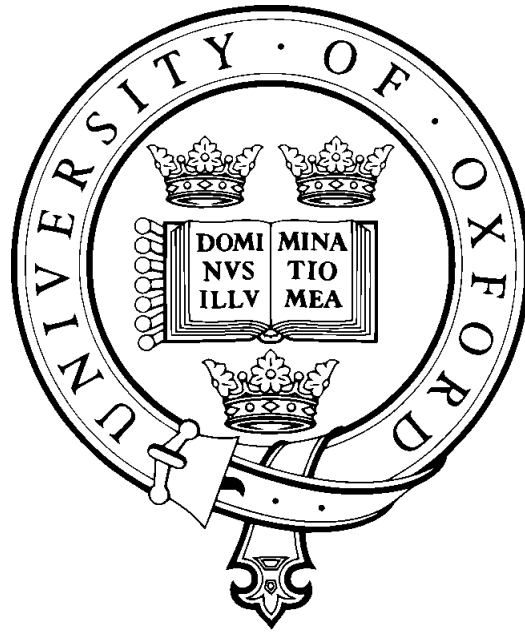


Monitoring and analysis of antenatal and postnatal changes in maternal vital signs



Rebecca M. Pullon

St Hugh's College

Supervised by

Professor Lionel Tarassenko

Submitted: Michaelmas Term 2016

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

Monitoring and analysis of antenatal and postnatal changes in maternal vital signs

Rebecca M. Pullon

Thesis submitted for the degree of Doctor of Philosophy

St Hugh's College

Michaelmas Term 2016

Abstract

Pregnancy-related complications affect approximately 15% of pregnancies and, if severe, can have long-term consequences. Timely recognition of physiological deterioration is known to reduce the prevalence and severity of complications. However pregnancy-associated changes in vital signs (blood pressure, heart rate, temperature, oxygen saturation, and respiratory rate) complicate the detection of abnormal physiology, and these changes are not well documented. This thesis describes the development of algorithms to ensure the collection of good-quality vital-sign data during the antenatal and postnatal stages of pregnancy, and the design of an evidence-based obstetric early warning score.

Vital-sign information from 1,000 pregnant women during pregnancy, labour, and after delivery was collected during the 4P study using pulse oximetry, oscillometry for blood pressure measurement and a tympanic thermometer. Dynamic time warping was used to assess beat-by-beat quality in the photoplethysmograph (PPG) waveform obtained from the pulse oximeter. The resulting signal quality index enabled the exclusion of poor-quality sections and their associated measurements of heart rate and peripheral oxygen saturation. A robust measurement of respiratory rate was obtained by combining information from the PPG waveform, and accelerometer and gyroscope waveforms from a smartphone. After processing, frequency-based techniques, such as Fourier analysis and auto-regressive modelling, and time-domain peak detection were fused to estimate respiratory rate. When compared with the reference respiratory rate obtained from midwife measurement, the lowest mean absolute error of 1.16 breaths per minute was obtained from respiratory rate estimates from the y -axis of the gyroscope.

Antenatal and postnatal reference ranges for each vital sign were developed with a standard polynomial multilevel (hierarchical) model using 10,000 vital sign measurements from 620 healthy women in the 4P study. Vital-sign trajectories confirmed known trends of blood pressure and heart rate changes during pregnancy, and provided new information about other vital-sign trends. Additional covariates were included to investigate the effect of parity and body mass index (BMI) on vital-sign trends. The outer centiles of the vital sign reference ranges were used to design an obstetric early warning score (C-ObsEWS) that took into account gestational age or time after delivery.

The investigations in this thesis contribute additional knowledge of pregnancy-associated vital-sign changes, and lead to an initial proposal for an evidence-based obstetric early warning score specific to the stage of pregnancy.

Acknowledgements

As with any piece of work, this thesis would not be the quality that it is without the input from and support of many colleagues, friends, and family. When I started out on the DPhil journey I was naive at what the process involved. Along the way I've had many rough patches, but I have learnt and grown so much that it is a different person leaving Oxford.

My most indebted thanks goes to my supervisor, Lionel Tarassenko. When I chose to join the 4P study I didn't know that I had also chosen a supervisor with much experience, wisdom and sensitivity. Thank you for providing me with enough direction that I (mostly) knew what I was doing, but pushed me to grow and learn things for myself along the way. Thank you for taking my suggestions seriously, and for being willing to read my work (sometimes overnight!) and provide thoughtful feedback.

Peter Watkinson, your help has been invaluable. Thank you for your vision and passion for clinical research and for encouraging me to have confidence in myself. You have always given me the time that I needed, now that I know to just turn up at your office instead of waiting for an email response. I have enjoyed our banter, and dreaming about what better clinical monitoring could look like.

Thank you to the rest of the 4P study team: Lucy MacKillop, thank you for patiently explaining the nuances of obstetric care and physiology to me over and over again; Jude Kemp, Clare Edwards, Fiona Kumar, and Fenella Roseman, you are the best research midwives I could ever hope to work with. I enjoyed popping into your office for tea, thinking up ways to improve the study, and working together to iron out all the mistakes in the 4P database. We got there in the end!

My thanks also goes to everyone at the IBME who I worked alongside day to day for the last four years. Kate Niehaus, you have been my partner in crime, my confident and companion. Thank you for being in the lab every day to keep me company, and for setting such a good example of striving to do the best you can. I have enjoyed our trips together, running alongside (behind) you in the London marathon, and going to Iceland just so we could focus on our thesis writing. Lise Loerup, you paved the way for me in the 4P study. I have so appreciated having you as a project partner; it has made such a difference being able to figure out things together with you. Thank you for your initiative. To Time Series Terrace, my row of adventure buddies, you have kept me going the last four years. Thank you for facilitating numerous work-breaks, great lunchtime chats, and for generally being great listening ears for work and non-work topics alike. I have learnt so much from you all. Dario Salvi and the rest of the postdoc office, thank you for taking me under your wing and helping me organise a clinical trial! I often came

with urgent requests and you always helped me figure out a solution or let me raid the postdoc equipment stores.

A number of other people have also journeyed with me. Simon Kerr, your constant support and encouragement has been invaluable. Thank you for putting up with my thesis frustrations and emotional struggles, you always knew how to cheer me up. Dave Wong, thank you for helping me to figure out the maths when I got stuck, and to have the confidence to try new things when I wanted to give up. Yours and Claires home was a source of refuge for me in my first two years. Christopher Vogel, thank you for understanding my questions about work and thesis writing even when I didnt know how to ask them, and for offering to proof-read almost everything I wrote! A huge thanks for the residents of 33 Argyle St for putting up with my disillusioned-with-DPhil self, my thesis-writing self, and then for proof-reading my entire thesis the night before I sent it to printers. I hope my improvements in grammar have made you proud! Thank you to my friends at Oxford Community Church for doing life with me. To all my rowing peers at St Hughs Boat Club, thank you for providing me with a way to push my body when my mind needed a break.

Finally, I would not be here if it were not for the support of my family. My brothers, Andrew and Jonathan, although I often feel a world away from you, I am so pleased that we are family and I look forward to living closer to you again. My sister-in-law Candra, I appreciate your thoughtfulness and your determination to foster a friendship across the world. I cant wait to show you Oxford. To my parents, you have been my biggest supporters. Thank you mum for being on the other end of the phone more times than I can count. Thank you dad for your wisdom and for reading parts of my thesis at the last minute when I needed emergency clinical input.

Thank you Lord God for your unconditional love and grace. The truth about who I am to you is the reason I persevered and still have hope in the future.

*“For the LORD is **good** and His love endures forever”*

Psalm 100:5

*“Give thanks to the LORD for He is **good**”*

Psalm 136:1

*“Taste and see that the LORD is **good**”*

Psalm 34:8

Contents

1	Introduction	1
2	Maternal health and vital sign monitoring	5
2.1	Maternal health and care in the UK	5
2.1.1	Causes of maternal mortality	7
2.1.2	The burden of pregnancy-related complications	7
2.2	Identifying physiological deterioration	12
2.2.1	Validation of track-and-trigger systems	13
2.2.2	A standardised early warning score for the general adult population	16
2.3	Introduction to vital signs	18
2.3.1	Vital signs and their measurement	18
2.3.2	Blood pressure and heart rate in the antenatal period: A system- atic review	21
2.3.3	Other vital sign changes	28
2.4	Conclusion	30
3	Description of dataset	31
3.1	The 4P study	31
3.1.1	Study population	32
3.1.2	Timeline for data collection	32
3.1.3	Vital sign data	33
3.1.4	Labour and delivery information	33
3.1.5	Other information collected	34
3.2	The 4P system	36
3.2.1	Equipment for vital sign collection	37
3.2.2	Tablet application for vital sign data collection	39
3.2.3	Website system	41
3.2.4	Data transmission and storage	43
3.3	Data used in this thesis	44
3.3.1	Study population	44
3.3.2	Pregnancy complications and delivery information	45
3.3.3	Vital sign measurements	47
3.4	Quality of the pulse oximeter recordings	52
3.4.1	Pulse oximeter hardware	53
3.4.2	Previous methods for artefact suppression	55
3.4.3	Signal quality index for PPG	55

3.4.4	SQI applied to 4P data	57
3.5	Conclusion	61
4	Respiratory rate estimation	62
4.1	Mechanics of breathing	63
4.1.1	Movement detection	63
4.1.2	Heart rate variability	63
4.2	Related work	65
4.2.1	Respiratory rate from accelerometers	65
4.2.2	Respiratory rate from the pulse oximeter	67
4.2.3	Fusing respiratory rate estimates	69
4.2.4	Respiratory rate in pregnancy	71
4.3	Dataset	72
4.3.1	Reference respiratory rate	73
4.4	Respiratory rate algorithms	78
4.4.1	The accelerometer and gyroscope waveforms	78
4.4.2	The PPG waveform	81
4.4.3	Respiratory Quality Index	83
4.4.4	Principal component analysis	86
4.4.5	Respiratory rate calculation	88
4.4.6	Evaluation of methods	90
4.4.7	Summary	92
4.5	Results	94
4.5.1	Respiratory quality	99
4.5.2	Comparison of devices	105
4.6	Discussion	106
5	Modelling vital sign trends	114
5.1	Reference ranges in obstetric care	114
5.2	Univariate model building	120
5.2.1	Dataset	120
5.2.2	Description of multilevel models	120
5.2.3	Functional forms of time considered	123
5.2.4	Methods for prediction	126
5.2.5	Assessing model fit	126
5.3	Population trends	127
5.3.1	Significance of results	131
5.4	Subject-specific trends	134
5.4.1	Subgroup investigation	135
5.5	Inclusion of additional covariates	139
5.5.1	Parity	140
5.5.2	BMI	144
5.6	Discussion	147

6	Development of an Obstetric Early Warning Score	150
6.1	Alerting systems in obstetrics	150
6.1.1	Identification of obstetric EWS in the literature	152
6.1.2	Methods of design	156
6.1.3	Methods of validation	158
6.1.4	Future direction for an obstetric early warning score	162
6.2	Centile-based early warning scores	164
6.2.1	Scoring bands of a centile-based ObsEWS	165
6.2.2	Re-visiting the 4P data	169
6.3	Women who developed sepsis	172
6.4	Discussion	178
7	Multivariate systems	181
7.1	The contribution of multivariate models	181
7.2	Multivariate multilevel models	182
7.2.1	Dataset	182
7.2.2	Description of the multivariate model	183
7.2.3	Functional form of time	184
7.3	Vital sign correlations	186
7.4	A continuous measure of abnormality	188
7.5	Discussion	192
8	Conclusions and future work	194
8.1	Summary of thesis	194
8.2	Future work	198
8.2.1	Validation of an obstetric early warning score	198
8.2.2	Further development of an early warning score	200
8.2.3	Monitoring during labour	201
8.3	Final remarks	201
	Appendices	202
A	Definition of healthy	202
B	Supplementary material: Pulse oximeter quality	203
B.1	Peak detection algorithm	203
B.2	SQI(Flags)	205
B.3	SQI(DTW)	206
B.4	Dynamic Time Warping (DTW)	210
C	Supplementary material: Multilevel models	212
C.1	Bayes rule and prior distributions	213
C.2	Inference using sampling methods	214
C.3	Stan software	216
C.4	Assessing convergence	217
C.5	WAIC Information Criterion	218
C.6	Multilevel models for aggregate data	219
C.7	Example of multilevel model development	220

C.7.1	Model building	220
C.7.2	Model comparison	224
C.7.3	Model selection	228
C.8	Equations for final univariate vital sign trends	230
References		233

Introduction

It was not long ago when having a baby was one of the most dangerous things a woman could do. Maternal health has been a global focus for the last 20 years, from which there have been huge worldwide reductions in the number of deaths occurring during pregnancy or in the six weeks after delivery (maternal mortality). The United Nation's 5th Millennium Development Goal aimed to reduce the worldwide maternal mortality ratio (the number of maternal deaths per live births) by 75% from the year 1990 to 2015. This ambitious goal was not met, partly hindered by lack of structured healthcare services in developing countries [1]. Yet the maternal mortality ratio still reduced globally by 44% from 1990 to 2015, with huge leaps in Southern Asia (a reduction of 64%) and sub-Saharan Africa (49%) [2]. In most developed countries, the rate of maternal mortality has now fallen to around 20 in 100,000 pregnancies [3].

Despite the reduction in maternal mortality, many women still experience complications during pregnancy which can endanger the health of the baby, and have long-term consequences for the mother. Pregnancy-related complications are estimated to affect around 15% of pregnancies globally, and include severe infection, loss of blood, high blood pressure, and gestational diabetes [4]. In developed countries, the risk of a woman developing a pregnancy complication is increasing due to demographic factors, such as increased age and weight at the time of pregnancy, and an increased prevalence of

chronic diseases, such as heart disease [5]. Women who are at risk of developing a pregnancy-related complication need access to good quality maternity services that can act promptly in the case of an emergency. A delay in diagnosis, incomplete assessment, and failure to detect deterioration are common reasons for maternal complications to become severe [6]. Current hospital systems to detect physiological deterioration for obstetric (pregnant or recently pregnant) populations vary by region and are usually only informed by local expert opinion. To further reduce the burden of health problems from pregnancy, the maternal health focus needs to include pregnancy-related complications. The focus of this thesis is the creation of an evidence-based alerting system for the monitoring of obstetric inpatients.

