last 4 beats
SpO ₂	3 Hz	Mean over last 4 beats
Artefact flags:		
Artefact (short-term)	75 Hz	Single pulse artefact
Artefact (out-of-track)	75 Hz	Consecutive pulse artefacts
Sensor alarm	75 Hz	No finger detected
Pulse amplitude quality	75 Hz	Single pulse artefact

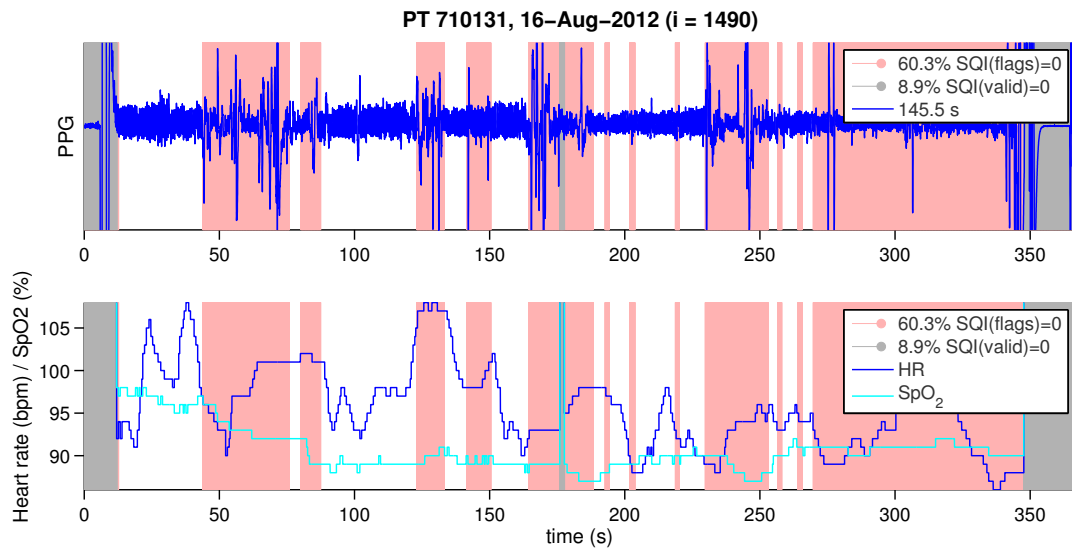
1. As artefacts were typically only flagged once per pulse, a time window (T_{pulse}) was defined immediately before and after every artefact to label the whole pulse as artefactual. This was set to $T_{pulse} = 0.67$ seconds, corresponding to the maximum pulse expected pulse length (corresponding to a minimum HR of 40 bpm).
2. If multiple artefacts were flagged within a short time window (T_{cons}), the whole time window was labelled as artefactual. This was set empirically to $T_{cons} = 5$ seconds.

C.2.2 Recordings of low SpO₂ not rejected

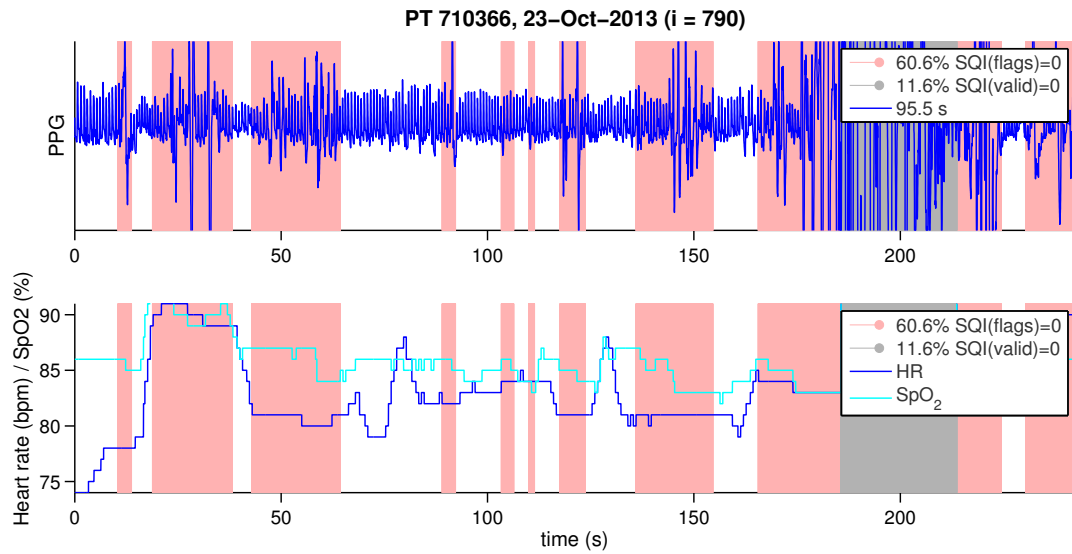
The PPG records that were not excluded by this threshold were visually examined. We concluded that, although they did contain a substantial amount of artefactual data, the low median SpO₂ values could also represent a true physiological state. Figure C.2a highlights the SpO₂ recordings that were deemed to be of sufficiently good quality to be included in the analysis. Each of these recordings are numbered and Figures C.2 and C.3 on the following pages show the associated recorded signals and output SpO₂ time series.



(a) Recordings with low $\text{SpO}_2 \leq 90\%$ that exceeded the required $th_p \geq 20\%$ of data with $\text{SQI}(\text{flags})=1$.

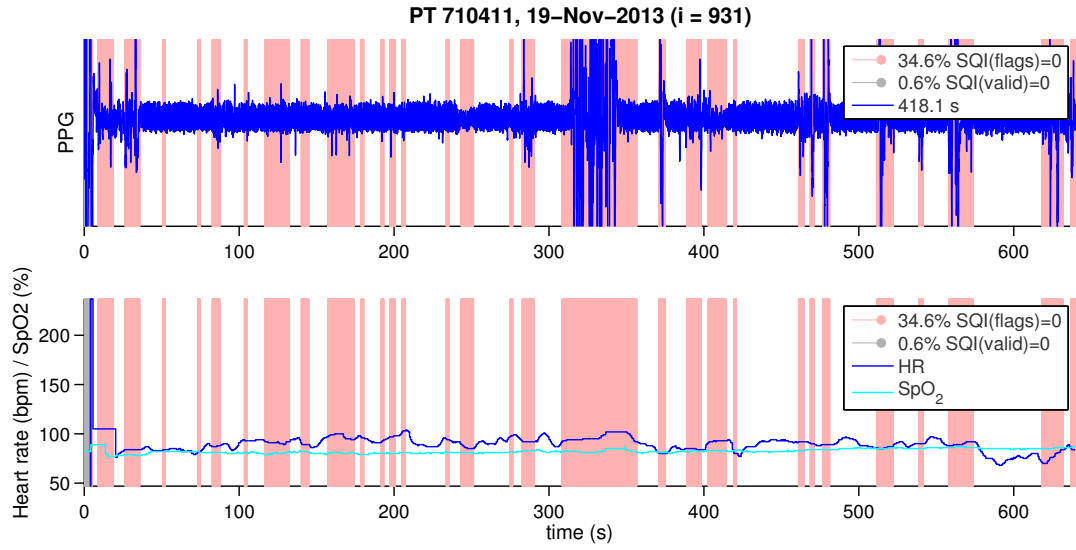


(b) Recording i=1490

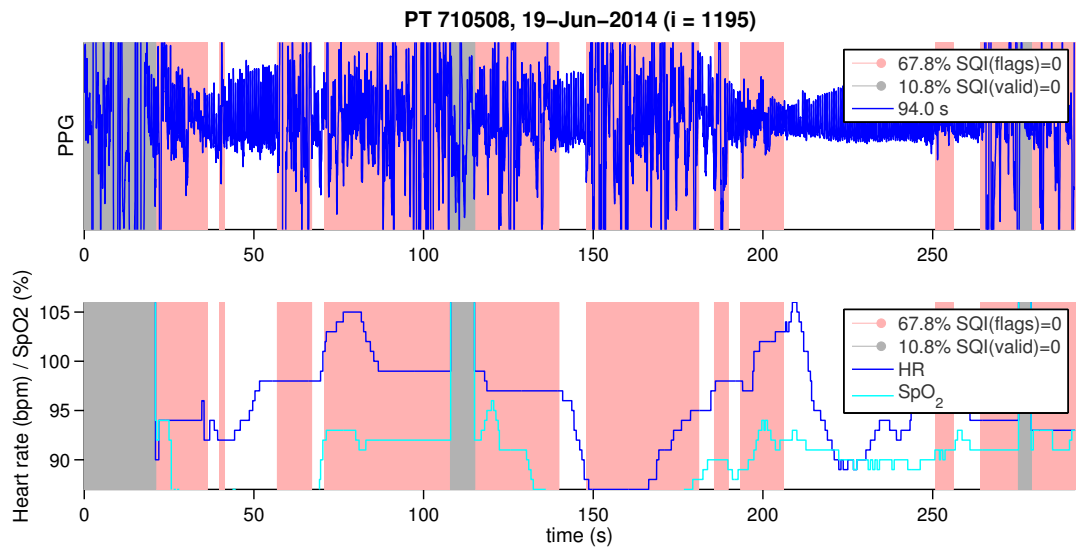


(c) Recording i=790

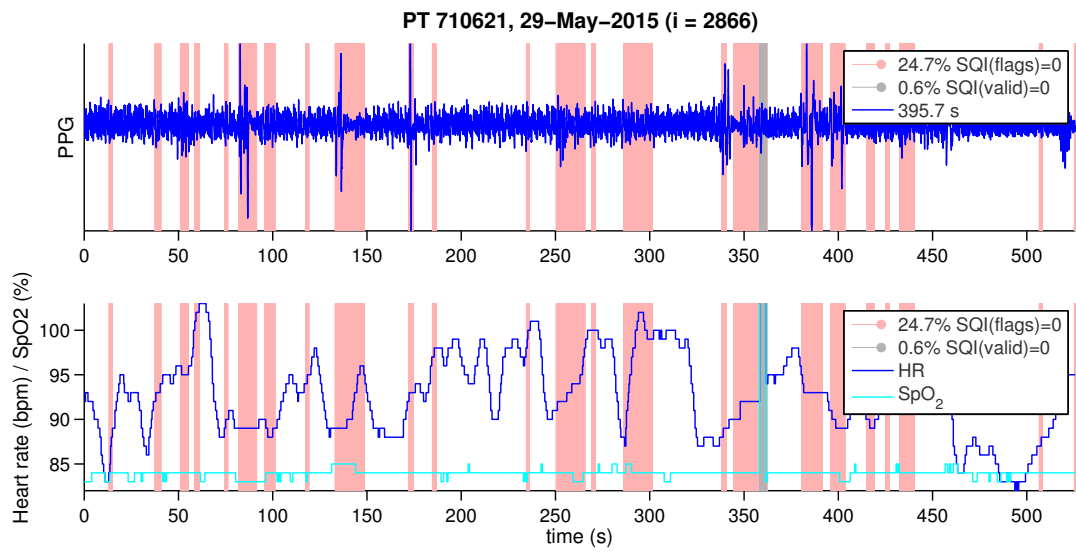
Figure C.2: Recordings with low values of $\text{SpO}_2 \leq 90\%$ that were not rejected (1 of 2)



(a) Recording i=931



(b) Recording i=1195



(c) Recording i=2866

Figure C.3: Recordings with low values of $\text{SpO}_2 \leq 90\%$ that were not rejected (2 of 2)

Appendix D

Supporting material: Centile estimation

D.1 Centile estimation

The information presented in this section is a supplement to Section 5.3.

D.1.1 Fitting of centiles to heart rate data

To select the appropriate polynomial order P of the multilevel model 3a and 3b fitted to each vital sign, we used a combination of criteria which included Akaike's Information Criterion (AIC), the Bayesian Information Criterion (BIC), log-likelihood ($\log \mathcal{L}$) and likelihood ratio tests to compare nested models. For each vital sign, we started with a polynomial order $P = 0$ and successively added fixed-effect terms. At each step, model fit was assessed using visual inspection of model fit and a likelihood ratio test to deem whether introduction of additional terms led to significant improvement in the log likelihood error metric.

After deciding on an appropriate order P (which is equivalent to determining the number of fixed effects in the model) we determined the number of random effects using a similar procedure where random effect terms were successively added and the model fit

compared. The best fitting model is the model with the lowest values of AIC, BIC and $\log \mathcal{L}$.

Table D.1 illustrates this process of model fitting for centiles fitted to the HR data. At each step (line in the table), the shorthand notation FP describes the number of fixed effects where P is the order of the polynomial expression (Ffp denotes fractional polynomial). The shorthand notation $Risc$ describes the random effects where i , s and c denote random intercept, random slope and random curvature, respectively. At each step, the final model is highlighted in colour.

Table D.1: Metrics for selection of number of fixed and random effects

Model name	df [†]	AIC	BIC	log $\mathcal{L}$	$\mathbf{p}^\dagger$	
Model 3a – Random intercept model						
F0Ri	3	191297	19314	−9645	—	
F1Ri	4	18946	18970	−9469	< .0001	(F0Ri)
F2Ri	5	18919	18948	−9454	< .0001	(F1Ri)
F3Ri	6	18913	18948	−9450	0.0056	(F2Ri)
F4Ri	7	18913	18954	−9450	0.14	(F3Ri)
F2Ri	$y_{ij} = \beta_0 + b_{0i} + \beta_1 x_{ij} + \beta_2 x_{ij}^2 + \beta_3 x_{ij}^3 + e_{ij}$					
Model 3b – Random slope model						
F3Ris	8	18880	18927	−9432	< .0001	(F3Ri)
F3Risc	11	18880	18945	−9429	0.10	(F3Ris)
F3Ris.ac ^{††}	9	18883	18935	−9432	0.54	(F3Ris)
F3Ris.hs ^{††}	9	18881	18934	−9431	0.26	(F3Ris)
F2Ris	$y_{ij} = \beta_0 + b_{0i} + \beta_1 x_{ij} + b_{1i} + \beta_2 x_{ij}^2 + \beta_3 x_{ij}^3 + e_{ij}$					
Model 3c – Fractional polynomial model						
FfpRi	5	18911	18941	−9451	—	
FfpRi	$y_{ij} = \beta_0 + b_{0i} + \beta_{p_1} x_{ij}^{p_1} + \beta_{p_2} x_{ij}^{p_2} + e_{ij}$					
[†] p values computed from a likelihood ratio test based on log $\mathcal{L}$ and degrees of freedom (df) [199]. This test may only be used to compare <i>nested</i> models (the model for comparison in brackets).						
^{††} matrix $\mathbf{R}_i$ defined to account for autocorrelation (ac) and heteroscedasticity (hs).						

D.1.2 Assessment of heteroscedasticity

To properly assess whether Models 3a and 3b accounted for heteroscedasticity (i.e. an increase or decrease in variance with gestational age), we plotted for each model the standardised observations (y_{ij}) and residuals (e_{ij}) against gestational age (Figure D.1).

For both plots, for the assumptions of normal distribution and constant variance to hold, the data points should fall approximately symmetrically about the line $y = 0$ and the variance should be constant across gestational ages. The slight tendency towards an increasing variance with gestational age for the observations in Model 3a is corrected in Model 3b through the inclusion of a random slope, which allowed for a heteroscedastic model variance.

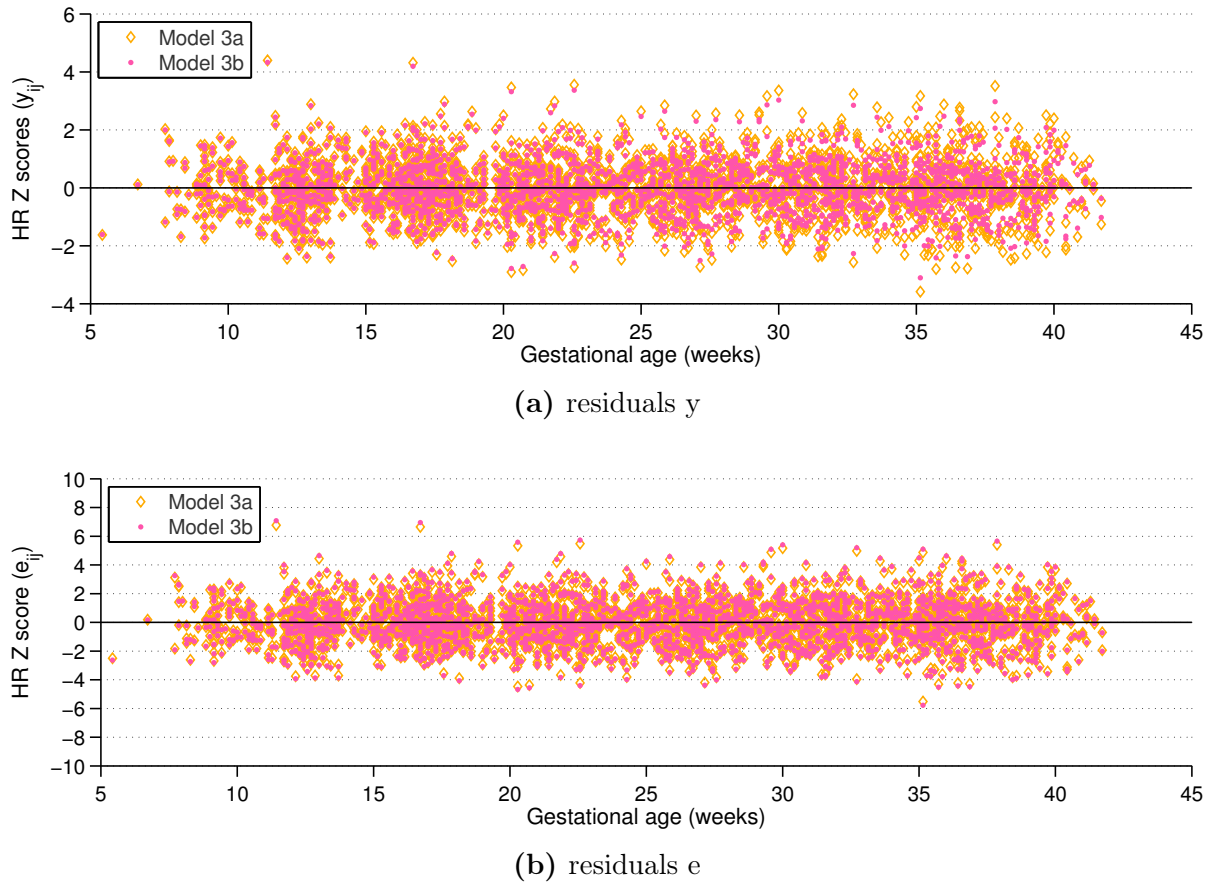


Figure D.1: Standardised (a) observations and (b) residuals of Models 3a (orange) and 3b (pink) with gestational age.