In order to create an alerting system, one needs to decide which physiological parameters to monitor, and have knowledge of their normal ranges during pregnancy. Guidelines in the UK suggest five vital signs for routine monitoring: blood pressure, heart rate, oxygen saturation, temperature and respiratory rate [7]. These measurements can be quickly obtained from non-invasive devices, thus are suitable for bedside monitoring and also for inclusion in an alerting system. A review of the current literature of vital sign trends in pregnancy reveals that pregnancy physiology remains largely unknown. Observational studies of blood pressure and heart rate during pregnancy have been regularly performed, but these are often small local studies and do not provide a consensus on population trends. Little information for the remaining vital signs exists, particularly in the period immediately after delivery.

The gap in knowledge regarding vital sign changes in pregnancy led to the design of the 4P (Pregnancy Physiology Pattern Prediction) study, a multi-centre, longitudinal observational study of 1,000 women experiencing a normal pregnancy. The study began in Oxford in 2012 and the last woman was recruited in December 2016. Data cleaning and analysis is expected to be completed by July 2017. Vital-sign data from the 4P study

(and demographic information for the women taking part) will form a large database of physiological data during pregnancy, delivery, and the two weeks after delivery.

Good-quality data is essential. The vital-sign measurement system in the 4P study relied on equipment regularly used in the maternity hospital to reflect the quality of current clinical practice. In addition, mobile technology was employed to enable remote monitoring by the woman at home, and midwife review from the clinic. Importantly, a robust, automatic measure of respiratory rate is required. Current clinical practice is midwife observation of chest wall movement for 15 or 30 seconds which does not provide adequate resolution of the number of breaths per minute [8]. Respiratory rate is the only vital sign not routinely measured with an electronic device. The acquisition of good-quality vital-sign data including accurate values of respiratory rate during pregnancy and after delivery will enable us to design an evidence-based obstetric alerting system as described in the later chapters of this thesis.

Overview of thesis:

The structure of this thesis is as follows. In Chapter 2 we provide background on maternity health and care in the UK, introduce the reader to the concept of an early warning score (EWS), and review the literature of known vital sign changes in pregnancy. In Chapter 3 we introduce the 4P study and the dataset used in this thesis. We also develop quality metrics for the pulse oximeter device to ensure good quality data. In Chapter 4 we develop a robust method for the measurement of respiratory rate using information from the pulse oximeter, and the in-built accelerometer and gyroscope in a mobile phone. In Chapter 5 we create reference ranges for each vital sign trajectory during pregnancy and in the two weeks after delivery using multilevel models. These models take into account the hierarchical structure of longitudinal data and allow the analysis of women's individual vital sign trends. In Chapter 6 we translate the vital-sign reference ranges into an obstetric alerting system that can be used in clinical care.

In Chapter 7 we explore the correlation between vital signs in a multivariate model, also based on the multilevel hierarchical structure. We develop a continuous measure of physiological abnormality and compare the results with the obstetric alerting system. Finally, in Chapter 8 we conclude and provide directions for future work.

Maternal health and vital sign monitoring

2.1 Maternal health and care in the UK

The most common reason for admission to hospital in England is to have a baby. In 2015 there were 697,852 live births in England and Wales, each of which cost the NHS¹ around £3,700 [9]. Pregnant women receive care from a range of healthcare professionals at multiple time points during pregnancy and after the baby is born (Figure 2.1).

The antenatal (or antepartum) period is the time from conception to delivery, or the period of fetal *gestation*. The stage of gestation is measured in weeks^{+days}, and lasts an average of 40⁺⁰ weeks. Expected delivery date (EDD) is calculated from crown-rump measurement (by ultrasound) between 10⁺⁰ and 13⁺⁶ [10]. The first point of contact for a pregnant woman with a healthcare professional is usually her GP at some time between 7 and 9 weeks gestation. She is then transferred to community midwifery care where she will receive around 7 home visits during her pregnancy. Ultrasound scans are offered at around 12 and 20 weeks gestation. Antenatal care is designed to be woman-centred, empowering the mother to make informed choices during her pregnancy [11].

The postnatal (or postpartum) period, also called the puerperium, is medically recog-

¹The national public health service in the UK is called the National Health Service (NHS), founded in 1948.

nised as a time of six weeks (42 days) after childbirth when the mother's organs return to their original non-pregnant position. The present day brings modern medical institutional systems that have largely replaced family, community, and religious experts in providing care for the mother during this time [12]. Most women in the UK give birth in hospitals or maternity centres (87% in 2012 [9]) where they spend just a couple of days before returning home. Postnatal care is primarily provided by the midwife, and includes routine clinical examination and observation of the woman and her baby, routine infant screening and ongoing information and support, most of which takes place in the woman's home [13].

Present-day guidelines for good maternity care are set by the National Institute of Care and Excellence (NICE) and the Royal College of Obstetricians and Gynaecologists (RCOG). These are often informed by international guidelines published by the World Health Organisation (WHO). *Obstetrics* is the field of study relating to pregnancy, childbirth and the six weeks after delivery.

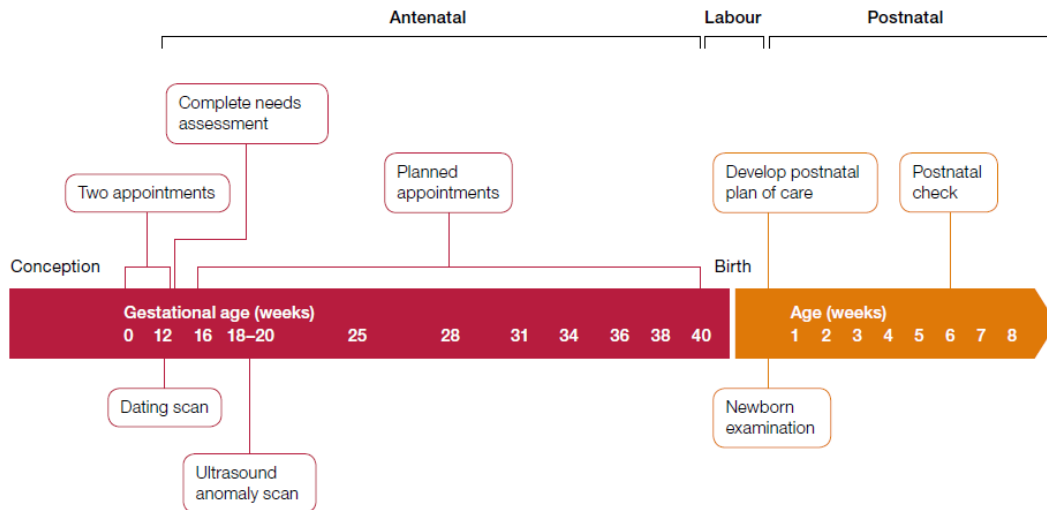


Figure 2.1: Summary of maternal care pathway in the UK. Reproduced from [9].

2.1.1 Causes of maternal mortality

Every triennium, the Confidential Enquiry into Maternal Deaths in the UK (CEMD) takes place. Started in 1952, the CEMD is the longest running maternal audit programme in the world, and has seen a reduction in rate of maternal mortality² from 90 per 100,000 maternities (1952-1954) to 8.54 (2012-2014) [14, 15]. The Enquiry is now run by MBRRACE-UK³ (as of 2014) and is updated every year. The latest CEMD (published in December 2016) covered the triennium 2012-2014 [14].

Maternal deaths can be *directly* related to obstetric complications of the pregnant state, or from an *indirect* cause such as a previous existing disease [3]. The main causes of maternal death in the last triennium were cardiac disease, sepsis and neurological conditions (Figure 2.2). 24% of the deaths occurred in the 24 hours after delivery, and 51% in the 6 weeks after delivery. The postpartum period is most dangerous for maternal health and yet little is known about the expected physiological changes during this time.

2.1.2 The burden of pregnancy-related complications

While the rate of maternal mortality in the UK and most other developed countries is now low, the burden of effects from pregnancy-related complications (morbidity) remains high [4, 16]. It is difficult to estimate the global impact of maternal morbidity, primarily because it is difficult to define. In 2011, the WHO attempted to define cases of severe maternal morbidity, labelling them “near-miss” deaths. Then in 2016, a WHO working group developed a wider definition of maternal morbidity to encompass those less severe cases [17]. Maternal morbidity is now globally defined as “*any health condition attributed to and/or complicating pregnancy and childbirth that has a negative impact*”

²Maternal mortality is the death of a woman while pregnant or within 42 days of the end of pregnancy from any cause related to or aggravated by the pregnancy or its management, but not from accidental or incidental causes.

³MBRRACE-UK is a national collaborative programme which involves the surveillance and investigation of maternal deaths, stillbirths and infant deaths.

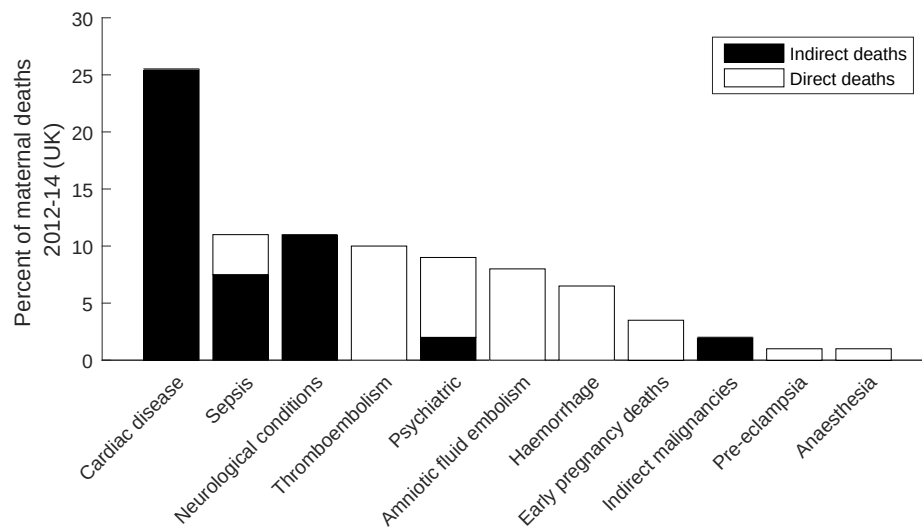


Figure 2.2: Direct and indirect causes of maternal deaths in the UK for 2012-2014. Data from [14]. Genital tract sepsis is the only form of sepsis considered a direct maternal death. Maternal deaths by suicide (a psychiatric cause) are now considered a direct maternal death (as of 2016).

on the woman’s wellbeing and/or functioning”. Along with this definition, the working group produced a framework through which morbidities and their associated symptoms and treatments could be recorded so that the impact both to the woman and the health system could be understood.

The long-term effects of pregnancy-related complications are of growing concern. Poor management of complications can increase severity and lead to chronic conditions. Obstetric complications are also associated with an increased risk of postpartum depression [18]. It is therefore a global interest to reduce both maternal mortality and morbidity, and reduce the severity of complications.

In the UK, the number of maternal morbidities remains high primarily as a result of demographic trends. These include an increasing maternal age, increasing rate of obesity (measured by Body Mass Index, BMI⁴), and a shift towards a more ethnically diverse maternity population [5]. A quarter of live births in the UK in 2015 were to non-UK born mothers [16]. Furthermore, there has been an increase in the prevalence

⁴BMI is weight divided by height squared, measured in kgm^{-2} .

of chronic conditions, such as cardiac disease and diabetes, which are associated with higher rates of maternal mortality and morbidity.

Women deemed to have a high-risk of maternal morbidity require greater clinical involvement. While low-risk women are typically cared for by their GP and local midwifery team, women with complex pregnancies should receive more frequent care from a multidisciplinary team [6]. In addition to a midwife, such teams may include an obstetrician, obstetric anaesthetist, surgeon, and other healthcare professionals as required (e.g. physiotherapists and nutritionists). If a woman develops a severe pregnancy complication, she will likely require admission to an obstetric intensive care unit (ICU).

Below we present definitions for three of the most common pregnancy-related complications relevant to the research described in this thesis: sepsis, postpartum haemorrhage, and hypertension [19, 20, 21].

Sepsis

Sepsis is a severe infection which can lead to physiological deterioration, and ultimately death. During the 19th century, maternal sepsis took on epidemic proportions, particularly in centres where women stayed for several days after delivery [22]. Rates declined with the introduction of antibiotics, yet it is still a leading cause of maternal death in the UK (although only genital tract sepsis is considered a direct maternal death). Sepsis is of particular concern in the postpartum period accounting for 82% of deaths from maternal sepsis in 2009-2011 in the UK [6]. It is thought that for every death, approximately 50 women have life-threatening morbidity from sepsis [23]. The most significant long-term effect of genital tract sepsis is infertility, estimated to affect 450,000 women globally each year [22].

The definition of sepsis was standardised in 1992 [24]. Sepsis is now described as a continuum in four stages: Systemic Inflammatory Response Syndrome (SIRS), sepsis, severe sepsis and septic shock.

- **SIRS** criteria are defined as two of: high or low temperature, high respiratory rate, high heart rate, or a high white blood cell count.
- **Sepsis** is the presence of an infection with SIRS.
- **Severe sepsis** is sepsis associated with organ dysfunction or low blood pressure (hypotension).
- **Septic shock** is sepsis associated with persistent hypotension that does not recover with fluid resuscitation.

A study of all severe sepsis incidents in the UK found that severe sepsis progressed very quickly from initial signs of SIRS; in 80% of cases it was less than 24 hours [23].

Postpartum haemorrhage

A postpartum haemorrhage (PPH) is excessive bleeding after delivery. It is the most common type of major obstetric haemorrhage and is the leading cause of worldwide maternal mortality [25]. Mortality due to PPH disproportionately affects developing regions (99.7% of cases), however the morbidity associated with severe haemorrhage remains a major problem worldwide [6]. Compared to a woman with no complications, a woman with severe PPH is more likely to develop acute renal failure, postpartum sepsis, severe anaemia or be admitted to ICU [26, 22]. In England, PPH affected 13% of all maternities from 2011-2012 [6], and severe PPH is estimated to affect 1-2% of worldwide maternities [27]. Both in the UK and internationally the incidence of recorded PPH is increasing.