D.2 Equations for final centiles

Table D.2 summarises the equations of the estimated mean and centiles for each of the six vital signs HR, SpO₂, SBP, DBP, Temp and RR.

Table D.2: Final centiles for all six vital signs. Please refer to Appendix B.3 for definition of symbols. The time covariate is defined as $t_{ij} = x_{ij} - 10$, where x_{ij} is gestational age in weeks

Heart rate (bpm), $D = \{72.85, -0.19; -0.19, 0.05\}$, $\sigma^2 = \{48.85\}$

$$E(y_{ij}) = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 t_{ij}^3 = 84.64 + 0.31 t_{ij} + 0.02 t_{ij}^2 - 7 \times 10^{-4} t_{ij}^3$$

$$SD(y_{ij}) = \sqrt{\text{Var}(b_{0i} + b_{1i} t_{ij} + e_{ij})} = \sqrt{72.85 - 2 \times 0.19 t_{ij} + 0.05 t_{ij}^2 + 48.85}$$

Oxygen saturation (%), $D = \{0.47\}$, $\sigma^2 = \{1.21\}$

$$E(y_{ij}) = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 = 97.86 - 0.06 t_{ij} + 6 \times 10^{-4} t_{ij}^2$$

$$SD(y_{ij}) = \sqrt{\text{Var}(b_{0i} + e_{ij})} = \sqrt{0.47 + 1.21}$$

Systolic blood pressure (mmHg), $D = \{85.31, -0.50; -0.50, 0.07\}$, $\sigma^2 = \{53.16\}$

$$E(y_{ij}) = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 = 116.12 - 0.43 t_{ij} + 0.02 t_{ij}^2$$

$$SD(y_{ij}) = \sqrt{\text{Var}(b_{0i} + b_{1i} t_{ij} + e_{ij})} = \sqrt{85.31 - 2 \times 0.50 t_{ij} + 0.07 t_{ij}^2 + 53.16}$$

Diastolic blood pressure (mmHg), $D = \{47.49, -0.42; -0.42, 0.3\}$, $\sigma^2 = \{24.66\}$

$$E(y_{ij}) = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 = 71.77 - 0.45 t_{ij} + 0.02 t_{ij}^2$$

$$SD(y_{ij}) = \sqrt{\text{Var}(b_{0i} + b_{1i} t_{ij} + e_{ij})} = \sqrt{47.49 - 2 \times 0.42 t_{ij} + 0.03 t_{ij}^2 + 24.66}$$

Temperature (C°), $D = \{0.04\}$, $\sigma^2 = \{0.06\}$

$$E(y_{ij}) = \beta_0 + \beta_1 t_{ij} = 36.83 - 6 \times 10^{-3} t_{ij}$$

$$SD(y_{ij}) = \sqrt{\text{Var}(b_{0i} + e_{ij})} = \sqrt{0.04 + 0.06}$$

Respiratory rate (rpm), $D = \{3.03\}$, $\sigma^2 = \{6.72\}$

$$E(y_{ij}) = \beta_0 + \beta_1 t_{ij} = 16.6 + 0.02 t_{ij}$$

$$SD(y_{ij}) = \sqrt{\text{Var}(b_{0i} + e_{ij})} = \sqrt{3.03 + 6.72}$$

Bibliography

- [1] G. Lewis (ed) The Confidential Enquiry into Maternal and Child Health (CEMACH), *Saving Mothers' Lives: reviewing maternal deaths to make motherhood safer: 2003-2005. The Seventh Report on Confidential Enquiries into Maternal Deaths in the United Kingdom.* London, UK: CEMACH, 2007.
- [2] Health and Social Care Information Centre, *Hospital Episode Statistics: NHS Maternity Statistics – England, 2013-14.* London, England: Health and Social Care Information Centre, 2015.
- [3] Report by the Comptroller and Auditor General, *Department of Health – Maternity services in England.* London, England: National Audit Office, 2013.
- [4] M. Knight, S. Kenyon, P. Brocklehurst, J. Neilson, J. Shakespeare, J.J. Kurinczuk (Eds.) on behalf of MBRRACEUK, *Saving Lives, Improving Mothers' Care - Lessons learned to inform future maternity care from the UK and Ireland Confidential Enquiries into Maternal Deaths and Morbidity 2009-12.* Oxford, UK: National Perinatal Epidemiology Unit, University of Oxford, 2014.
- [5] G. Kayem, J. Kurinczuk, G. Lewis, S. Golightly, P. Brocklehurst, and M. Knight, "Risk factors for progression from severe maternal morbidity to death: A national cohort study," *PLoS One*, vol. 6, p. e29077, 2011.
- [6] N. Heslehurst, J. Rankin, J. Wilkinson, and C. Summerbell, "A nationally representative study of maternal obesity in england, uk: trends in incidence and demographic inequalities in 619,323 births, 1989-2007," *International Journal of Obesity*, vol. 34, no. 3, pp. 420–428, 2010.
- [7] M. Nair, J. Kurinczuk, P. Brocklehurst, S. Sellers, G. Lewis, and M. Knight, "Factors associated with maternal death from direct pregnancy complications: a uk national case-control study," *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 122, no. 5, pp. 653–662, 2015.
- [8] NHS Commissioning Board, *Commissioning Maternity Services: A Resource Pack to support Clinical Commissioning Groups.* London, England: NHS England, 2012.
- [9] N. J. Kassebaum, A. Bertozzi-Villa, M. S. Coggeshall, K. A. Shackelford, C. Steiner, K. R. Heuton, D. Gonzalez-Medina, R. Barber, C. Huynh, D. Dicker, T. Templin, T. M. Wolock, and A. A. Ozgoren, "Global, regional, and national levels and causes of maternal mortality during 1990–2013: a systematic analysis for the global burden of disease study 2013," *Lancet*, vol. 384, no. 9947, pp. 980–1004, 2014.

- [10] National Institute for Health and Care Excellence, *Hypertension in Pregnancy: The Management of Hypertensive Disorders During Pregnancy. NICE Clinical Guidelines, No. 107*. London, UK: RCOG Press, 2010.
- [11] P. R. James and C. Nelson-Piercy, "Management of hypertension before, during, and after pregnancy," *Heart*, vol. 90, pp. 1499–1504, 2004.
- [12] M. Brown, M. Lindheimer, M. de Swiet, A. Van Assche, and J. Moutquin, "The classification and diagnosis of the hypertensive disorders of pregnancy: statement from the international society for the study of hypertension in pregnancy (issHP)," *Hypertension in Pregnancy*, vol. 20, pp. 9–14, 2001.
- [13] J. A. Turner, "Diagnosis and management of pre-eclampsia: an update," *International Journal of Women's Health*, vol. 2, pp. 327–337, 2010.
- [14] Centre for Maternal and Child Enquiries (CMACE), "Saving mothers' lives: reviewing maternal deaths to make motherhood safer: 2006-08. the eighth report on confidential enquiries into maternal deaths in the united kingdom." *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 118(Suppl.1), pp. 1–203, 2011.
- [15] J. E. A. K. Bamfo, "Managing the risks of sepsis in pregnancy," *Best Practice & Research Clinical Obstetrics and Gynaecology*, vol. 27, pp. 583–595, 2013.
- [16] J. M. Ford and H. Scholefield, "Sepsis in obstetrics: cause, prevention, and treatment," *Current Opinion in Anaesthesiology*, vol. 27, pp. 253–258, 2014.
- [17] S. Paterson-Brown and C. Howell, *Managing Obstetric Emergency and Trauma - The MOET course manual*, 3rd ed. Cambridge, UK: Cambridge University Press, 2014.
- [18] M. Knight, "Antenatal pulmonary embolism: risk factors, management and outcomes," *BJOG*, vol. 115, no. 4, pp. 453–461, 2008.
- [19] C. Nelson-Piercy, P. MacCallum, and L. Mackillop, *Reducing the Risk of Venous Thromboembolism during Pregnancy and the Puerperium - Green-top Guideline No. 37a*. London, UK: Royal College of Obstetricians and Gynaecologists, April 2015.
- [20] P. Sultan, N. Arulkumaran, and A. Rhodes, "Provision of critical care services for the obstetric population," *Best Practice & Research Clinical Obstetrics and Gynaecology*, vol. 27, pp. 803–809, 2013.
- [21] M. Waterstone, S. Bewley, and C. Wolfe, "Incidence and predictors of severe obstetric morbidity: case-control study," *BMJ*, vol. 322, no. 7294, pp. 1089–1093, 2001.
- [22] A. McGlennan and K. Sherratt, "Charting change on the labour ward," *Anaesthesia*, vol. 68, no. 4, pp. 338–342, 2013.
- [23] NHS England, "National maternity review," <http://www.england.nhs.uk/ourwork/futurehhs/mat-review/>, 2016, [Online; accessed 22-Feb-2016].
- [24] National Institute for Health and Care Excellence, "Clinical guideline: Acute illness in adults in hospital: recognising and responding to deterioration," <http://nice.org.uk/guidance/cg50>, 2010, [Online; accessed 4-Dec-2015].
- [25] M. Scrutton and I. Gardner, "Maternal critical care in the uk: Developing the service," *International Journal of Obstetric Anaesthesia*, vol. 21, pp. 291–293, 2012.