The RCOG define PPH as a blood loss of 500ml or more from the genital tract after delivery of the baby. If delivery was by caesarean-section⁵, the threshold for PPH is 1000ml. Severe PPH is more than 2000ml of blood loss after delivery. The term “primary” is used to describe PPH that occurs within 24 hours of delivery, and “secondary” is a haemorrhage occurring between 24 hours and 12 weeks after delivery [28, 29].

⁵A caesarean-section is where the baby is extracted by an operation, rather than passing through the birth canal.

Gestational hypertension and pre-eclampsia

Hypertension means elevated blood pressure, indicated by either a high systolic blood pressure or a high diastolic blood pressure⁶. Gestational hypertension is the new onset of hypertension in pregnancy (after 20 weeks gestation), and affects approximately 10% of pregnancies [30]. Women who are nulliparous⁷, aged 40 years or older, have a BMI of more than 35 kgm⁻², or carrying a multiple pregnancy are at most risk of developing gestational hypertension [31].

In 2010, NICE set out guidelines to define and manage hypertension in pregnancy. Mild hypertension is diagnosed at a blood pressure higher than 140/90mmHg (systolic/-diastolic). Treatment is initiated at 150/100mmHg (or 140/90mmHg with organ failure) and severe hypertension is diagnosed at 160/110mmHg. Pre-eclampsia is gestational hypertension with the additional presence of significant proteinuria⁸. Severe pre-eclampsia is pre-eclampsia with severe hypertension or biochemical or haematological impairment [31]. Gestational hypertension, pre-eclampsia, and severe pre-eclampsia, are collectively referred to as the Hypertensive Disorders of Pregnancy (HDP).

Gestational hypertension can be managed with routine monitoring and anti-hypertensive medication. However, once proteinuria develops (i.e. pre-eclampsia), the outcome for the mother deteriorates. Pre-eclampsia can lead to renal failure, stroke, cardiac arrest, liver failure and eclampsia (seizures) [32]. Long-term consequences for women with HDP are also concerning, such as an increased risk of chronic hypertension and cardiovascular disease [31]. Good blood pressure management in the postpartum period to reduce blood pressure to pre-pregnancy levels is essential to reduce the prevalence of long-term effects [33].

⁶The terms systolic blood pressure and diastolic blood pressure are detailed in Section 2.3.1.

⁷Parity is the number of previous births at $\geq 20^{+0}$ weeks gestation. Nulliparous refers to women with a parity of 0, multiparous refers to women with a parity ≥ 1 .

⁸Protein in the urine. A significant amount is more than 300mg recorded in 24 hours.

2.2 Identifying physiological deterioration

Maternal morbidity, like maternal mortality, is associated with a high rate of preventability. With timely identification and intervention, the prevalence of pregnancy-related complications, and their severity, can be reduced.

It is well known that physiological deterioration precedes critical illness [34, 35]. In the previous section we presented three of the most common pregnancy-related complications: maternal sepsis, postpartum haemorrhage, and hypertension. Vital sign deterioration is an indication for all three of these complications. Since 2007, CEMD reports have recommended routine monitoring of the five vital signs advocated by NICE: blood pressure, heart rate, temperature, oxygen saturation and respiratory rate⁹ [36, 7]. The system for detecting deterioration must be coupled with a system for escalation of care to deal with in-hospital emergencies. Rapid Response Systems are hospital-wide systems that combine the detection and then treatment of patients in crisis, before adverse consequences arise.

To improve early recognition of unexpected deterioration, track-and-trigger warning systems were introduced in general wards (non-ICU wards). Their purpose is to secure the timely presence of skilled clinical help for those patients with deteriorating physiology [37]. Track-and-trigger systems measure and record a physiological parameter (track) and initiate an appropriate response if a pre-determined threshold is reached (the trigger). Gao *et al.* [38] performed a systematic review of all track-and-trigger systems published in the literature up until 2007, and identified two main groups of criteria. *Single-parameter systems* consist of a single cut-off value for each parameter, such that a response is initiated for any one abnormal measurement [39]. These systems are used by Medical Emergency Teams, common in Australia and the USA. *Aggregate*

⁹The measurement of vital signs and their expected changes during pregnancy will be introduced in Section 2.3.

scoring systems assign weights to each physiological variable depending on its degree of abnormality. The weights for each parameter are then summed to produce an aggregate score which is compared with a trigger threshold for initiating a clinical response [40]. A variation of the aggregate scoring system, where a response is also triggered if an individual parameter is scored at the highest level, is sometimes called a combination system.

Single-parameter systems trigger a single-response strategy, while aggregate scoring systems produce a “score” which allows for a graded response strategy. Aggregate scoring systems are based on the evidence that simultaneous disturbances in multiple physiological parameters tend to be more commonly observed preceding acute illnesses than a significant disturbance of a single physiological parameter in isolation [40]. The most common form of track-and-trigger system used in the UK is the Early Warning Score (EWS), an aggregate scoring system [38].

Track-and-trigger systems have been adapted for different settings, such as paediatrics and maternity, and for different hospitals [41]. Adaptions of thresholds are usually driven by expert clinical opinion, and considerable variation exists between the types of systems in use. This variation indicates the need for data-driven approaches to develop and validate standardised systems. The UK has recently developed and implemented a standardised track-and-trigger system for the general adult population (NEWS, see Section 2.2.2).

2.2.1 Validation of track-and-trigger systems

For track-and-trigger systems to be an effective and useful clinical tool, they must direct constrained clinical resources to those patients who will benefit most from immediate intervention. Validation usually involves assessing whether a track-and-trigger system can predict which patients, if no intervention is performed, would be likely to experience an undesired outcome. In these studies, the track-and-trigger system is treated as a

binary classifier where the outcome is positive (adverse event) or negative (absence of an adverse event). Clinical outcomes commonly assessed are: death within a given period (usually 24 hours) from the vital sign measurements, admission to ICU, or development of a specific disease or complication. In obstetrics the rate of maternal mortality is low, so an appropriate measure could be admission to ICU which would identify most pregnancy-related complications. The validation process requires a large dataset with correct outcome labels.

The metrics commonly used to evaluate a binary classifier, such as a track-and-trigger system, are described by the confusion matrix in Table 2.1. The predicted outcome (whether or not the track-and-trigger system triggered a response) is compared with the true outcome, the adverse event. If the adverse event was correctly identified it is a true positive (TP), if the absence of an event was correctly identified it is a true negative (TN). If an adverse event was predicted but it did not occur, the prediction is a false positive (FP) and conversely if an adverse event occurred but was not predicted, it is marked as a false negative (FN). There are four commonly derived metrics:

- **Sensitivity (true positive rate):** the proportion of patients with the adverse event who generated an alert, the true positive rate.
- **Specificity (true negative rate):** the proportion of patients without the adverse event who did not generate an alert. $(1 - \text{specificity})$ is the false alarm rate, i.e. the proportion of patients that generated an alert unnecessarily.
- **Positive predictive value (PPV):** the proportion of positive cases that were correctly identified.
- **Negative predictive value (NPV):** the proportion of negative cases that were correctly identified.

The *prevalence* of a condition is the frequency of its occurrence in a population. Sensitivity and specificity do not depend on the prevalence of the event being tested for. However, the positive predictive value does depend on the prevalence of the condition.

		True outcome		
		Positive	Negative	
Predicted outcome	Positive	True Positive (TP)	False Positive (FP)	$PPV = \frac{TP}{TP + FP}$
	Negative	False Negative (FN)	True Negative (TN)	$NPV = \frac{TN}{TN + FN}$
		Sensitivity $TP = \frac{TP}{TP + FN}$	Specificity $TN = \frac{TN}{TN + FP}$	

Table 2.1: Confusion matrix of metrics commonly used to summarise performance of a binary classifier.

The lower the prevalence of true positives, the lower the positive predictive value, for the same sensitivity and specificity.

The ideal track-and-trigger system would have a high sensitivity, such that all patients who triggered are correctly identified as having the outcome, and a high specificity which indicates few misleading alerts or triggers. In practice this is rarely the case, so a suitable balance needs to be struck between sensitivity and specificity [42]. This is commonly investigated using a Receiver-Operating-Characteristics (ROC) curve [43]. The ROC curve is a plot of the true positive rate (sensitivity) against the false positive rate (1 - specificity), as the trigger threshold is varied. 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. An efficient scoring system is characterised by an AUROC close to 1. Figure 2.3 illustrates a typical ROC curve.

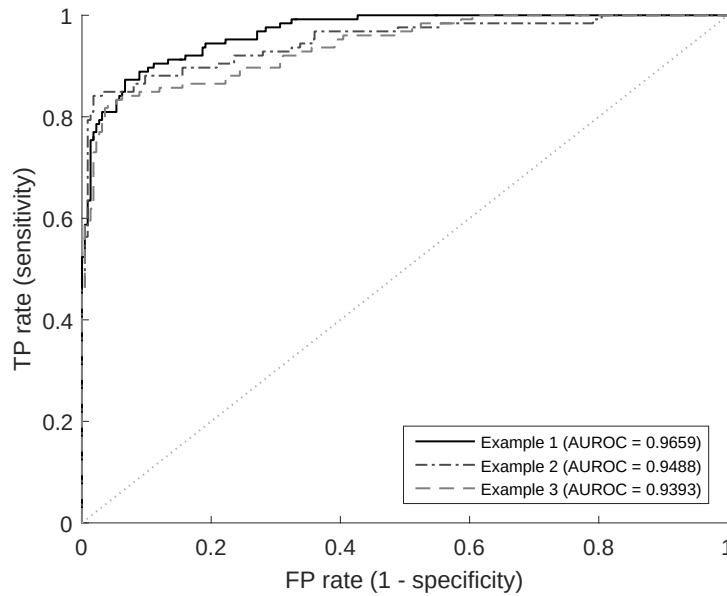


Figure 2.3: Example of Receiver-Operating-Characteristics (ROC) curves. A random classifier would have a AUROC of 0.5 as indicated by the dotted line.

2.2.2 A standardised early warning score for the general adult population

An example of a standardised track-and-trigger system is the UK National Early Warning System (NEWS) for the adult non-pregnant population, published in 2012. Before the development of NEWS, several track-and-trigger systems had been developed and implemented locally around the UK. However, inconsistencies in the scoring method caused different responses to acute illnesses, and confusion for staff transferring between local healthcare centres. The need for a national system had been recognised for several years [44].

The NEWS system (see Figure 2.4) is based on the VitalPAC EWS (ViEWS) system which was developed on a large vital sign dataset of acute medical admissions from Portsmouth Hospitals NHS Trust, UK [45]. Six parameters were included in NEWS: systolic blood pressure, heart rate, temperature, oxygen saturation, respiratory rate, and level of consciousness. The researchers applied a seven band scoring set for each

National Early Warning Score (NEWS)*							
PHYSIOLOGICAL PARAMETERS	3	2	1	0	1	2	3
Respiration Rate	≤8		9 - 11	12 - 20		21 - 24	≥25
Oxygen Saturations	≤91	92 - 93	94 - 95	≥96			
Any Supplemental Oxygen		Yes		No			
Temperature	≤35.0		35.1 - 36.0	36.1 - 38.0	38.1 - 39.0	≥39.1	
Systolic BP	≤90	91 - 100	101 - 110	111 - 219			≥220
Heart Rate	≤40		41 - 50	51 - 90	91 - 110	111 - 130	≥131
Level of Consciousness				A			V, P, or U

*The NEWS initiative flowed from the Royal College of Physicians' NEWS Development and Implementation Group (NEWSDIG) report, and was jointly developed and funded in collaboration with the Royal College of Physicians, Royal College of Nursing, National Outreach Forum and NHS Training for Innovation

Figure 2.4: Scoring metric of the UK National Early Warning Score (NEWS) for the general adult (non-pregnant) population [44].

vital sign that ranged from 3 (most abnormal, below the normal value) to 0 (normal) to 3 (most abnormal above the normal value) except for oxygen saturation and level of consciousness, which are one-sided variables. NEWS was validated using the large Portsmouth database of 198,755 data sets from 35,585 completed and consecutive patient episodes. The primary outcome was death within 24 hours of a given observation set, which occurred for 3,133 (8.8%) of patient episodes. The AUROC for NEWS was 0.89 which is a higher value than the AUROCs of most existing EWS systems and consistent with the performance of ViEWS [44].

The boundary for the NEWS aggregate trigger threshold was evaluated in comparison to another EWS system, the Modified Early Warning Score. The first level of response for NEWS was initiated at an aggregate score of 5 or 6, and a second level of response requiring emergency assessment was initiated at an aggregate score of 7 or more. These thresholds caused an alert for 20% and 10% respectively of data sets collected prospectively over a year on acute medical wards. The designers of NEWS concluded that their score was likely to be more sensitive than current systems, but was also more specific at detecting acute clinical deterioration [44]. The NEWS score has since been validated in

other environments [25, 46, 47].

The NEWS system - a standardised, validated, evidence-based EWS - is an important milestone for hospital care. There has been similar recognition for the need of an evidence-based, standardised obstetric score, as systems for the non-pregnant population do not account for physiological changes in pregnancy [48]. Such a system does not yet exist for the obstetric population. In Chapter 6 we review current obstetric EWS in the literature and propose a new obstetric alerting system.

2.3 Introduction to vital signs

NICE suggests five vital signs as important for routine monitoring of patients: blood pressure, heart rate, temperature, oxygen saturation and respiratory rate [7]. These are the input parameters for the NEWS system, which also includes an assessment of the level of consciousness as the sixth parameter. Here we describe each vital sign and their method of measurement, and then review their known changes during the antenatal and postnatal periods of pregnancy.