- [26] M. A. F. Pimentel, “Modeling of vital sign data from post-operative patients,” Ph.D. dissertation, University of Oxford, Oxford, UK, 2015.
- [27] Royal College of Physicians, Report of a working party, *National Early Warning Score (NEWS): Standardising the assessment of acute- illness severity in the NHS*. London, UK: Royal College of Physicians, July 2012.
- [28] D. Prytherch, G. Smith, P. Schmidt, and P. Featherstone, “ViEWS—towards a national early warning score for detecting adult inpatient deterioration,” *Resuscitation*, vol. 81, no. 8, pp. 932–937, 2010.
- [29] J. R. Lappen, M. Keene, M. Lore, W. A. Grobman, and D. R. Gossett, “Existing models fail to predict sepsis in an obstetric population with intrauterine infection,” *American Journal of Obstetrics & Gynecology*, vol. 203, pp. 573.e1–5, 2010.
- [30] C. Carle, P. Alexander, M. Columb, and J. Johal, “Design and internal validation of an obstetric early warning score: secondary analysis of the intensive care national audit and research centre case mix programme database,” *Anaesthesia*, vol. 67, pp. 12–18, 2012.
- [31] E. R. Seiler, “Obstetrical chart,” *The Journal of the Royal College of General Practitioners*, vol. 20, pp. 59–60, 2000.
- [32] R. A. Isaacs, M. Y. K. Wee, D. E. Bick, S. Beake, Z. A. Sheppard, S. Thomas, V. Hundley, G. B. Smith, E. van Teijlingen, and P. W. Thomas, “A national survey of obstetric early warning systems in the united kingdom: five years on,” *Anaesthesia*, vol. 69, pp. 687–692, 2014.
- [33] R. D. J. Swanton, S. Al-Rawi, and M. Y. K. Wee, “A national survey of obstetric early warning systems in the united kingdom,” *International Journal of Obstetric Anesthesia*, vol. 18, no. 3, pp. 253–257, 2009.
- [34] D. Bick, J. Sandall, M. Furuta, M. Wee, R. Isaacs, G. Smith, and S. Beake, “A national cross sectional survey of heads of midwifery services of uptake, benefits and barriers to use of obstetric early warning systems (ews) by midwives,” *Midwifery*, vol. 30, no. 11, pp. 1140–1146, 2014.
- [35] S. Singh, A. McGlennan, A. England, and R. Simons, “A validation study of the cemach recommended modified early obstetric warning system (meows),” *Anaesthesia*, vol. 68, pp. 354–367, 2013.
- [36] The Maternal Critical Care Working Group, *Providing equity of critical and maternity care for the critically ill pregnant or recently pregnant woman*. London, UK: Royal College of Obstetricians and Gynaecologists, July 2011.
- [37] F. Baskett, “Epidemiology of obstetric critical care,” *Best Practice & Research Clinical Obstetrics and Gynaecology*, vol. 22, no. 5, pp. 763–774, 2008.
- [38] S. E. Edwards, W. A. Grobman, J. R. Lappen, C. Winter, R. Fox, E. Lenguerrand, and T. Draycott, “Modified obstetric early warning scoring systems (moews): validating the diagnostic performance for severe sepsis in women with chorioamnionitis,” *American Journal of Obstetrics & Gynecology*, vol. 212, pp. 536.e1–8, 2015.

- [39] P. Palatini, A. Benetos, G. Grassi, S. Julius, S. E. Kjeldsen, G. Mancia, K. Narkiewicz, G. Parati, A. C. Pessina, L. M. Ruilope, and A. Zanchetti, "Identification and management of the hypertensive patient with elevated heart rate: Statement of a european society of hypertension consensus meeting," *Journal of Hypertension*, vol. 24, no. 4, pp. 603–610, 2006.
- [40] G. Chamberlain and F. B. Pipkin, *Clinical Physiology in Obstetrics*. MA, USA: Blackwell Science Inc., 1998.
- [41] A. Mahendru, T. Everett, I. Wilkinson, C. Lees, and C. McEniery, "A longitudinal study of maternal cardiovascular function from preconception to the postpartum period," *Journal of Hypertension*, vol. 32, p. 0, 2014.
- [42] C. L. Stanfield, *Principles of Human Physiology*, 4th ed. New York, USA: Pearson, 2010.
- [43] A. Plüddemann, M. Thompson, C. Heneghan, and C. Price, "Pulse oximetry in primary care: primary care diagnostic technology update," *The British Journal of General Practice*, vol. 61, no. 586, pp. 358–359, 2011.
- [44] J. R. Kambam, R. E. Handte, W. U. Brown, and B. E. Smith, "Effect of normal and preeclamptic pregnancy on the oxyhemoglobin dissociation curve," *Anesthesiology*, vol. 65, pp. 426–427, 1986.
- [45] A. Templeton and G. R. Kelman, "Maternal blood gases, (PAO₂ - PaO₂), physiological shunt and VD/VT in normal pregnancy," *British Journal of Anaesthesia*, vol. 48, no. 10, pp. 1001–1004, 1976.
- [46] G. Grindheim, M. E. Estensen, E. Langesaeter, L. A. Rosseland, and K. Toska, "Changes in blood pressure during healthy pregnancy: a longitudinal cohort study," *Journal of Hypertension*, vol. 30, no. 2, pp. 342–50, 2012.
- [47] M. De Swiet, *Medical Disorders in Obstetric Practice*. Wiley, 2008.
- [48] N. P. O'Grady, P. S. Barie, J. G. Bartlett, T. Bleck, K. Carroll, A. C. Kalil, P. Linden, D. G. Maki, D. Nierman, W. Pasculle, and H. Masur, "Guidelines for evaluation of new fever in critically ill adult patients: 2008 update from the american college of critical care medicine and the infectious diseases society of america," *Critical Care Medicine*, vol. 36, no. 4, pp. 1330–1349, 2008.
- [49] D. J. Tuffnell, P. C. Buchan, D. Albert, and S. Tyndale-Biscoe, "Fetal heart rate responses to maternal exercise, increased maternal temperature and maternal circadian variation," *Journal of Obstetrics and Gynaecology*, vol. 10, no. 5, pp. 387–391, 1990.
- [50] D. R. Goldhill, S. A. White, and A. Sumner, "Physiological values and procedures in the 24 h before icu admission from the ward," *Anaesthesia*, vol. 54, no. 6, pp. 529–534, 1999.
- [51] C. W. Seymour, J. M. Kahn, C. R. Cooke, T. R. Watkins, S. R. Heckbert, and T. D. Rea, "Prediction of critical illness during out-of-hospital emergency care," *JAMA*, vol. 304, no. 7, pp. 747–754, 2010.
- [52] M. Elliott and A. Coventry, "Critical care: the eight vital signs of patient monitoring," *British Journal of Nursing*, vol. 21, no. 10, pp. 621–625, 2012.

- [53] *WristOx₂ Model 3150 Bluetooth wrist worn pulse oximeter OEM specification and technical information*, http://www.nonin.com/documents/IFUManuals/3150_ENG.pdf, Nonin Medical, Inc., Minnesota, USA, 2011, [Online; accessed 4-Dec-2015].
- [54] D. J. Meredith, D. Clifton, P. Charlton, J. Brooks, C. W. Pugh, and L. Tarassenko, "Photoplethysmographic derivation of respiratory rate: a review of relevant physiology," *Journal of Medical Engineering & Technology*, vol. 36, no. 1, pp. 1–7, 2012.
- [55] G. Ogedegbe and T. Pickering, "Principles and techniques of blood pressure measurement," *Cardiology Clinics*, vol. 28, no. 4, pp. 571–586, 2010.
- [56] G. Parati, G. S. Stergiou, R. Asmar, G. Bilo, P. de Leeuw, Y. Imai, K. Kario, E. Lurbe, A. Manolis, T. Mengden, E. O'Brien, T. Ohkubo, P. Padfield, P. Palatini, T. G. Pickering, J. Redon, M. Revere, L. M. Ruilope, A. Shennan, J. A. Staessen, A. Tisler, B. Waeber, A. Zanchetti, and G. Mancia, "European society of hypertension practice guidelines for home blood pressure monitoring," *Journal of Human Hypertension*, vol. 24, no. 12, pp. 779–785, 2010.
- [57] E. O'Brien, J. Petrie, W. Littler, M. de Swiet, P. L. Padfield, D. G. Altman, M. Bland, A. Coats, and N. Atkins, "The British Hypertension Society protocol for the evaluation of blood pressure measuring devices," *Journal of Hypertension*, vol. 11, pp. 677–679, 1993.
- [58] *Fora IR20 Ear Thermometer Owner's Manual*, 3rd ed., <http://www.foracare.com/downloads/Owner-booklet-FORA-IR20-EN-ES.pdf>, Fora Care, California, USA, 2012, [Online; accessed 4-Dec-2015].
- [59] K. K. Giuliano, A. J. Giuliano, S. S. Scott, E. MacLachlan, E. Pysznik, S. Elliot, and D. Woytowicz, "Temperature measurement in critically ill adults: a comparison of tympanic and oral methods," *American Journal of Critical Care*, vol. 9, no. 4, pp. 254–261, 2000.
- [60] World Health Organization, "Fourth programme report, 1988-1989: ARI programme for control of acute respiratory infections." <http://apps.who.int/iris/handle/10665/58546>, Geneva, Switzerland, 1990, [Online; accessed 4-Dec-2015].
- [61] W. Bianchi, A. F. Dugas, Y.-H. Hsieh, M. Saheed, P. Hill, C. Lindauer, A. Terzis, and R. E. Rothman, "Revitalizing a vital sign: Improving detection of tachypnea at primary triage," *Annals of Emergency Medicine*, vol. 61, no. 1, pp. 37–43, 2013.
- [62] Z. V. Edmonds, W. R. Mower, L. M. Lovato, and R. Lomeli, "The reliability of vital sign measurements," *Annals of Emergency Medicine*, vol. 39, no. 3, pp. 233–237, 2002.
- [63] J. Hogan, "Why don't nurses monitor the respiratory rates of patients?" *British journal of nursing*, vol. 15, no. 9, pp. 489–492, 2006.
- [64] R. J. Cooper and S. M. Green, "Don't hyperventilate over triage respiratory rates," *Annals of Emergency Medicine*, vol. 61, no. 1, pp. 44–45, 2013.
- [65] P. B. Lovett, J. M. Buchwald, K. Starmann, and P. Bijur, "The vexatious vital: Neither clinical measurements by nurses nor an electronic monitor provides accurate measurements of respiratory rate in triage," *Annals of Emergency Medicine*, vol. 45, no. 1, pp. 68–76, 2005.

- [66] C. Winter, J. Crofts, C. Laxton, S. Barnfield, and T. Draycott, *PROMPT course manual*, 2nd ed. Cambridge, UK: Cambridge University Press, 2012.
- [67] D. Moher, A. Liberati, J. Tetzlaff, and D. Altman, "The PRISMA group. preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *BMJ*, vol. 339, p. b2535, 2009.
- [68] D. Moher, L. Shamseer, M. Clarke, D. Ghersi, A. Liberati, M. Petticrew, P. Shekelle, and L. Stewart, "Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement," *Systematic Reviews*, vol. 4, no. 1, p. 1, 2015.
- [69] L. Loerup, R. M. Pullon, J. Birks, S. Fleming, L. H. Mackillop, and P. J. Watkinson, "Trends of vital signs with gestational age in normal pregnancies: a systematic review protocol," *BMJ Open*, vol. 6, p. e008769, 2016.
- [70] C. Ioannou, K. Talbot, E. Ohuma, I. Sarris, J. Villar, A. Conde-Agudelo, and A. Papageorgiou, "Systematic review of methodology used in ultrasound studies aimed at creating charts of fetal size," *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 119, no. 12, pp. 1425–1439, 2012.
- [71] P. F. Whiting, A. W. Rutjes, M. E. Westwood, S. Mallett, J. J. Deeks, J. B. Reitsma, M. M. Leeflang, J. A. Sterne, and P. M. Bossuyt, "QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies," *Annals of Internal Medicine*, vol. 155, no. 8, pp. 529–536, 2011.
- [72] A. P. Jones, R. D. Riley, P. R. Williamson, and A. Whitehead, "Meta-analysis of individual patient data versus aggregate data from longitudinal clinical trials," *Clinical Trials*, vol. 6, pp. 16–27, 2009.
- [73] S. Fleming, M. Thompson, R. Stevens, C. Heneghan, A. Pluddemann, I. Maconochie, L. Tarassenko, and D. Mant, "Normal ranges of heart rate and respiratory rate in children from birth to 18 years of age: a systematic review of observational studies," *The Lancet*, vol. 377, no. 9770, pp. 1011 – 1018, 2011.
- [74] K. J. Ishak, R. W. Platt, L. Joseph, J. A. Hanley, and J. J. Caro, "Meta-analysis of longitudinal studies," *Clinical Trials*, vol. 4, no. 5, pp. 525–539, 2007.
- [75] S. Fleming, "Measurement and fusion of non-invasive vital signs for routine triage of acute paediatric illness," Ph.D. dissertation, University of Oxford, Oxford, UK, 2010.
- [76] A. S. Bryk and S. W. Raudenbush, "Application of hierarchical linear models to assessing change," *Psychological Bulletin*, vol. 101, pp. 147–158, 1987.
- [77] V. N. Demidov, "Comparative evaluation of the changes in the cardiac activity and hemodynamics of healthy women at the end of pregnancy while lying on their sides and on their backs," *Voprosy okhrany materinstva i detstva*, vol. 23, no. 3, pp. 71–75, 1978, [Russian].
- [78] N. Ochskenbein-Kolble, M. Roos, T. Gasser, R. Huch, A. Huch, and R. Zimmermann, "Cross sectional study of automated blood pressure measurements throughout pregnancy," *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 111, no. 4, pp. 319–25, 2004.
- [79] C. Macdonald-Wallis, J. Silverwood, A. Fraser, M. Nelson, K. Tilling, A. Lawlor, and L. B. de, "Gestational-age-specific reference ranges for blood pressure in pregnancy: findings from a prospective cohort," *Journal of Hypertension*, vol. 33, no. 1, p. 96, 2015.

- [80] H. Strevens, K. Kristensen, J. Langhoff-Roos, and D. Wide-Svensson, "Blood pressure patterns through consecutive pregnancies are influenced by body mass index," *American Journal of Obstetrics & Gynecology*, vol. 187, no. 5, pp. 1343–8, 2002.
- [81] M. Ishikuro, T. Obara, H. Metoki, T. Ohkubo, M. Yamamoto, K. Akutsu, K. Sakurai, N. Iwama, M. Katagiri, K. Yagihashi, N. Yaegashi, S. Mori, M. Suzuki, S. Kuriyama, and Y. Imai, "Blood pressure measured in the clinic and at home during pregnancy among nulliparous and multiparous women: the boshi study," *American Journal of Hypertension*, vol. 26, no. 1, pp. 141–8, 2013.
- [82] G. A. N. Khan, N. Ishrat, and N. A. Sabzposh, "Blood pressure pattern in pregnant women of different body mass index in three trimesters of pregnancy," *Indian Journal of Public Health Research and Development*, vol. 4, no. 3, pp. 98–102, 2013.
- [83] E. Jasovic-Siveska and V. Jasovic, "Prediction of mild and severe preeclampsia with blood pressure measurements in first and second trimester of pregnancy," *Ginekologia Polska*, vol. 82, no. 11, pp. 845–50, 2011.
- [84] C. Macdonald-Wallis, D. A. Lawlor, A. Fraser, M. May, S. M. Nelson, and K. Tilling, "Blood pressure change in normotensive, gestational hypertensive, preeclamptic, and essential hypertensive pregnancies," *Hypertension*, vol. 59, no. 6, pp. 1241–8, 2012.
- [85] O. Turan, C. DePaco, N. Kametas, A. Khaw, and K. Nicolaides, "Effect of parity on maternal cardiac function during the first trimester of pregnancy," *Ultrasound in Obstetrics & Gynecology*, vol. 32, pp. 849–854, 2008.
- [86] Office for National Statistics, "Births in England and Wales in 2014," http://www.ons.gov.uk/ons/dcp171778_410897.pdf, 2015, [Online; accessed 4-Dec-2015].
- [87] E. Falaschetti, J. Mindell, C. Knott, and N. Poulter, "Hypertension management in england: a serial cross-sectional study from 1994 to 2011," *Lancet*, vol. 383, pp. 1912–1919, 2014.
- [88] R. Gaillard and V. W. Jaddoe, "Assessment of maternal blood pressure development during pregnancy," *Journal of Hypertension*, vol. 33, pp. 61–62, 2015.
- [89] M. Egger, G. D. Smith, and D. G. Altman, *Statistical Methods for Examining Heterogeneity and Combining Results from Several Studies in Meta-Analysis*. Oxford, UK: John Wiley & Sons, 2008.
- [90] E. Borghi, M. de Onis, C. Garza, J. V. den Broeck, E. A. Frongillo, Grummer-Strawn, S. V. Buuren, H. Pan, L. Molinari, R. Martorell, W. Onyango, and J. C. Martines, "Construction of the world health organization child growth standards: selection of methods for attained growth curves," *Statistics in Medicine*, vol. 25, pp. 247–265, 2006.
- [91] A. M. VanDerGraaf, G. G. Zeeman, H. Groen, C. Roberts, and G. A. Dekker, "Non-invasive assessment of maternal hemodynamics in early pregnancy," *Pregnancy Hypertension*, vol. 3, no. 4, pp. 261–269, 2013.
- [92] The INTERBIO–21st Consortium, "Interbio–21st study protocol," www.interbio21.org.uk, Oxford, UK, 2012, [Online; accessed 21-Jan-2016].
- [93] K. Butt and K. Lim, "Determination of gestational age by ultrasound," *Journal of Obstetrics and Gynaecology Canada*, vol. 36, no. 2, pp. 171–181, 2014.