2.3.1 Vital signs and their measurement

Blood pressure: Blood pressure (BP) is a measure of the pressure exerted on the wall of the arteries by the blood, in millimetres of mercury (mmHg). Since blood flow is pulsatile, both the maximum and minimum blood pressure are usually recorded. These are during systole (contraction of the heart muscle) and diastole (relaxation of the heart muscle) and referred to as systolic blood pressure (SBP) and diastolic blood pressure (DBP) respectively. BP is influenced by age, BMI, and other environmental factors such as the time of day the measurement is performed.

The clinical gold standard for BP measurement is by Korotkoff sounds [49] using a stethoscope (to listen to the sounds) and sphygmomanometer (to measure the pressure). This is a manual method which requires trained clinical personnel. Recently, automated

blood pressure machines have become widely adopted. These are based on oscillometry techniques whereby an algorithm calculates SBP and DBP from the mean arterial pressure (MAP), the pressure at which maximum amplitude oscillations occur through the cuff applied to the upper arm [50]. Designing a generic algorithm that accurately calculates SBP and DBP from MAP is challenging. Automated machines generally underestimate blood pressure, and therefore may not capture the rise in blood pressure during the last third of pregnancy.

Heart rate: Heart rate (HR) is the frequency of heart contractions, measured as 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. In non-pregnant adults, a normal resting HR is typically between 50 and 90 bpm. HR is usually measured in hospital environments using an electronic device, such as a pulse oximeter [51]. These devices provide an automated, continuous measure of HR.

Temperature: Temperature (Temp) is a measure of core body heat, measured in degrees Celsius ($^{\circ}\text{C}$). The ability of the body to maintain its core temperature within the boundaries required for optimal functionality 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, menstrual cycle or heavy exercise.

To accurately measure core body Temp, a probe must be inserted near to the centre of the body, usually the pulmonary artery. Tympanic thermometers (infra-red) are more appropriate for routine bedside monitoring, and are now standard on the general hospital ward and in primary care [52].

Oxygen saturation: Oxygen saturation is a measure of the amount of oxygen in the blood, defined as the percentage of haemoglobin that is bound to oxygen (ox-haemoglobin, HbO_2). As blood is circulated around the body, oxygen is unloaded from

the haemoglobin molecules.

The gold standard measurement of oxygen saturation is calculated by direct arterial measurement, but this is invasive and not practical for immediate feedback. Oxygen saturation is therefore usually monitored as peripheral oxygen saturation (SpO_2) at the extremities (finger, toe or earlobe) by a pulse oximeter [51]. In healthy adults, normal levels of SpO_2 are between 95 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 occur.

Respiratory rate: Respiratory rate (RR) is the measure of the number of respirations in one minute, reported in respirations per minute (rpm). One respiration consists of inhalation (air movement into the lungs) and exhalation (expulsion of carbon dioxide from the lungs). Resting RR for healthy adults is between 8 and 20 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 [53].

RR is most accurately monitored by capnography, a continuous measure of concentration (or partial pressure) of carbon dioxide in the expired air. Capnography requires a mask over the patient's nose and mouth so that all expired air is captured, so is not desirable outside ICU or the research laboratory. Other methods of RR measurement include electrical impedance plethysmography (electrodes placed around the abdomen) and inductance plethysmography (a band with strain gauges strapped around the abdomen). However, all these methods require set-up time and/or exposure of the abdomen which are not considered suitable for routine measurements. In hospitals, outside ICUs, respiratory rate is usually measured by manual observation of chest wall movement [54].

Respiratory rate is the only vital sign that is not routinely measured with an automatic or electronic device. It is also the only vital sign under voluntary control, such that knowledge of measurement confounds the result. In Chapter 4 we develop a robust

electronic measure of respiratory rate from an accelerometer and pulse oximeter device that is appropriate to be used in pregnancy.

2.3.2 Blood pressure and heart rate in the antenatal period: A systematic review

We carried out a systematic review of the literature to determine the normal physiological trends during pregnancy. This was performed by the author (RP) in partnership with Lise Loerup (LL), and clinically led by Dr Peter Watkinson (PW) and Dr Lucy MacKillop (LM). The protocol for this review is published in [55], and the final paper has been submitted for publication. The review is registered with PROSPERO¹⁰. Here we present the method in brief, and the results relevant to the work of this thesis.

Search strategy

We searched MEDLINE, EMBASE and CINAHL, from inception until June 2015 with guidance from a trained librarian. We searched for papers that had measured any of the six vital signs recommended for clinical assessment (SBP, DBP, HR, Temp, SpO₂, and/or RR) during the antepartum, intrapartum or postpartum stages of pregnancy. We included cross-sectional, case-control and longitudinal studies that contained at least fifty participants. Gestational age must have been specified to within 16 weeks. We further defined inclusion criteria to ensure that the extracted data represented vital sign measurements from an unselected population of pregnant women. For further details, refer to [55]. After screening, two reviewers (RP and LL) extracted data from selected papers, and performed a quality assessment in line with QUADAS-2 [56]. Examples of good-quality studies are those that clearly describe the type of study, exclude women at high risk of pregnancy complications, clearly describe the pregnancy dating method, report gestational age to the day, and use equipment approved for use in pregnancy.

¹⁰PROSPERO registration number CRF42014009673

We found few studies that measured Temp, SpO₂, or RR, and few studies that reported measurements in the intrapartum or postpartum periods. Therefore, we restricted our analysis to BP and HR in the antepartum period only.

To combine the extracted data into a single trend during pregnancy, we used a multilevel model adjusted for use on aggregate data [57]. We introduce multilevel models for individual patient data in Chapter 5 with specifications described in Appendix C. The adjustment of multilevel models for aggregate data is described in Appendix C.6. Only the mean trajectory was investigated, as heterogeneity in the data means variance is difficult to accurately estimate. We used a quadratic polynomial to describe the population mean trajectory.

Results

The database search returned 1,702 publications. Full text was retrieved for 251 studies. Of these, 35 studies met our pre-specified selection criteria. These studies are summarised in Table 2.2. BP data was extracted from 31 studies, and HR data extracted from nine studies (five studies contained both BP and HR data). 30 studies were performed prospectively¹¹. Quality of the studies (as a percentage), ranged from 5.6 % to 84.2 %. The majority of studies (22/35) had a quality score less than 50%. The most common reasons for poor-quality studies were a poorly described study design, and missing details of equipment used for vital sign measurement.

We included 276 summary measurements of BP, which represented a total of 222,797 paired SBP and DBP measurements from 38,962 women. The 23 longitudinal studies reported measurements of BP from 36 separate groups of women (median of four time points per group). A further eight studies reported cross-sectional measurements. Only one study, by Ochsenbein-Kolble *et al.* [58], reported multiple cross-sectional measurements. A mercury sphygmomanometer (manual BP measurement) was used in 16

¹¹A *prospective* study is where the study is designed before data collection begins. This is opposite to a *retrospective* study where the study proposal/question is formed after the collection of the data set.

Year	Author	Country	Study design	Study type	BP measurement	HR measurement	Measurement position	Total women (n)	Groups (n)	Time points per group (n)	Quality score (%)
1969	MacGillivray <i>et al.</i>	UK	LT	P	manual	-	sitting	226	1	7	31.6
1992	Okonofua <i>et al.</i>	Nigeria	LT	P	manual	-	sitting	189	1	6	36.8
1992	Peterson <i>et al.</i>	USA	LT	P	not specified	-	not specified	356	1	6	10.5
1993	Halligan <i>et al.</i>	Ireland	LT	P	manual	-	sitting	98	1	4	63.2
1996	Churchill <i>et al.</i>	UK	LT	P	manual	-	sitting	209	1	3	78.9
1996	van Oppen <i>et al.</i>	The Netherlands	LT	P	-	ECG	sitting	50	1	4	50.0
1999	Matkin <i>et al.</i>	USA	LT	P	not specified	-	not specified	2,853	2	35	10.5
2001	Strevens <i>et al.</i>	Sweden	LT	R	manual	-	sitting	600	1	10	44.4
2002	Iwasaki <i>et al.</i>	Japan	LT	R	automated	-	sitting	1,570	1	3	16.7
2002	Lo <i>et al.</i>	New Zealand	LT	P	manual	-	sitting	101	1	4	52.6
2002	Strevens <i>et al.</i>	Sweden	LT	R	manual	-	sitting	498	3	9	33.3
2006	Gu <i>et al.</i>	China	LT	R	-	other	lateral	182	2	8	41.2
2007	Lauszus <i>et al.</i>	Denmark	LT	P	manual	-	not specified	59	1	2	31.6
2007	Miller <i>et al.</i>	USA	LT	P	manual	-	sitting	1,733	1	3	31.6
2008	Hilmert <i>et al.</i>	USA	LT	P	automated	-	sitting	164	1	3	47.4
2010	Bakker <i>et al.</i>	The Netherlands	LT	P	automated	-	sitting	7,106	3	3	57.9
2011	Jasovic-Siveska <i>et al.</i>	Macedonia	LT	P	manual	-	not specified	400	3	4	15.8
2011	Timmermans <i>et al.</i>	The Netherlands	LT	P	automated	-	not specified	3,187	3	3	21.1
2012	Grindheim <i>et al.</i>	Norway	LT	P	automated	other	lateral	57	1	4	57.9
2012	Lee <i>et al.</i>	USA	LT	P	not specified	-	sitting	1,684	1	2	21.1
2012	Macdonald-Wallis <i>et al.</i>	UK	LT	R	manual	-	not specified	13,016	1	17	27.8
2013	Ishikuro <i>et al.</i>	Japan	LT	P	automated	-	sitting	587	3	8	42.1
2013	Khan <i>et al.</i>	India	LT	P	manual	-	sitting	200	3	3	26.3
2014	Farias <i>et al.</i>	Brazil	LT	P	automated	-	sitting	199	1	3	42.1
2014	Mahendru <i>et al.</i>	UK	LT	P	automated	automated	lateral	54	1	3	84.2
1978	Demidov <i>et al.</i>	Russia	CS	P	not specified	not specified	lateral	50	1	1	22.2
1990	Tuffnell <i>et al.</i>	UK	CS	P	not specified	not specified	sitting	50	1	1	5.6

Continued on next page ...

Year	Author	Country	Study design	Study type	BP measurement	HR measurement	Measurement position	Total women (n)	Groups (n)	Time points per group (n)	Quality score (%)
1994	Ekholm <i>et al.</i>	Finland	CS	P	manual	ECG	sitting/supine	94	2	1	33.3
1994	Tape <i>et al.</i>	USA	CS	P	manual	-	sitting	59	1	1	61.1
2003	Elvan-Taspinar <i>et al.</i>	The Netherlands	CS	P	manual	-	recumbent	54	1	1	55.6
2004	Ochsenbein-Kolble <i>et al.</i>	Switzerland	CS	P	automated	-	sitting	3,234	1	37	66.7
2007	Ngan <i>et al.</i>	China	CS	P	not specified	-	lateral	58	1	1	38.9
2008	Turan <i>et al.</i>	UK	CS	P	-	echo	lateral	4,689	2	1	76.5
2013	van Der Graaf <i>et al.</i>	Australia	CS	P	manual	-	not specified	225	2	1	83.3
2015	Carpenter <i>et al.</i>	UK	CS	P	-	ECG	supine	99	1	1	52.9

LT: Longitudinal design; CS: cross-sectional design.

P: Prospective study; R: Retrospective study.

Table 2.2: *Summary of studies selected for inclusion in the systematic review of BP and HR trends in the antepartum period. Year is year of publication.*

studies, five of which reported that the device had been calibrated prior to the study taking place. An automated BP monitor was used in nine studies. Two of these studies used a monitor approved for use in pregnancy.

We included 34 summary measurements of HR, which represented a total of 7,123 measurements from 5,317 women. The four longitudinal studies reported measurements from five separate groups of women. A further five studies reported cross-sectional measurements of HR at a single time point. The method of HR measurements was reported as electrocardiography (ECG; 3 studies), echocardiography (echo; 1 study), finger arterial pressure sensor (1 study, Grindheim *et al.*), radioulnar pulse wave monitoring (1 study, Gu *et al.*), reading from an automated BP monitor (1 study), or not specified (2 studies).

Significant heterogeneity was present for SBP, DBP and HR in all trimesters. The modelled trajectories for BP and HR, using a multilevel model for aggregate data, are shown in Figure 2.5. We were not able to adjust for maternal characteristics, such as age and weight, as such data were often missing from the publications, or reported using different summary measures at different gestational ages. Both SBP and DBP changed significantly across pregnancy (for combined manual and automatic readings). Mean (95% CI) SBP at 10 weeks gestation was 110.5 (109.1, 112.0) mmHg. Mean (95% CI) SBP was lowest at 16 weeks, 110.1 (108.7, 111.6) mmHg, rising to 116.3 (114.4, 118.2) mmHg at 40 weeks gestation; a change in SBP of 6.2 (4.9, 7.5) mmHg from 16 to 40 weeks ($p < 0.001$). Mean (95% CI) DBP was 64.6 (63.2, 66.1) at 10 weeks gestation. DBP was lowest at 19 weeks gestation, a mean (95% CI) of 63.3 (62.0, 64.7), rising to 70.7 (69.0, 72.4) mmHg at 40 weeks gestation; a change of 7.4 (6.3, 8.4) mmHg from 19 to 40 weeks ($p < 0.001$).

Data was available from 25 studies to allow us to assess the effects of parity on SBP and DBP. Nulliparous women had a slightly higher BP on average, but this was not statistically significant.