- [94] A. Reinders, A. C. Cuckson, J. T. M. Lee, and A. H. Shennan, "An accurate automated blood pressure device for use in pregnancy and pre-eclampsia: the Microlife 3BTO-A," *BJOG: An International Journal of Obstetrics and Gynaecology*, vol. 112, no. 7, pp. 915–920, 2005.
- [95] J. Villar, A. T. Papageorgiou, R. Pang, E. O. Ohuma, L. C. Ismail, F. C. Barros, A. Lambert, M. Carvalho, Y. A. Jaffer, E. Bertino, M. G. Gravett, D. G. Altman, M. Purwar, I. O. Frederick, J. A. Noble, C. G. Victora, Z. A. Bhutta, and S. H. Kennedy, "The likeness of fetal growth and newborn size across non-isolated populations in the intergrowth-21st project: the fetal growth longitudinal study and newborn cross-sectional study," *The Lancet Diabetes & Endocrinology*, vol. 2, no. 10, pp. 781–792, 2014.
- [96] R. C. Hermida and D. E. Ayala, "Prognostic value of office and ambulatory blood pressure measurements in pregnancy," *Hypertension*, vol. 40, no. 3, pp. 298–303, 2002.
- [97] M. Brown, A. Robinson, and M. Jones, "The white coat effect in pregnancy: much ado about nothing?" *BJOG: An International Journal of Obstetrics and Gynaecology*, vol. 106, pp. 474–480, 1999.
- [98] J. Bland and D. Altman, "Measuring agreement in method comparison studies," *Statistical methods in medical research*, vol. 8, pp. 135–160, 1999.
- [99] —, "Statistical methods for assessing agreement between two methods of clinical measurements," *The Lancet*, vol. 8476, no. 1, pp. 307–310, 1986.
- [100] E. Ohuma, L. Hoch, C. Cosgrove, H. Knight, L. Cheikh-Ismail, L. Juodvirsiene, A. Papageorgiou, H. Al-Jabri, M. Domingues, P. Gilli, N. Kunnawar, N. Musee, F. Roseman, A. Carter, M. Wu, and D. Altman, "Managing data for the international, multicentre INTERGROWTH-21st Project," *BJOG: An International Journal of Obstetrics and Gynaecology*, vol. 120, no. s2, pp. 64–70, 2013, for the International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st).
- [101] J. Brooke, "Sus – a quick and dirty usability scale," <http://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html>, Reading, UK, 1986, [Online; accessed 21-Jan-2016].
- [102] D. Altman and E. Ohuma, "Statistical considerations for the development of prescriptive fetal and newborn growth standards in the INTERGROWTH-21st project," *BJOG: An International Journal of Obstetrics and Gynaecology*, vol. 120, no. s2, pp. 71–76, 2013, for the International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st).
- [103] G. B. Smith, D. R. Prytherch, P. E. Schmidt, and P. I. Featherstone, "Review and performance evaluation of aggregate weighted 'track and trigger' systems," *Resuscitation*, vol. 77, no. 2, pp. 170–179, 2008.
- [104] M. Folke, L. Cernerud, M. Ekström, and B. Hök, "Critical review of non-invasive respiratory monitoring in medical care," *Medical & Biological Engineering & Computing*, vol. 41, pp. 377–383, 2003.
- [105] F. Al-Khalidi, R. Saatchi, D. Burke, H. Elphick, and S. Tan, "Respiration rate monitoring methods: A review," *Pediatric Pulmonology*, vol. 46, no. 6, pp. 523–529, 2011.

- [106] W. Karlen, S. Raman, J. M. Ansermino, and G. A. Dumont, "Multiparameter respiratory rate estimation from the photoplethysmogram," *Biomedical Engineering, IEEE Transactions on*, vol. 60, no. 7, pp. 1946–1953, 2013.
- [107] J. A. Hirsch and B. Bishop, "Respiratory sinus arrhythmia in humans: how breathing pattern modulates heart rate," *American Journal of Physiology - Heart and Circulatory Physiology*, vol. 241, no. 4, pp. 620–629, 1981.
- [108] F. Yasuma and J. Hayano, "Respiratory sinus arrhythmia: why does the heartbeat synchronize with respiratory rhythm?" *Chest*, vol. 125, pp. 683–690, 2004.
- [109] Electrophysiology, Task Force of the European Society of Cardiology the North American Society of Pacing, "Heart rate variability: Standards of measurement, physiological interpretation, and clinical use," *Circulation*, vol. 93, no. 5, pp. 1043–1065, 1996.
- [110] J. Pumprla, K. Howorka, D. Groves, M. Chester, and J. Nolan, "Functional assessment of heart rate variability: physiological basis and practical applications," *International Journal of Cardiology*, vol. 84, no. 1, pp. 1 – 14, 2002.
- [111] N. Montano, A. Porta, C. Cogliati, G. Costantino, E. Tobaldini, K. R. Casali, and F. Iellamo, "Heart rate variability explored in the frequency domain: A tool to investigate the link between heart and behavior," *Neuroscience & Biobehavioral Reviews*, vol. 33, no. 2, pp. 71 – 80, 2009.
- [112] P. D. Hung, S. Bonnet, R. Guillemaud, E. Castelli, and P. T. N. Yen, "Estimation of respiratory waveform using an accelerometer," in *Conference proceedings: 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2008*, 2008, pp. 1493–1496.
- [113] D. H. Phan, S. Bonnet, R. Guillemaud, E. Castelli, and N. Y. P. Thi, "Estimation of respiratory waveform and heart rate using an accelerometer," in *Conference proceedings: IEEE Engineering in Medicine and Biology Society*, 2008, pp. 4916–4919.
- [114] A. Jin, B. Yin, G. Morren, H. Duric, and R. M. Aarts, "Performance evaluation of a tri-axial accelerometry-based respiration monitoring for ambient assisted living," *Conference proceedings : IEEE Engineering in Medicine and Biology Society*, vol. 2009, pp. 5677–5680, 2009.
- [115] G. Z. Liu, Y. W. Guo, Q. S. Zhu, B. Y. Huang, and L. Wang, "Estimation of respiration rate from three-dimensional acceleration data based on body sensor network," *Telemedicine Journal and E-health*, vol. 17, no. 9, pp. 705–711, 2011.
- [116] G. B. Drummond, A. Bates, J. Mann, and D. K. Arvind, "Validation of a new non-invasive automatic monitor of respiratory rate for postoperative subjects," *British Journal of Anaesthesia*, vol. 107, no. 3, pp. 462–469, 2011.
- [117] A. Bates, M. J. Ling, J. Mann, and D. K. Arvind, "Respiratory rate and flow waveform estimation from tri-axial accelerometer data," *IEEE Body Sensor Networks International Conference*, pp. 144–50, 2010.
- [118] A. Bates, M. J. Ling, C. Geng, A. Turk, and D. K. Arvind, "Accelerometer-based respiratory measurement during speech," *Body Sensor Networks International Conference*, pp. 95–100, 2011.

- [119] J. Mann, R. Rabinovich, A. Bates, S. Giavedoni, W. MacNee, and D. K. Arvind, "Simultaneous activity and respiratory monitoring using an accelerometer," *IEEE Body Sensor Networks International Conference*, pp. 139–143, 2011.
- [120] G. B. Drummond, A. Bates, J. Mann, and D. K. Arvind, "Characterization of breathing patterns during patient-controlled opioid analgesia," *British Journal of Anaesthesia*, vol. 111, no. 6, pp. 971–978, 2013.
- [121] T. Ono, H. Takegawa, T. Ageishi, M. Takashina, H. Numasaki, M. Matsumoto, and T. Teshima, "Respiratory monitoring with an acceleration sensor," *Physics in Medicine and Biology*, vol. 56, pp. 6279–6289, 2011.
- [122] T. Pechprasarn and S. Pongnumkul, "Estimation of respiratory rate from smartphone's acceleration data," *2013 10th IEEE International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, vol. 2013, pp. 1–5, 2013.
- [123] A. Schafer and J. Vagedes, "How accurate is pulse rate variability as an estimate of heart rate variability?: A review on studies comparing photoplethysmographic technology with an electrocardiogram," *International journal of cardiology*, vol. 166, no. 1, pp. 15–29, 2013.
- [124] L. Nilsson, A. Johansson, and S. Kalman, "Respiration can be monitored by photoplethysmography with high sensitivity and specificity regardless of anaesthesia and ventilatory mode," *Acta Anaesthesiologica Scandinavica*, vol. 49, no. 8, pp. 1157–1162, 2005.
- [125] L. Nilsson, A. Johanson, and S. Kalman, "Monitoring of respiratory rate in postoperative care using a new photoplethysmographic technique," *Journal of clinical monitoring and computing*, vol. 16, no. 4, pp. 309–315, 2000.
- [126] J. Y. A. Foo and J. Stephen, "Estimation of breathing interval from the photoplethysmographic signals in children," *Physiological Measurement*, vol. 26, no. 6, p. 1049, 2005.
- [127] K. Nakajima, T. Tamura, and H. Miike, "Monitoring of heart and respiratory rates by photoplethysmography using a digital filtering technique," *Medical engineering & physics*, vol. 18, no. 5, pp. 365–372, 1996.
- [128] D. Wertheim, C. Olden, E. Savage, and P. Seddon, "Extracting respiratory data from pulse oximeter plethysmogram traces in newborn infants," *Archives of Disease in Childhood - Fetal and Neonatal Edition*, vol. 94, no. 4, pp. F301–F303, 2009.
- [129] K. H. Shelley, A. A. Awad, R. G. Stout, and D. G. Silverman, "The use of joint time frequency analysis to quantify the effect of ventilation on the pulse oximeter waveform," *Journal of clinical monitoring and computing*, vol. 20, no. 2, pp. 81–87, 2006.
- [130] D. Wertheim, C. Olden, L. Symes, H. Rabe, and P. Seddon, "Monitoring respiration in wheezy preschool children by pulse oximetry plethysmogram analysis," *Medical & biological engineering & computing*, vol. 51, pp. 965–970, 2013.
- [131] R. Takalo, H. Hytti, and H. Ihalainen, "Tutorial on univariate autoregressive spectral analysis," *Journal of clinical monitoring and computing*, vol. 19, no. 6, pp. 401–410, 2005.