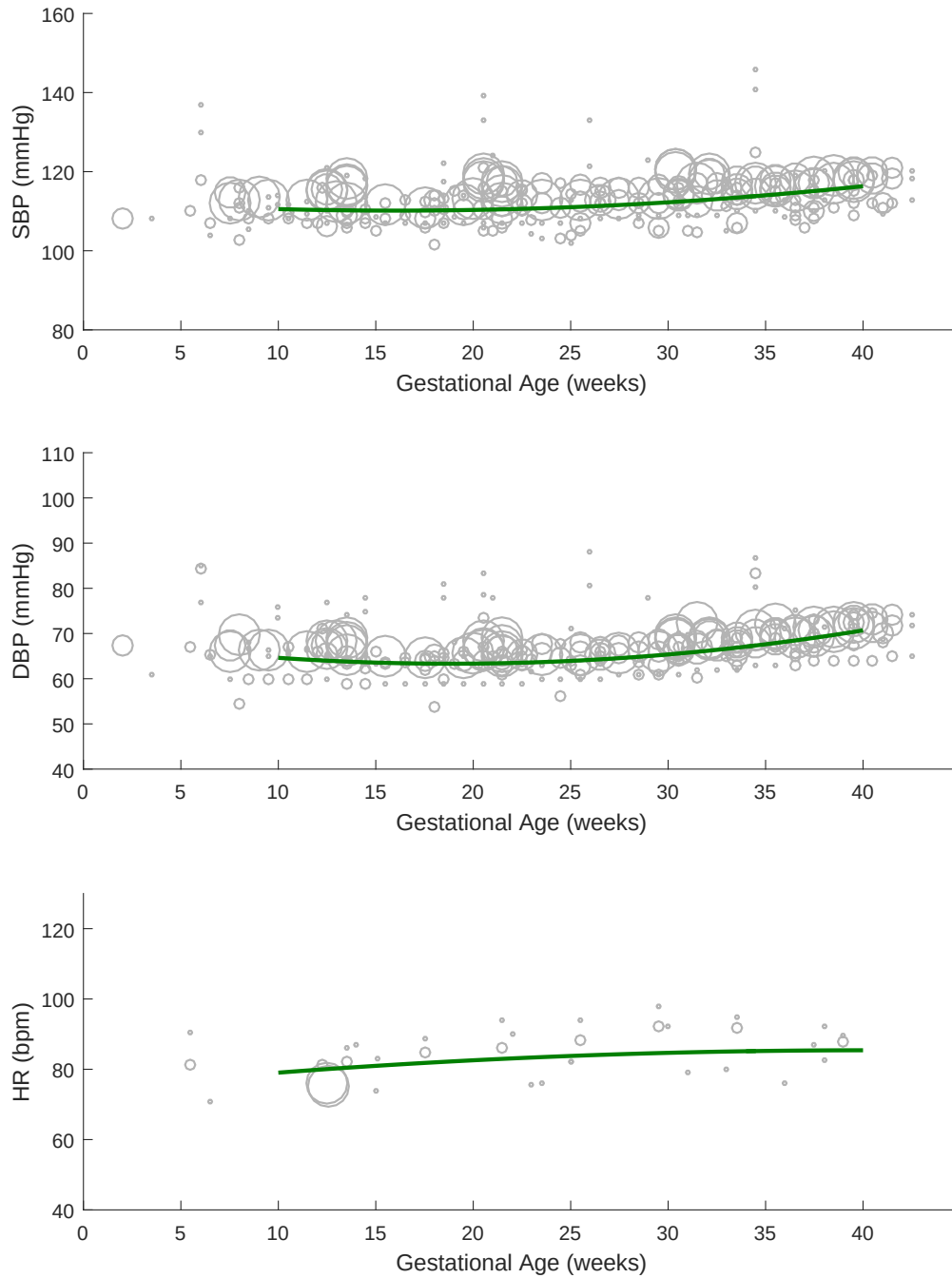


Figure 2.5: Mean trajectory of SBP (top), DBP (middle), and HR (bottom) during pregnancy, modelled by synthesis of data available in the literature. The raw data points are shown as grey circles, where the size of the circle reflects the number of participants in the study. Modelling was performed using a quadratic polynomial multilevel model, adjusted for aggregate data.

Discussion

The systematic review described here has synthesised the available evidence base to inform our understanding of changes in BP and HR during pregnancy. Our analysis showed that SBP and DBP dip very slightly in mid-pregnancy and then rise in the third trimester. The absolute change during the antepartum period was larger for DBP (7.4 mmHg) than SBP (6.2 mmHg). HR steadily increased during pregnancy.

Although several studies of BP and HR during the antepartum period were identified, they suffered from small sample sizes (fewer than 50 participants) and therefore were excluded from our analysis. Of the 35 studies included, many were of poor methodological quality. Three aspects were of particular concern. Firstly, a lack of accuracy in measuring and reporting of gestational age was common (23/35 studies with high bias for this quality criterion). Secondly, few BP measurements were performed with a calibrated device, or a device validated for use in pregnancy. Thirdly, the selection of appropriate populations to model was not robust. The poor quality of many of these studies likely contributes to the high between-study variability in mean BP and HR estimates. While the lack of accuracy in gestational age calculation may have affected the timing of observed trends, the choice of measurement device and population will have affected the mean and spread of the reported data. Combining such heterogeneous data is challenging, and only the mean trajectories can be calculated accurately.

This investigation suggests the need for more high-quality, large cohort studies of low-risk pregnant women from which to understand physiological changes during pregnancy and after delivery. Previous work in adult early warning scores shows that relatively small changes in thresholds have large changes in the ability of such a score to detect deterioration [40].

Note on postnatal trends

Of the 23 longitudinal studies selected for inclusion in the systematic review of antenatal vital sign changes, four also performed a postnatal measurement. Halligan *et al.* [59] (1993) measured BP at six weeks after delivery, Churchill *et al.* [60] (1996) measured BP at 12 weeks after delivery, Grindheim *et al.* [61] (2012) measured BP and HR at six months after delivery, and Mahendru *et al.* [62] (2014) measured BP and HR at 14 to 17 weeks after delivery. Data from these studies represent such a wide time range after delivery that it is challenging to obtain a complete understanding of BP and HR changes during this time. The four studies do not agree on whether BP increases or decreases in the weeks/months after delivery. The HR data shows that HR falls by 10 to 12 bpm some time during the three to six months after delivery [61, 62]. To further understand vital sign changes in the postnatal period, studies focusing on this area need to be performed.

2.3.3 Other vital sign changes

Temperature: Most studies of temperature and pregnancy are performed during labour. One study by Hartgill *et al.* [63] was identified that studied Temp longitudinally through pregnancy and the postpartum period. Temp was reported to decrease during pregnancy from 37.1 °C at 8 weeks gestation to 36.7 °C at 36 weeks gestation. At 12 weeks after delivery, Temp was measured at 36.4 °C and thought to stabilise at 36.7 °C from 24 weeks after delivery. Data in this study was collected from only 15 participants, and we could not identify a study that had reproduced the results. Obstetric textbooks do not specify an expected change in Temp [64].

Oxygen saturation: Little is known about how SpO₂ changes in pregnancy. It has been shown that the affinity of haemoglobin to oxygen decreases in order to meet an increase in demand from the placenta and fetus [65]. This means a right shift in the

oxygen dissociation curve (Figure 2.6) such that a greater partial pressure of oxygen is required to maintain the same level of SpO_2 . However, clinical professionals do not expect this to have an affect on SpO_2 during pregnancy. The only study we identified that measured SpO_2 in any stage of pregnancy was by Deckhardt *et al.* [66]. They observed mean SpO_2 at term was 96%, similar to the non-pregnant population.

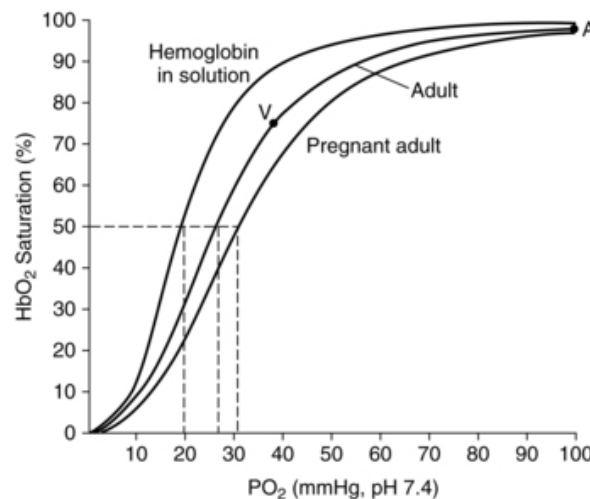


Figure 2.6: The dissociation of oxygen from haemoglobin is described by a sigmoid curve that depends on the partial pressure of oxygen. During pregnancy this curve is shifted to the right. Reproduced from [65].

Respiratory rate: An improved respiratory efficacy is required during pregnancy to ensure adequate oxygenation. Pregnant woman adjust their breathing by breathing more deeply (tidal volume increases 30-40% during pregnancy), but at the same frequency. RR is therefore thought to remain unchanged at around 14 to 17 rpm [67, 68]. After delivery the respiratory system rapidly returns to its pre-pregnancy state, a direct result of loss of progesterone (a hormone) and reduced intra-abdominal pressure (after delivery of the baby) which allows increased excursion of the diaphragm. RR remains constant [65].

Blood pressure and heart rate: During pregnancy, blood volume increases, which is balanced by a decrease in stroke volume and an increase in HR. BP is not affected by the change in blood volume during pregnancy. Instead, the mid-pregnancy dip observed

by the BP studies included in the systematic review (Section 2.3.2) is caused by hormonal changes. BP in pregnancy is affected by similar factors as for the non-pregnant population, such as age and BMI [61, 69, 70]. Parity is also thought to influence BP [71, 72].

Immediately after delivery, HR remains elevated (for 24 hours), then steadily returns to pre-pregnancy values by 10 days after delivery [73]. In the week after delivery, up to 2 litres of body fluid is mobilised and excreted. It is hypothesised that BP increases between day 3 and 6 after delivery to compensate for the loss of fluid, although no conclusive studies could be found [65].

2.4 Conclusion

In this chapter we have discussed the burden of pregnancy-related complications, and how their prevalence and severity can be reduced by timely recognition of physiological deterioration to enable clinical intervention. However, the pregnancy-associated changes in vital signs are known to complicate the recognition of physiological deterioration, and these changes are not well understood. In particular, there has been little investigation into vital sign changes in the immediate postpartum period. The next chapter introduces the 4P study which aims to gather a database of vital sign information during pregnancy and in the two weeks after delivery from women experiencing a normal pregnancy.

Description of dataset

The limited knowledge of vital sign changes in pregnancy led to the design of the 4P (Pregnancy Physiology Pattern Prediction) study, an ongoing longitudinal observational study based in Oxford UK. The 4P study recruited women in early pregnancy and monitored them until two weeks after delivery. Maternal vital sign measurements and information about medication and complications were collected at regular intervals during the study. In this chapter we introduce the 4P study, present the populations and vital sign distributions used for analysis in this thesis, and assess the data quality of pulse oximeter recordings.

3.1 The 4P study

The 4P study has been executed in two phases. The first phase ran from 1 August 2012 to 16 May 2016, as part of the INTERBIO-21st study¹ based at the John Radcliffe Hospital Maternity Centre, Oxford, UK (Oxford University Hospitals NHS Foundation Trust)². In July 2015 the second phase began as an independent study³. The second phase included additional centres in Newcastle (Royal Victoria Infirmary, Newcastle Hospitals NHS Foundation Trust) and London (St Thomas' Hospital, Guy's and St

¹www.interbio21.org.uk

²Oxford Research Ethics Committee (ref: 08/H0606/139)

³National Research Ethics Service committee South East Coast - Brighton and Sussex (ref: 14/LO/1312)

Thomas' Hospital NHS Foundation Trust).

3.1.1 Study population

The 4P study aimed to recruit 1,000 women. This target was reached in December 2016. Women were recruited in early pregnancy and required to have an ultrasound dating scan between 9^{+0} and 13^{+6} weeks of gestation. The full inclusion criteria are presented in Table 3.1.

Recruitment rate was determined by the availability of research midwives. Once the pre-arranged monthly recruitment target was reached, recruitment paused. The recruitment target for Phase 1 of the 4P study was 25 a month; for Phase 2, recruitment targets per month per site were: 15 for Oxford, 10 for Newcastle, and 8 for London.

Recruitment criteria
At least 16 years of age (parental consent to be obtained for participants under 18 years old)
Naturally conceived pregnancy
Singleton pregnancy
Intent to deliver at the hospital of the recruiting centre
BMI < 35 kgm ⁻²
Ultrasound dating scan < 14 weeks
Able to understand written or spoken English or has their own interpreter willing to translate
Willing and able to use the home monitoring equipment

Table 3.1: *Recruitment criteria for the 4P study*

3.1.2 Timeline for data collection

Women were recruited to the 4P study in early pregnancy, usually at their first antenatal ultrasound scan, and monitored by research midwives at home visits every four to six weeks until delivery. After delivery, women were given a self-monitoring kit to continue monitoring their own vital signs once a day for the next 14 days. The midwives visited the women twice during this postnatal period to take additional readings. At each midwife visit, vital sign data and other pregnancy-related information were collected from the women. Intrapartum vital sign data and delivery information were extracted

in retrospect from hospital notes (paper charts). The study timeline is shown in Figure 3.1.

Gestational age was calculated from crown-rump measurement obtained by ultrasound prior to 14 weeks of gestation, using the formula developed from the INTERGROWTH-21st dataset [74]. After delivery, time was measured as number of calendar days after delivery, where the day of delivery is day zero. We chose calendar days, rather than hours, after delivery because our time scale is a number of days (around 14) which becomes cumbersome to calculate in hours.

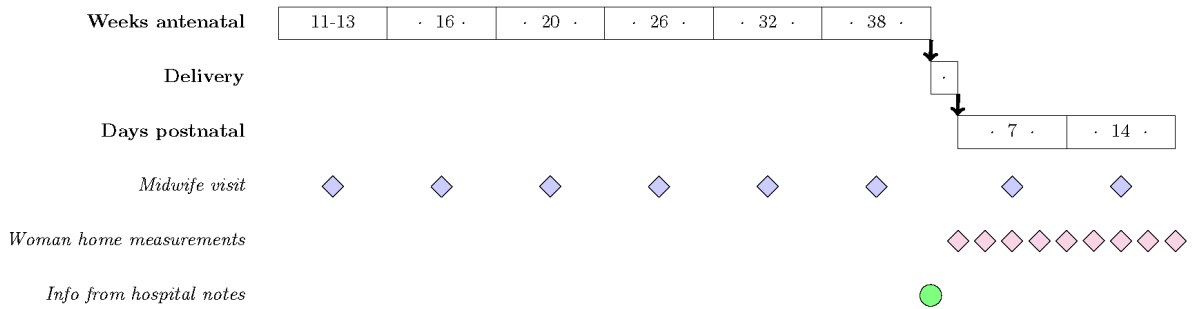


Figure 3.1: Timeline for vital sign data collection in the 4P study. Note that women were expected to perform daily measurements for two weeks after delivery; a total of 14 sets of measurements.