- [132] S. G. Fleming and L. Tarassenko, "A comparison of signal processing techniques for the extraction of breathing rate from the photoplethysmogram," *International Journal of Biological and Medical Sciences*, vol. 2, pp. 232–236, 2007.
- [133] S. Fleming, L. Tarassenko, M. Thompson, and D. Mant, "Non-invasive measurement of respiratory rate in children using the photoplethysmogram," in *Conference proceedings : IEEE Engineering in Medicine and Biology Society*, 2008, pp. 1886–1889.
- [134] J. Lee and K. H. Chon, "Respiratory rate extraction via an autoregressive model using the optimal parameter search criterion," *Annals of Biomedical Engineering*, vol. 38, no. 10, pp. 3218–3225, 2010.
- [135] J. Lee, J. P. Florian, and H. Ki, "Respiratory rate extraction from pulse oximeter and electrocardiographic recordings," *Physiological Measurement*, vol. 32, no. 11, p. 1763, 2011.
- [136] P. Leonard, N. Grubb, P. Addison, D. Clifton, and J. Watson, "An algorithm for the detection of individual breaths from the pulse oximeter waveform," *Journal of clinical monitoring and computing*, vol. 18, no. 5-6, pp. 309–312, 2004.
- [137] P. S. Addison and J. N. Watson, "Secondary transform decoupling of shifted nonstationary signal modulation components: Application to photoplethysmography," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 2, no. 1, pp. 43–57, 2004.
- [138] P. Addison, J. Watson, M. Mestek, and R. Mecca, "Developing an algorithm for pulse oximetry derived respiratory rate (rroxi): a healthy volunteer study," *Journal of clinical monitoring and computing*, vol. 26, no. 1, pp. 45–51, 2012.
- [139] D. Clifton, J. G. Douglas, P. S. Addison, and J. N. Watson, "Measurement of respiratory rate from the photoplethysmogram in chest clinic patients," *Journal of clinical monitoring and computing*, vol. 21, no. 1, pp. 55–61, 2007.
- [140] P. A. Leonard, D. Clifton, P. S. Addison, J. N. Watson, and T. Beattie, "An automated algorithm for determining respiratory rate by photoplethysmogram in children," *Acta Paediatrica*, vol. 95, no. 9, pp. 1124–1128, 2006.
- [141] P. A. Leonard, J. G. Douglas, N. R. Grubb, D. Clifton, P. S. Addison, and J. N. Watson, "A fully automated algorithm for the determination of respiratory rate from the photoplethysmogram," *Journal of clinical monitoring and computing*, vol. 20, no. 1, pp. 33–36, 2006.
- [142] K. V. Madhav, M. R. Ram, E. H. Krishna, K. N. Reddy, and K. A. Reddy, "Estimation of respiratory rate from principal components of photoplethysmographic signals," in *Biomedical Engineering and Sciences (IECBES), 2010 IEEE EMBS Conference on*, 2010, pp. 311–314.
- [143] Y. Zhou, Y. Zheng, C. Wang, and J. Yuan, "Extraction of respiratory activity from photoplethysmographic signals based on an independent component analysis technique: Preliminary report," *Instrumentation Science & Technology*, vol. 34, no. 5, pp. 537–545, 2006.
- [144] L. Mason, "Signal processing methods for non-invasive respiration monitoring," Ph.D. dissertation, University of Oxford, Oxford, UK, 2002.

- [145] S. A. Shah, “Vital sign monitoring and data fusion for paediatric triage,” Ph.D. dissertation, University of Oxford, Oxford, UK, 2012.
- [146] C. Orphanidou, S. Fleming, S. Shah, and L. Tarassenko, “Data fusion for estimating respiratory rate from a single-lead ecg,” *Biomedical Signal Processing and Control*, vol. 8, pp. 98–105, 2013.
- [147] S. J. Orfanidis, *Introduction to Signal Processing*. New Jersey, USA: Prentice-Hall, 1995.
- [148] S. Nemati, A. Malhotra, and G. D. Clifford, “Data fusion for improved respiration rate estimation,” *EURASIP Journal on Advances in Signal Processing*, vol. 2010, no. 10, 2010.
- [149] J.-W. Yoon, Y.-S. Noh, Y.-S. Kwon, S.-B. Park, W.-K. Kim, and H.-R. Yoon, “Improvement of dynamic respiration monitoring through sensor fusion of accelerometer and gyro-sensor,” *Journal of Electrical Engineering and Technology*, vol. 9, no. 1, pp. 334–343, 2014.
- [150] P. Gandhi, H. Mehta, A. Gokhale, C. Desai, P. Gokhale, and C. Shah, “A study on cardiac autonomic modulation during pregnancy by non-invasive heart rate variability measurement,” *International Journal of Medicine and Public Health*, vol. 4, no. 4, pp. 441–445, 2014.
- [151] W. Karlen, C. Brouse, E. Cooke, J. Ansermino, and G. Dumont, “Respiratory rate estimation using respiratory sinus arrhythmia from photoplethysmography,” in *Conference proceedings of the 33rd annual International Conference of the IEEE EMBS*, 2011, pp. 1201–1204.
- [152] W. Karlen, H. Gan, M. Chiu, D. Dunsmuir, G. Zhou, G. A. Dumont, and J. M. Ansermino, “Improving the accuracy and efficiency of respiratory rate measurements in children using mobile devices,” *PLoS ONE*, vol. 9, no. 6, p. e99266, 2014.
- [153] A. Papageorgiou, E. Ohuma, D. Altman, T. Todros, L. Cheikh-Ismail, A. Lambert, Y. A. Jaffer, E. Bertino, M. Gravett, M. Purwar, J. Noble, R. Pang, C. Victora, F. Barros, M. Carvalho, L. Salomon, Z. Bhutta, S. Kennedy, and J. Villar, “International standards for fetal growth based on serial ultrasound measurements: the fetal growth longitudinal study of the INTERGROWTH-21st Project,” *Lancet*, vol. 384, pp. 869–879, 2014, for the International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st).
- [154] L. Salomon, C. Diaz-Garcia, J. Bernard, and Y. Ville, “Reference range for cervical length throughout pregnancy: non-parametric lms-based model applied to a large sample,” *Ultrasound in Obstetrics & Gynecology*, vol. 33, pp. 459–464, 2009.
- [155] R. Bakker, E. Steegers, J. Mackenbach, A. Hofman, and V. Jaddoe, “Maternal smoking and blood pressure in different trimesters of pregnancy: The generation r study,” *Journal of Hypertension*, vol. 28, pp. 2210–2218, 2010.
- [156] M. L. Thompson, M. A. Williams, and R. S. Miller, “A longitudinal study of maternal hemodynamics during normal pregnancy,” *Paediatric and Perinatal Epidemiology*, vol. 23, pp. 253–263, 2009.
- [157] E. Bertino, S. Milani, C. Fabris, and M. D. Curtis, “Neonatal anthropometric charts: what they are, what they are not,” *Archives of Disease in Childhood. Fetal and Neonatal Edition*, vol. 92, pp. F7–F10, 2007.

- [158] L. Tarassenko, D. A. Clifton, M. R. Pinsky, M. T. Hravnak, J. R. Woods, and P. J. Watkinson, "Centile-based early warning scores derived from statistical distributions of vital signs," *Resuscitation*, vol. 82, pp. 1013–1018, 2011.
- [159] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*. Cambridge (MA), USA: MIT Press, 2006.
- [160] T. J. Cole and P. J. Green, "Smoothing reference centile curves: the lms method and penalized likelihood," *Statistics in Medicine*, vol. 11, pp. 1305–1319, 1992.
- [161] D. Wong, "Identifying vital sign abnormality in acutely-ill patients," Ph.D. dissertation, University of Oxford, Oxford, UK, 2011.
- [162] A. Indrayan, "Demystifying lms and bcpe methods of centile estimation for growth curves and other health parameters," *Indian Pediatrics*, vol. 51, pp. 37–43, 2014.
- [163] M. Stasinopoulos, "gamlss: Generalised Additive Models for Location Scale and Shape," <https://cran.r-project.org/web/packages/gamlss/index.html>, 2016, [Online; accessed 21-Jan-2016].
- [164] R. A. Rigby and D. M. Stasinopoulos, "Smooth centile curves for skew and kurtotic data modelled using the box-cox power exponential distribution," *Statistics in Medicine*, vol. 23, pp. 3053–3076, 2004.
- [165] T. Gasser, W. Kohler, H. Muller, A. Kneip, R. Largo, L. Molinari, and A. Prader, "Velocity and acceleration of height growth using kernel estimation," *Annals of Human Biology*, vol. 11, no. 5, pp. 397–411, 1984.
- [166] S. Guo, A. F. Roche, R. N. Baumgartner, W. C. Chumlea, and A. S. Ryan, "Kernel regression for smoothing percentile curves: reference data for calf and scapular skinfold thicknesses in maxican americans," *The American Journal of Clinical Nutrition*, vol. 51, pp. 908S–16S, 1990.
- [167] J. D. Singer and J. B. Willett, *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York, USA: Oxford University Press Inc, 2003.
- [168] P. Royston and D. Altman, "Regression using fractional polynomials of continuous covariates – parsimonious parametric modeling," *Applied Statistics*, vol. 43, pp. 429–467, 1994.
- [169] K. Tilling, C. Macdonald-Wallis, D. A. Lawlor, R. A. Hughes, and L. D. Howe, "Modelling childhood growth using fractional polynomials and linear splines," *Annals of Nutrition & Metabolism*, vol. 65, pp. 129–138, 2014.
- [170] S. van Buuren and M. Fredriks, "Worm plot: a simple diagnostic device for modelling growth reference curves," *Statistics in Medicine*, vol. 20, pp. 1259–1277, 2001.
- [171] M. Wand and M. Jones, *Kernel Smoothing. Monographs on Statistics and Applied Probability*. London: Chapman and Hill, 1995.
- [172] S. Rabe-Hesketh and A. Skrondal, *Multilevel and longitudinal modeling using Stata. Volume I: Continuous responses*, 3rd ed. Texas, USA: Stata Press, 2012.
- [173] A. Pereira and B. P. Krieger, "Pulmonary complications of pregnancy," *Clinics in chest medicine*, vol. 25, no. 2, pp. 299–310, 2004.