3.1.3 Vital sign data

Vital sign information (BP, HR, Temp, SpO₂, and RR) was collected at every antenatal and postnatal midwife visit using the equipment described in Section 3.2.1. During the home monitoring phase, women recorded their own vital sign information with the exception of RR. Correct use of home monitoring equipment was taught during a 10-minute training session at the final antenatal appointment.

3.1.4 Labour and delivery information

During labour and the women's hospital stay, vital sign measurements were regularly taken by hospital staff and recorded on paper charts. The partogram is used to record

progress during labour including maternal vital signs (see Figure 3.2), and the local obstetric early warning score chart (called MEOWS in Oxford) used on the general obstetric wards (see Figure 3.3). If the woman received an epidural or anaesthetic, these were recorded on additional charts. Charts were manually transcribed to the 4P database by research midwives.

In addition to vital sign information, the labour information (time of each labour stage onset, reason for inducing labour or performing caesarean section if relevant), delivery information (mode of delivery, estimated total blood loss, weight and status of newborn) and hospital information (date of discharge from hospital, haemoglobin/haematocrit, postnatal blood transfusions) were extracted from the woman's hospital notes.

3.1.5 Other information collected

At recruitment, baseline information was collected about the woman such as demographic information (date of birth, height, weight, ethnicity), obstetric history (parity, gravidity⁴, last menstrual period), medical history (such as smoking status, diabetes, cardiac disease, respiratory disease, kidney disease, hepatic disease), pregnancy-related health, and current medications. Pregnancy-related health includes whether the woman has developed any complications during this pregnancy, such as gestational diabetes, gestational hypertension, pre-eclampsia (see Section 2.1.2), or severe vomiting requiring hospitalisation. At each subsequent follow-up visit during pregnancy, and at the final postnatal visit, current health, pregnancy-related health, and prescribed medications were updated. Midwives entered this information directly onto the 4P website (synchronised with the 4P database) using their computer tablet while with the participant.

After the home monitoring phase was completed, the women were asked to provide feedback on the usability of the mobile health system through a standardised usability questionnaire.

⁴The number of previous pregnancies.

Maternal condition															
Time (left line = 1 hr)															
190															
Pulse rate • 180															
170															
BP recordings 160															
150															
140															
130															
120															
110															
100															
90															
80															
70															
60															
50															
Maternal temperature °C															
Pressure areas checked															
Maternal position															
Coping strategies:															
Pool temperature															
Urine volume															
Urine analysis															
Contractions: 10	5														
	4														
Weak	3														
Moderate	2														
Strong	1														
Syntocinon mls/hr															
Drugs and intravenous fluids															

Figure 3.2: The section of the partogram (used during labour) that records maternal vital sign values.

Section A – General observations																
Date:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Time:																
Gestational age / days postnatal:																
Looks unwell	Yes															
	No															
Temperature °C	39.5															
	39															
	38.5															
	38															
	37.5															
	37															
	36.5															
	36															
	35.5															
Maternal heart rate b.p.m.	170															
	160															
	150															
	140															
	130															
	120															
	110															
	100															
	90															
	80															
Systolic blood pressure mmHg	200															
	190															
	180															
	170															
	160															
	150															
	140															
	130															
	120															
	110															
Diastolic blood pressure mmHg	130															
	120															
	110															
	100															
	90															
	80															
	70															
	60															
	50															
	40															
Respiratory Rate (please indicate actual rate)	>30															
	21-30															
	11-20															
	0-10															
Urine Analysis	Protein															
	Other															
Section A Total Score																

Figure 3.3: The first page of the Oxford MEOWS chart used in obstetric wards.

3.2 The 4P system

The system for data collection, particularly vital sign data, evolved during the 4P study. The main changes occurred between the first and second phases of the study, as new sites were added (Newcastle and London) and the INTERBIO-21st study came to a close so could no longer aid recruitment and antenatal visits in Oxford. We have labelled the different systems from 0 to 3, as described in Table 3.2. System 0 was the initial system for vital sign data collection, used in the clinic for antenatal measurements. System 1 was developed for postnatal home monitoring in Phase 1. System 2 was an adjustment of system 1 that catered for multiple sites as required in Phase 2. Additional functions (such as a 60-second respiratory count and manual HR measure) were incorporated to give system 3, so system 2 was only used for the first four months of Phase 2. System 3 is the current system used in all 4P sites for both antenatal and postnatal stages of data collection.

	PHASE 1 (Aug 2012 - May 2016)		PHASE 2 (July 2015 - Sep 2017)	
	Antenatal measurements	Postnatal measurements	Version 1	Version 2
System	0	1	2	3
Duration of manual RR measure	15s	30s	30s	60s
Duration of manual HR measure	-	-	-	60s
Duration of pulse oximeter recording	Several minutes	60s	60s	120 - 300s
Time-stamps of respirations recorded	no	no	no	yes
Duplicate recordings performed	no	no	no	Oct 2015 to Mar 2016
Sites	Oxford	Oxford	Oxford, London, Newcastle	Oxford, London, Newcastle

Table 3.2: During the 4P study the system for data collection has evolved, particularly for vital sign data collection.

3.2.1 Equipment for vital sign collection

In this section we describe the equipment for vital sign data collection in the 4P study. We selected commercial medical devices regularly used in the hospital. These devices are interfaced to a Samsung Galaxy Tablet on which a custom-designed application guides the midwife and participants through data collection. The tablet application is described in Section 3.2.2.

Heart rate and oxygen saturation: We used a Nonin Bluetooth finger pulse oximeter (WristOx₂ 3150 finger pulse oximeter, Nonin Medical, Minnesota, USA) to measure HR and SpO₂. Pulse oximeters measure blood volume changes at the periphery, and are used to record continuous SpO₂ on hospital wards. The Nonin model transmits data via Bluetooth so manual transcription is not required. In 4P system 0, we recorded the pulse oximeter trace for the entire appointment duration, usually seven to eight minutes. In system 1, 2, and 3, the pulse oximeter duration was limited (Table 3.2). We describe the pulse oximeter in more detail and explain how we obtain a robust estimate of HR and SpO₂ from the pulse oximeter time-series, the photoplethysmograph (PPG), in Section 3.4.

In the second phase of the 4P study, midwives performed an additional HR measurement by pulse palpation for 60 seconds. We refer to this measure as Manual HR to distinguish it from the automatically-calculated HR from the pulse oximeter.

Blood pressure: We selected a Microlife automated blood pressure machine (BP 3BT0-A(2) and WatchBP Home, Microlife, Taiwan), that was validated for use in pregnancy, for BP collection in the 4P study [75, 76]. An automated machine was selected over the clinical gold standard Korotkoff sounds because it allowed measurement by non-expert users, as required during the home monitoring stage when women self-monitor at home.

Temperature: We use the GeniusTM 2 tympanic (ear) thermometer (MedTronic, Min-

neapolis, USA, formerly Covidien) which is commonly used in UK hospitals. In addition, system 0 also used a Bluetooth tympanic thermometer (IR20, Fora, California, USA) which transferred readings automatically, thus avoiding manual transcription errors. However, the Bluetooth thermometer was difficult to use and not thought suitable for home monitoring, so only the Genius thermometer was used in system 1, 2, and 3.

Respiratory rate: On hospital wards, RR is usually measured by manual observation of chest wall movement. We therefore used a manual chest wall respiratory count in the 4P study. Midwives counted respirations during the pulse oximeter recording. Respirations were counted for 15, 30, or 60 seconds depending on the 4P system used (see Table 3.2).

We also recorded accelerometer information from a Samsung Galaxy S3 smartphone (Samsung, South Korea) placed on the woman's chest while reclining (Figure 3.4). Respiratory rate is retrospectively estimated from the accelerometer recording using algorithms described in Chapter 4.

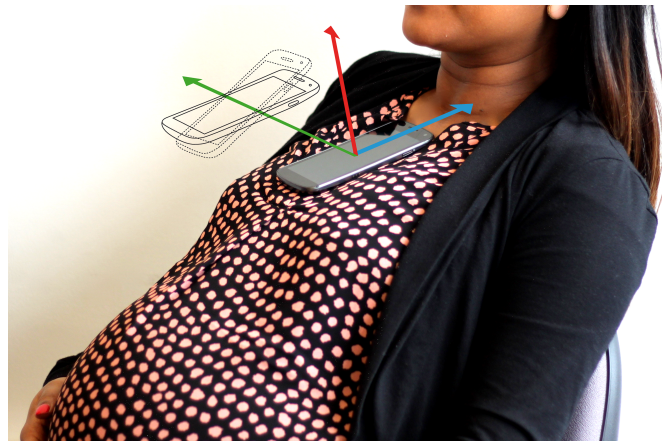


Figure 3.4: *Position of smartphone on woman's chest during respiratory rate recording.*

Summary of equipment models: Since the 4P study was spread over several years, multiple equipment orders were required. Where available, the same model of device was selected. However, this was not always possible. Of note is that the blood pressure monitor Microlife BP 3BT0-A(2) ceased production in mid-2014 with no updated model.

After correspondence with the manufacturers, we selected the Microlife Watch BP Home which used the same oscillometric algorithm as our previously selected device, and was also validated for use in pregnancy [76]. Manufacturers and models for equipment at each 4P study site are detailed in Table 3.3 and shown in Figure 3.5.

Site(s)	Oxford	Newcastle and London
Tablet	Samsung Galaxy Tab 2.0	Samsung Galaxy Tab 4.0
HR and SpO ₂	Nonin® Bluetooth WristOx ₂ 3150	Nonin® Bluetooth WristOx ₂ 3150
BP	Microlife BP 3BT0-A(2)	Microlife Watch BPhome
Temp	Covidien Genius™2	Covidien Genius™2
Smartphone	Samsung Galaxy Nexus GT-I9250	Samsung Galaxy S3

Table 3.3: *Equipment for vital sign measurement.*



Figure 3.5: *Equipment used for vital sign measurements in the 4P study (Oxford models shown).*

3.2.2 Tablet application for vital sign data collection

An Android tablet application (app) was developed to aid vital sign data collection. Two versions of the app were produced. The first was for use in the hospital clinic during

antenatal appointments of Phase 1 of the study (system 0). Here, vital sign information was recorded by INTERBIO-21st staff at antenatal ultrasound appointments. The app displayed a single screen on which vital sign measurements could be entered in any order. Data from the finger pulse oximeter and phone accelerometer were often recorded for the duration of the appointment.

The second version of the app was designed for women to use during self-monitoring (as non-expert users). This version was also adopted for postnatal midwife visits (system 1), and then for Phase 2 of the 4P study (systems 2 and 3). There were thus two user settings for the app: as a midwife and as a participant.

The tablet app was designed to be intuitive to navigate. The first screen displays a note to participants reminding them that their information is not being monitored in real-time by a midwife, so to seek medical help if they feel unwell. The app then explains how to measure each vital sign, moving to a new screen after a value is entered (see schematic of tablet application in Figure 3.6). In the midwife version of the app, the Manual HR and RR measurements are taken during the pulse oximeter recording.

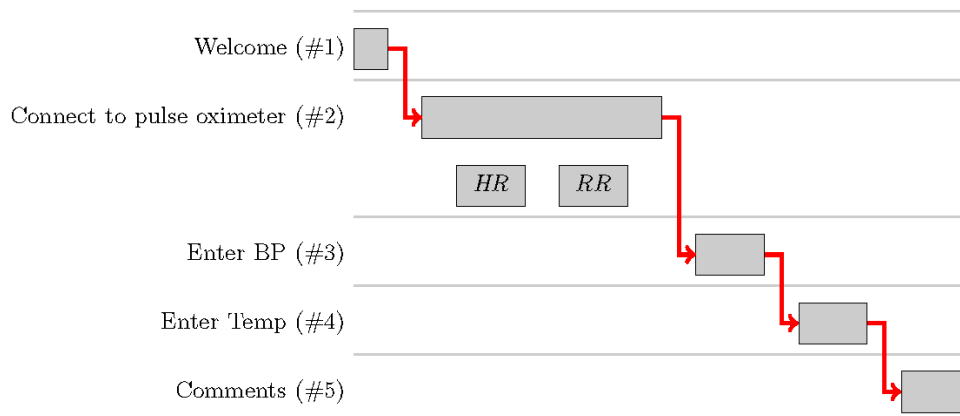


Figure 3.6: Progression of screens in the second version of the 4P tablet application. For the midwife version of the app, Manual HR (60s) and RR (60s) are taken during the pulse oximeter recording (participants do not perform a Manual HR or RR measurement). The movement from screen #2 to #3 is automatic when the pulse oximeter completes (after 60s for participants; at conclusion of RR for midwives). All other transitions require user input.

During measurement of respirations, the midwives indicate the time of each breath using a keypad button on the tablet. At the end of the app, there is space for comments. Users could not move backwards in the application, so if they noticed they had made a mistake or accidentally skipped a screen they often wrote a note in the comments section. This was helpful for midwife review of missing or unusual data. Screenshots of the second version of the app are shown in Figure 3.7. At the end of home monitoring, the tablet was factory reset before being assigned to the next participant.