- [174] C. Bishop, *Pattern Recognition and Machine Learning*. New York, USA: Springer Science+Business Media LLC, 2006.
- [175] R. Gaillard, R. Bakker, S. P. Willemsen, A. Hofman, E. A. Steegers, and V. W. Jaddoe, "Blood pressure tracking during pregnancy and the risk of gestational hypertensive disorders: The generation r study," *European Heart Journal*, vol. 32, pp. 3088–3097, 2011.
- [176] F. F. Lauszus, A. Rosgaard, T. Lousen, O. W. Rasmussen, T. M. Klebe, and J. G. Klebe, "Precision, consistency, and reproducibility of blood pressure in diabetic and non-diabetic pregnancy: the appraisal of repeated measurements," *Acta Obstetrica et Gynecologica Scandinavica*, vol. 86, no. 9, pp. 1063–70, 2007.
- [177] E. W. Steyerberg, K. G. M. Moons, D. A. van der Windt, J. A. Hayden, P. Perel, S. Schroter, R. D. Riley, H. Hemingway, and D. G. Altman, "Prognosis research strategy (PROGRESS) 3: Prognostic model research," *PLoS Med*, vol. 10, no. 2, p. e1001381, 2013, for the PROGRESS Group.
- [178] H. Strevens, D. Wide-Swensson, and I. Ingemarsson, "Blood pressure during pregnancy in a swedish population; impact of parity," *Acta Obstetrica et Gynecologica Scandinavica*, vol. 80, pp. 824–829, 2001.
- [179] C. Matkin, J. Britton, S. Samuels, and B. Eskenazi, "Smoking and blood pressure patterns in normotensive pregnant women," *Paediatric and Perinatal Epidemiology*, vol. 13, no. 1, pp. 22–34, 1999.
- [180] F. E. Okonofua, J. A. Balogun, N. A. Amienheme, and S. P. O'Brien, "Blood pressure changes during pregnancy in nigerian women," *International Journal of Cardiology*, vol. 37, no. 3, pp. 373–9, 1992.
- [181] G. B. Smith, D. R. Prytherch, P. E. Schmidt, P. I. Featherstone, and B. Higgins, "Review and performance evaluation of single-parameter 'track and trigger' systems," *Resuscitation*, vol. 79, no. 1, pp. 11–21, 2008.
- [182] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2006.
- [183] Obstetric Anaesthetists' Association, "MEOWS," <http://www.oaa-anaes.ac.uk/UI/Content/Content.aspx?ID=186>, 2009, [Online; accessed 21-Jan-2016].
- [184] P. J. Maguire, A. C. O'Higgins, K. A. Power, N. Daly, A. McKeating, and M. J. Turner, "Maternal bacteremia and the irish maternity early warning system," *International Journal of Gynecology and Obstetrics*, vol. 129, pp. 142–145, 2015.
- [185] D. M. Austin, L. Sadler, C. McLintock, C. McArthur, V. Masson, C. Farquhar, and S. Rhodes, "Early detection of severe maternal morbidity: A retrospective assessment of the role of an early warning score system," *Australian and New Zealand Journal of Obstetrics and Gynaecology*, vol. 54, pp. 152–155, 2014.
- [186] G. B. Smith and D. R. Prytherch, "Obstetric early warning scores: much more work required," *Anaesthesia*, vol. 68, pp. 773–778, 2013.
- [187] E. O'Brien, R. Asmar, L. Beilin, Y. Imai, J. M. Mallion, G. Mancina, T. Mengden, M. Myers, P. Padfield, P. Palatini, G. Parati, T. Pickering, J. Redon, J. Staessen, G. Stergiou, and P. Verdecchia, "European society of hypertension recommendations for conventional,

- ambulatory and home blood pressure measurement,” *Journal of Hypertension*, vol. 21, no. 5, pp. 821–848, 2003, on behalf of the European Society of Hypertension Working Group on Blood Pressure Monitoring.
- [188] C. P. Subbe, “Centile-based early warning scores derived from statistical distributions of vital signs,” *Resuscitation*, vol. 82, pp. 969–970, 2011.
 - [189] C. Macdonald-Wallis, R. J. Silverwood, B. L. de Stavola, H. Inskip, C. Cooper, K. M. Godfrey, S. Crozier, A. Fraser, S. M. Nelson, D. A. Lawlor, and K. Tilling, “Antenatal blood pressure for prediction of pre-eclampsia, preterm birth, and small for gestational age babies: development and validation in two general population cohorts,” *BMJ*, vol. 351, 2015.
 - [190] C. M. Albright, T. N. Ali, V. Lopes, D. J. Rouse, and B. L. Anderson, “The Sepsis in Obstetrics Score: a model to identify risk of morbidity from sepsis in pregnancy,” *American Journal of Obstetrics and Gynecology*, vol. 211, no. 39, pp. e1–8, 2014.
 - [191] P. Schulam, F. Wigley, and S. Saria, “Clustering longitudinal clinical marker trajectories from electronic health data: Applications to phenotyping and endotype discovery,” 2015, [Online; accessed 21-Jan-2016].
 - [192] P. Schulam and S. Saria, “A framework for individualizing predictions of disease trajectories by exploiting multi-resolution structure,” in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 748–756.
 - [193] C. Proust-Lima and J. M. G. Taylor, “Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach,” *Biostatistics*, vol. 10, no. 3, pp. 535–549, 2009.
 - [194] N. R. Cook, “Use and misuse of the Receiver Operating Characteristic curve in risk prediction,” *Circulation*, vol. 115, no. 7, pp. 928–935, 2007.
 - [195] R. Pullinger, S. Wilson, R. Way, M. Santos, D. Wong, D. Clifton, J. Birks, and L. Tarassenko, “Implementing an electronic observation and early warning score chart in the emergency department: a feasibility study,” *European Journal of Emergency Medicine*, 2016, [in press].
 - [196] N. Mackintosh, K. Watson, S. Rance, and J. Sandall, “Value of a modified early obstetric warning system (MEOWS) in managing maternal complications in the peripartum period: an ethnographic study,” *BMJ Quality & Safety*, vol. 23, no. 1, pp. 26–34, 2014.
 - [197] S. J. Wilson, D. Wong, D. Clifton, S. Fleming, R. Way, R. Pullinger, and L. Tarassenko, “Track and trigger in an emergency department: an observational evaluation study,” *Emergency Medicine Journal*, vol. 30, no. 3, pp. 186–191, 2013.
 - [198] N. M. Laird and J. H. Ware, “Random-effects models for longitudinal data,” *Biometrics*, vol. 38, no. 4, pp. 963–974, 1982.
 - [199] J. Pinheiro and D. Bates, *Mixed-Effects Models in S and S-PLUS*, ser. Statistics and Computing. New York, USA: Springer New York, 2000.
 - [200] D. A. Harville, “Maximum likelihood approaches to variance component estimation and to related problems,” *Journal of the American Statistical Association*, vol. 72, no. 358, pp. 320–338, 1977.

- [201] H. Wu and J.-T. Zhang, *Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed-Effects Modeling Approaches*. New Jersey, USA: John Wiley & Sons, Inc, 2006.
- [202] Y. Sakamoto, M. Ishiguro, and G. Kitagawa, *Akaike information criterion statistics*. MI, USA: KTK Scientific Publishers, 1986.
- [203] G. Schwarz, “Multilevel mixed linear model analysis using iterative generalized least squares,” *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.
- [204] C. Rasmussen, “Documentation for gpml matlab code version 3.6,” <http://www.gaussianprocess.org/gpml/code/matlab/doc/>, 2015, [Online; accessed 4-Dec-2015].
- [205] M. Pimentel, M. Santos, M. Maraci, C. Arteta, J. Domingos, D. Clifton, and G. Clifford, “A \$5 smart blood pressure system,” in *Appropriate Healthcare Technologies for Low Resource Settings (AHT 2014)*, Sept 2014, pp. 1–4.