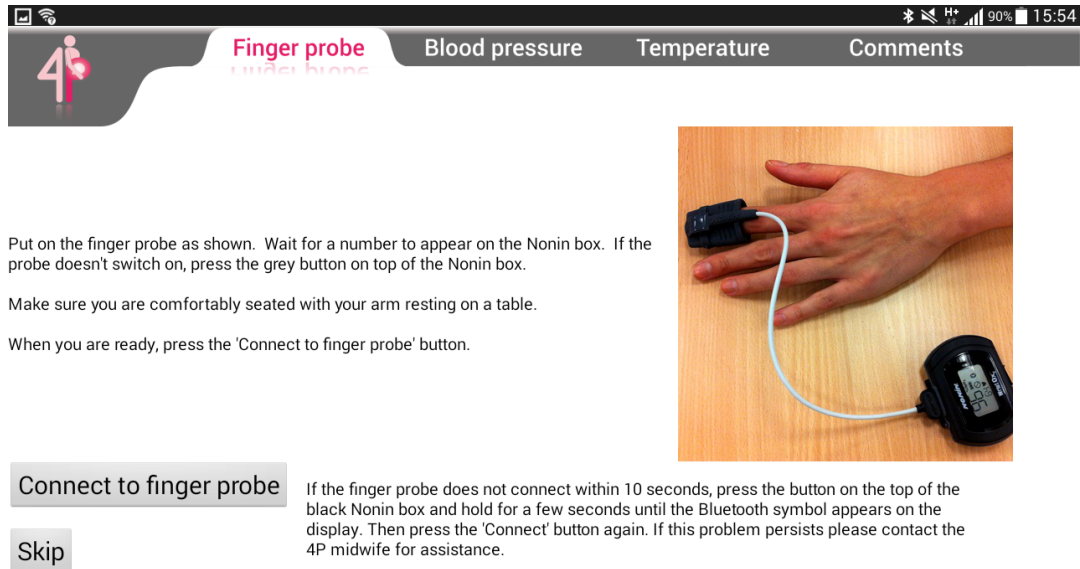
3.2.3 Website system

The 4P website was customised as a portal for the research midwives to manage the 4P study. The website allowed the midwives to create participant ID numbers, review vital sign data, and enter other participant information relevant to the study. Each midwife had her own login and could create and view participants recruited at her site. Participants were initially viewed under “Antenatal Participants”. After delivery, they moved to the homescreen where a calendar allowed the midwives to monitor the progress of postnatal self-monitoring.

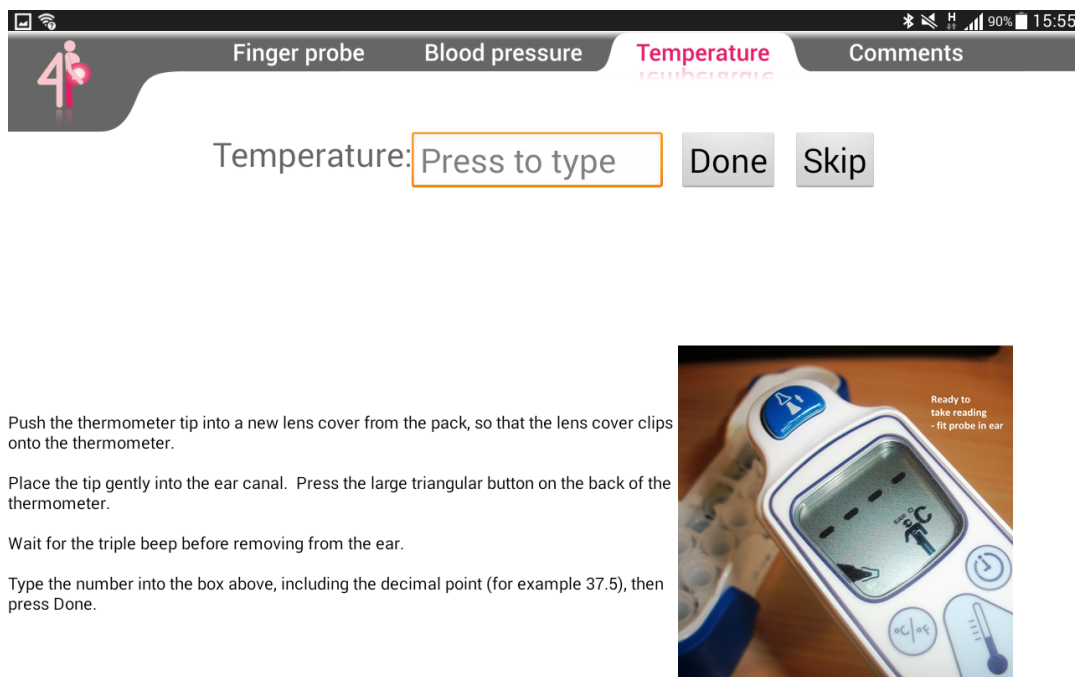
Reviewing vital sign data: The website accesses the 4P database to display vital sign information that has been entered into the tablet app (see screenshot in Figure 3.8).

Case report forms: Additional participant information, such as demographic information, complications developed, and medication, were entered onto the website at every visit using a series of case report forms.

Intrapartum charts: Vital sign data and birth information recorded on paper charts during delivery were manually entered onto the website. Charts included the maternal EWS, partogram, anaesthetics, and epidural charts. Midwives photocopied the patient notes (after removing patient identifiable information) so charts could be entered retrospectively.



(a) Screen #2 (participant version): connection of pulse oximeter to tablet via bluetooth for HR and SpO_2 measurement.



(b) Screen #4: input of Temp measurement.

Figure 3.7: Screenshots from the participant version of the 4P tablet application.

PARTICIPANT 6003

Participant ID	6003	Site	Oxford
Date of birth	270179	Personal mobile phone	
Expected delivery date	20 Jan 2015	Delivery date	21 Jan 2015
Kit assigned	Yes	Passcode	248199463

Notes

EDIT DETAILS

COMPLETE STUDY

ADD NOTE

Data received

	Date		Heart rate	SpO2	Blood pressure	Temperature	Respiration	Comments
	Fri 23 Jan 2015 11:22		61/min	98%	120/80	36.1 C (Genius)		test
	Fri 23 Jan 2015 11:25		60/min	98%	124/85	37.0 C (Genius)	5 breaths/min	test

Figure 3.8: Website screenshot: Vital sign data review

3.2.4 Data transmission and storage

At the end of a vital sign data recording session, the data stored in the Android tablet were automatically transmitted to the 4P database, hosted on a secure server residing behind the NHS firewall. Data were transmitted primarily via the 3G/4G network, or by a secure Wi-Fi network if available in the hospitals. After successful transmission, the data was deleted from the tablet's internal memory. Data stored on the 4P database are anonymised by the assignment of a four-digit unique identifier (Participant ID) that is 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. A master list that links the names of the women with their Participant ID was securely stored in a locked location only accessible to clinical research staff.

3.3 Data used in this thesis

The 4P study has been described here in full to provide motivation for the analysis carried out in this thesis, and to describe the mobile health system developed for vital sign data collection. For the remainder of this thesis we will examine the 4P data collected in the antepartum and postpartum periods only, for those women who had delivered by 15 July 2016 (to allow sufficient time for data analysis before thesis submission).

3.3.1 Study population

During Phase 1 (1 August 2012 to 16 May 2016), 568 women were recruited to the 4P study. A further 144 women were recruited to Phase 2 of the 4P study and had delivered by 15 July 2016. Vital sign measurements and other information were therefore recorded and available for 712 women.

For the investigations in this thesis we will only consider the 620 women who were *healthy*, *normal*, or *low-risk* (defined in Appendix A). Those women with a history of chronic disease, as reported on their first antenatal appointment, were separated from the healthy population. Women were not excluded if they later developed complications or medical conditions during pregnancy. Demographic information for the two groups of women, along with reasons for exclusion where appropriate, is displayed in Table 3.4. Additional information only collected during Phase 2 is also presented.

A large proportion of the women recruited were of White British ethnicity (76.4% in Phase 2). The most diverse population was from the London site, where 19.2% of the women were of Black ethnicity, 11.5% Asian, and 3.9% mixed race⁵.

Of the 620 women who were recruited as *healthy*, 27 women discontinued the study during antenatal monitoring, and a further 55 discontinued during the postnatal monitoring phase. In addition, 15 women declined home monitoring (but still allowed post-

⁵Ethnicity was categorised according to the definitions in the UK 2011 census. Ethnicity information was only collected in Phase 2 of the 4P study.

	Healthy women n = 620	Excluded women n = 92
<i>Baseline information, median (IQR)</i>		
Age (years)	31.8 (4.8)	31.0 (5.6)
Weight at recruitment (kg)	67.2 (12.8)	69.7 (14.1)
BMI at recruitment (km/m ²)	24.5 (4.2)	25.7 (5.0)
Gestational age at recruitment (weeks)	12.9 (2.5)	12.4 (1.9)
<i>Additional information, n (%)</i>		
Nulliparous	256 (41.3%)	33 (35.9%)
Current smoker	45 (7.3%)	13 (14.1%)
<i>Chronic conditions, n (%)</i>		
Hypertension	-	16 (17.4%)
Malignancy/cancer	-	5 (5.4%)
Cardiac disease	-	21 (22.8%)
Respiratory disease	-	68 (73.9%)
<i>Additional information Phase 2, n (%)</i>		
	n = 132	n = 12
Co-living with smoker	21 (15.9%)	6 (50.0%)
White British ethnicity	98 (74.2%)	12 (100.0%)

Table 3.4: Baseline information for women in the 4P study split into two groups: healthy women, and those with a history of chronic disease excluded from further analysis.

natal midwife visits). The reasons for discontinuation at each stage of the study are summarised in the flowchart in Figure 3.9.

3.3.2 Pregnancy complications and delivery information

Several women developed pregnancy-related complications during pregnancy or in the two weeks after delivery. The most prevalent complication was gestational hypertension (5.3%). Only two women developed severe pre-eclampsia (0.3%). The main complications and other medical conditions developed during pregnancy are detailed in Table 3.5.

Delivery information was available for those women who completed antenatal monitoring (n = 593). Most women delivered between 38⁺⁰ and 42⁺⁰ weeks of gestation, as expected for a normal pregnancy. 45 women (7.6%) delivered pre-term, defined as be-

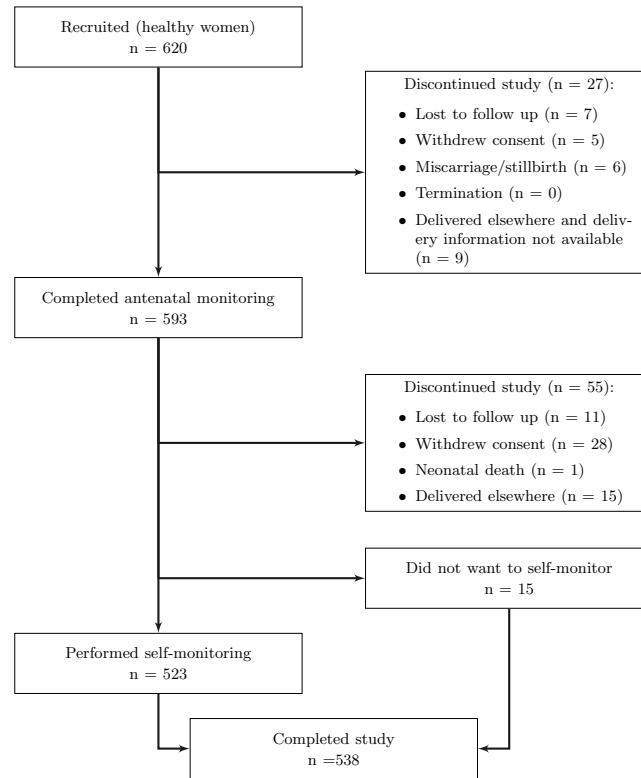


Figure 3.9: Flow chart of number of women recruited to the 4P study and reasons for withdrawal at each stage.

Healthy women n = 620	
<i>Pregnancy-related complications, n(%)</i>	
Gestational hypertension	33 (5.3%)
Pre-eclampsia	11 (1.8%)
Severe pre-eclampsia	2 (0.3%)
Gestational diabetes ⁶	24 (3.9%)
<i>Other conditions, n(%)</i>	
Diabetes	29 (4.7%)
Cardiac disease	2 (0.3%)
Respiratory disease	8 (1.3%)
Respiratory tract infection	32 (5.2%)
Urinary Tract Infection (UTI) requiring antibiotics	90 (14.5%)

Table 3.5: Pregnancy-related complications and other medical conditions developed during pregnancy for the women in the 4P study.

	Healthy women with delivery information n = 593
<i>Delivery information, median (IQR)</i>	
Gestational age delivered (wks of gestation)	39.9 (2.0)
Newborn weight (g)	3350 (689)
<i>Delivery mode, n(%)</i>	
Spontaneous Vaginal Delivery (SVD)	372 (62.7%)
Assisted vaginal delivery (e.g. use of forceps)	98 (16.5%)
Caesarean section	106 (17.9%)
Breech	1 (0.2%)
Not recorded	16 (2.7%)
<i>Additional information Phase 2, median (IQR)</i>	
Blood loss during delivery (mL)	400 (250)
Length of hospital stay after delivery (days)	1 (1)

Table 3.6: *Delivery information for women in the 4P study*

fore 37⁺ weeks. The most common mode of delivery was Spontaneous Vaginal Delivery (62.7%). Other delivery information is detailed in Table 3.6.

3.3.3 Vital sign measurements

Vital sign recordings were taken at regular intervals during pregnancy and after delivery. We refer to a set of vital sign measurements (BP, HR, Temp, SpO₂, and RR) as a vital sign *diary*.

Data quality and cleaning

Before analysis, the vital sign data was *cleaned* to ensure it was as complete and as high-quality as possible. The number of vital sign diary adjustments are presented in Table 3.7. We made the following decisions (codes refer to Table 3.7):

- If a missing vital sign value was entered in the comments field of the tablet app (last screen), the value was manually entered. [code: miss]

⁶The new onset of diabetes (characterised by a high blood sugar) during pregnancy. Gestational diabetes is not thought to have an effect on vital signs.

Code	Explanation	Action taken	Phase 1 (n)	Phase 2 (n)
sd	Vital signs split over two diaries with no overlap	Diaries merged	67	46
rep	Incomplete diary. Another complete diary was received on the same day	Incomplete diary excluded	169	22
miss	Missing vital sign information was entered in comments	Value added	22	12
wro	Incorrect vital sign information was corrected in comments (eg. typing error)	Value adjusted	7	12
id	Incorrect participant ID entered by the midwife	ID adjusted	42	6

Table 3.7: *Reasons for excluding or merging diaries, or correcting vital sign values.*

- If comments referred to an incorrectly entered vital sign and provided the correct value, this was adjusted. [code: wro]
- If two diaries were received in quick succession (within one hour) and one diary contained some vital sign information (for example the HR and SpO₂ information from the pulse oximeter) and the second diary contained the remaining vital sign information (for example BP and Temp), these diaries were merged into a single diary. [code: sd]
- If two diaries were received on the same day and only one contained a complete set of vital sign measurements, the incomplete diary was discarded. [code: rep].
- If more than one complete diary was received on the same day, all diaries were included.
- If a midwife entered the incorrect Participant ID (usually communicated by email to the technical team), the ID was corrected. [code: id]
- Vital sign diaries that had been recorded as a duplicate (from 23 October 2015 to 1 March 2016) were omitted from the data set.
- If women recorded an *antenatal* home monitoring measurement, these were included. (Antenatal monitoring is usually performed by the midwife; women are

given a self-monitoring kit at their final antenatal appointment in preparation for postnatal monitoring. They occasionally started self-monitoring before delivery. There were 70 vital sign diaries submitted by 34 women in the antenatal period.)

- Temperature measurements made by the Fora Bluetooth thermometer were excluded from analysis. The Fora thermometer was only used in system 0 (alongside the Genius), whereas information from the Genius was available for systems 1, 2, and 3. For consistency we only consider measurements from the Genius.
- If a repeat BP reading was performed, only the second BP measurement was considered for analysis. (The 4P tablet app automatically requested a repeat reading if a BP measurement was high: SBP > 140 mmHg or DBP > 90 mmHg).
- Quality of the pulse oximeter trace was analysed to ensure good quality HR and SpO₂ measurements. Poor-quality measurements were discarded. This process is described in Section 3.4.

Antenatal diaries

During the antepartum period, 3,673 vital sign diaries from 620 women were collected by the midwives during antenatal home visit appointments. The distribution of diaries across weeks of gestation is presented in Figure 3.10 below. There is a notable peak at 12 to 13 weeks when most women are recruited at their first antenatal ultrasound scan, and again at 20 weeks when midwives were able to see the women in hospital for their second ultrasound scan.

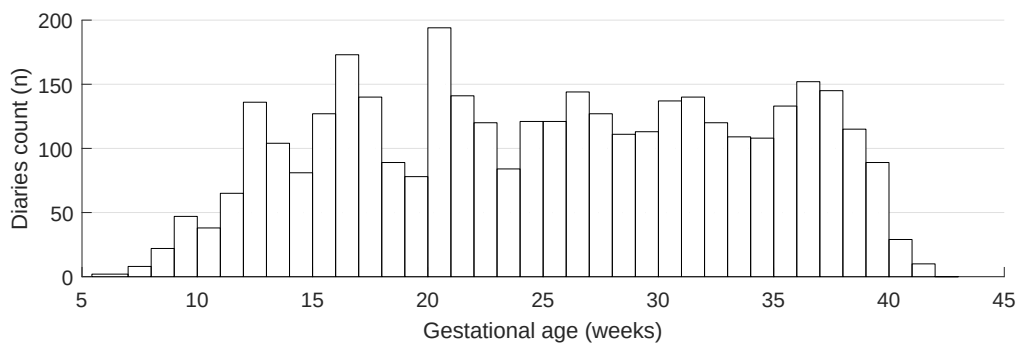


Figure 3.10: *Number of vital sign diaries recorded per week of gestation.*

Each woman had a median (IQR) of 6 (1) vital sign diaries recorded by midwives during the antenatal period (Figure 3.11). Women who withdrew from the study during the antenatal period often had fewer readings than women who completed this stage.

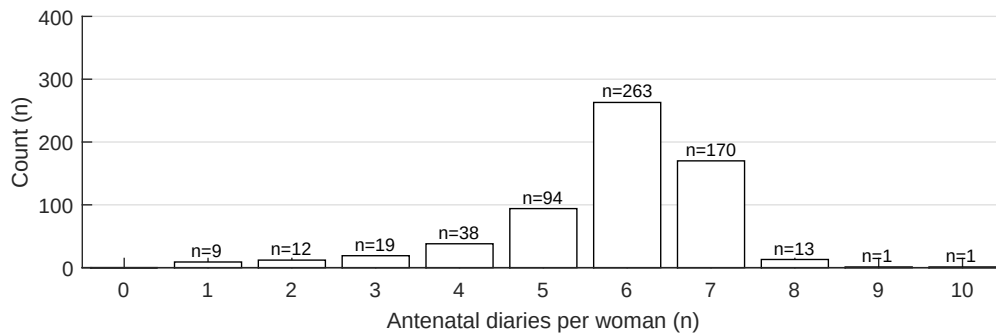


Figure 3.11: *Number of antenatal vital sign diaries recorded per woman.*

Postnatal diaries

In the postnatal period, 6,327 vital sign diaries were recorded from 538 women, mostly through self-monitoring. Most measurements were received between days 1 and 14 after delivery, as requested in the 4P study protocol. Sometimes midwives could not arrange their final home visit by day 14, so measurements were made later when possible. Figure 3.12 shows the distribution of vital sign diaries after delivery. The volume of vital sign diaries is much higher in the postnatal period than in the antenatal period (around 400 diaries per day after delivery compared to 100 per week of gestation) due to the self-monitoring measurements contributed by the women.

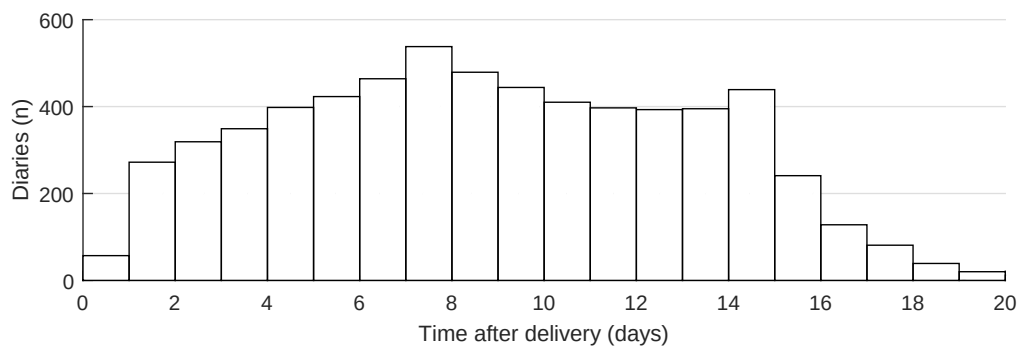


Figure 3.12: *Number of vital sign diaries recorded per day after delivery.*

Midwives usually performed two postnatal home visits. Women were usually compliant with postnatal self-monitoring at home. Of the 15 measurements requested (days 0 to 14 after delivery), women submitted a median (IQR) of 11 (4) diaries. Most women made between eight and 14 recordings. Figure 3.13 shows the number of vital sign diaries per woman, as recorded by the midwife and by the woman herself. Only women who self-monitored ($n = 523$) are included in the plot of measurements by the woman (15 women declined self-monitoring; see Figure 3.9). 76 women did not have any postnatal measurements by a midwife. 54 of those women withdrew during the postnatal period. The other 22 are missing because the women moved out of the area after delivery, and so could not receive midwife visits (however, they still performed self-monitoring), or the midwife did not have access to her own kit during the visit so used the participant's kit instead.

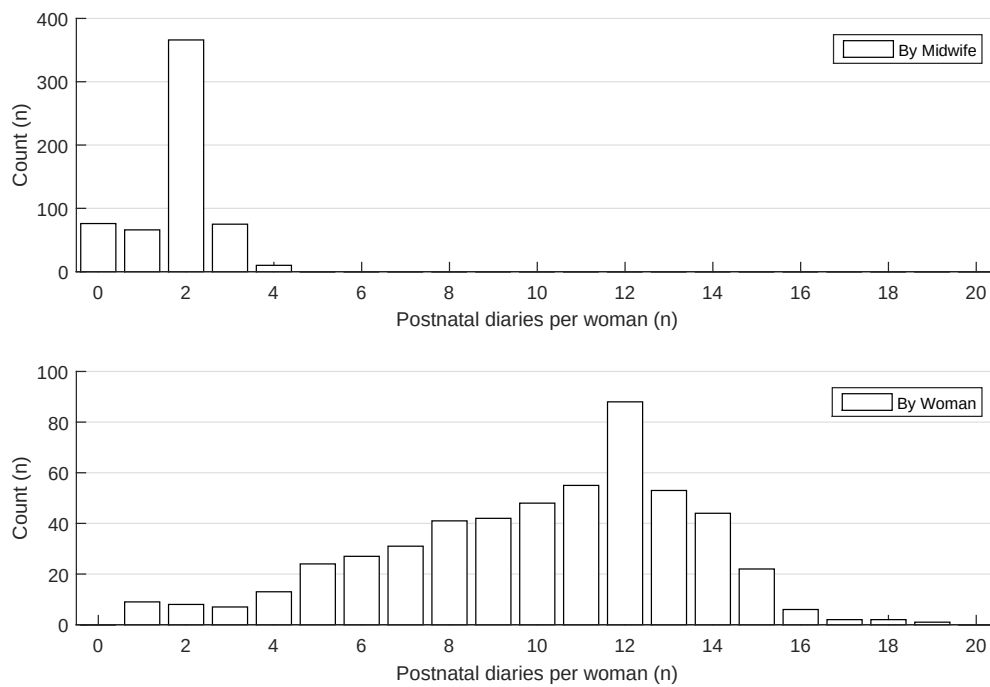


Figure 3.13: Number of postnatal vital sign diaries recorded per woman. **Top:** Measurements by midwife during home visits. **Bottom:** Self-monitoring measurements by woman. Note the different vertical axes between the top and bottom graphs.

Complete diaries

A vital sign diary was marked as *complete* if all vital signs were recorded (except for diaries submitted by the women at home where RR was not required). 2,930 (81.3%) diaries in the antenatal period, and 5,680 (89.8%) diaries in the postnatal period were complete. The vital sign missing the most from antenatal diaries was temperature because at the beginning of the 4P study, the brand of thermometer was still under discussion so some measurements were missed. In the postnatal period, the pulse oximeter measurement was most often missing because some women struggled to get the device to connect to the tablet via Bluetooth (despite clear instructions). Few RR measurements were recorded in the postnatal period compared to other vital signs because women were not able to record RR using self-monitoring.

The total number of vital sign diaries recorded by the midwife and women, at each stage of pregnancy, are shown in Table 3.8.

	Antenatal		Postnatal	
	Midwife n = 3603	Women n = 70	Midwife n = 1063	Women n = 5264
HR/SpO ₂	3,385 (93.9%)	66 (94.3%)	1,033 (97.2%)	4,895 (93.0%)
BP	3,584 (99.5%)	69 (98.6%)	1,063 (100%)	5,211 (99.0%)
Temp	3,091 (85.8%)	68 (97.1%)	1,051 (98.9%)	5,054 (96.0%)
RR	3,503 (97.2%)	-	1,035 (97.4%)	-
Complete diaries	2,866 (79.5%)	64 (91.4%)	998 (93.9%)	4,682 (88.9%)

Table 3.8: Number of diaries which contained each vital sign. Complete diaries are those where all vital signs were recorded.

3.4 Quality of the pulse oximeter recordings

In the 4P study, HR and SpO₂ are measured by the Nonin Bluetooth finger pulse oximeter (WristOx₂ 3150, Nonin Medical, Minnesota, USA). The time series signal obtained from the pulse oximeter is called the photoplethysmograph (PPG). A typical PPG trace

is displayed in Figure 3.14. The pulsatile component corresponds to cardiac changes in blood volume. The secondary, smaller peak is the dicrotic notch, and reflects the elasticity of the arteries and pulse velocity. This notch disappears with age as arteries get stiffer and pulse velocity decreases.

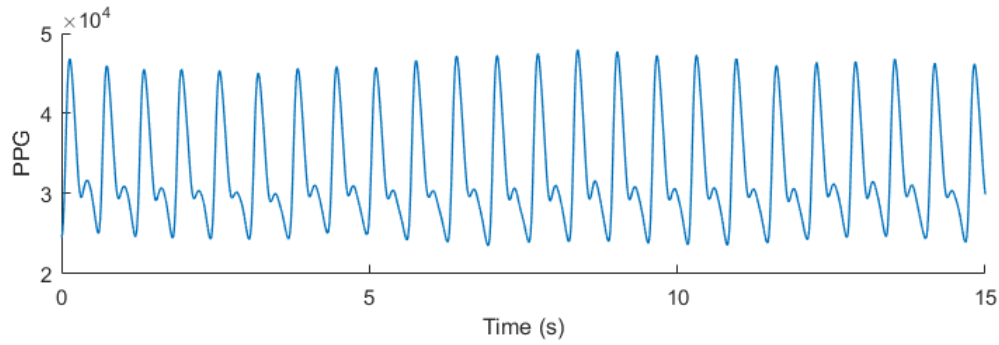


Figure 3.14: *A typical PPG waveform from the pulse oximeter.*

The quality of the PPG waveform is often compromised due to its sensitivity to motion, ambient light, and poor blood perfusion [77]. Such artefacts give rise to errors in the calculation of HR and SpO₂ and can lead to false alarms in clinical environments. It is therefore important to identify and deal with poor quality sections of the PPG signal before analysis. This is particularly vital for home monitoring recordings where healthcare professionals are not available to check the PPG waveform quality or the positioning of the pulse oximeter finger probe. In this section we discuss the pulse oximeter hardware, previous solutions for managing PPG artefacts, and describe a Signal Quality Index (SQI) applied to the PPG for exclusion of poor quality sections.

3.4.1 Pulse oximeter hardware

A pulse oximeter consists of a light source (commonly an LED) and a light detector (photodiode), where the detector is either placed directly across from the light source (transmission plethysmography) or next to the light source (reflectance plethysmography) [51, 78].

Haemoglobin that is fully saturated with oxygen (HbO₂) has a different absorption

coefficient to de-oxygenated blood (Hb). The difference in absorption changes with the wavelength of light (see Figure 3.15). Blood oxygen saturation levels can therefore be measured using the difference in absorption between two wavelengths of light, typically red and near infra-red [79]. Most commercial devices only display the near infra-red signal since it is less sensitive to changes in Hb saturation and therefore more stable over time.

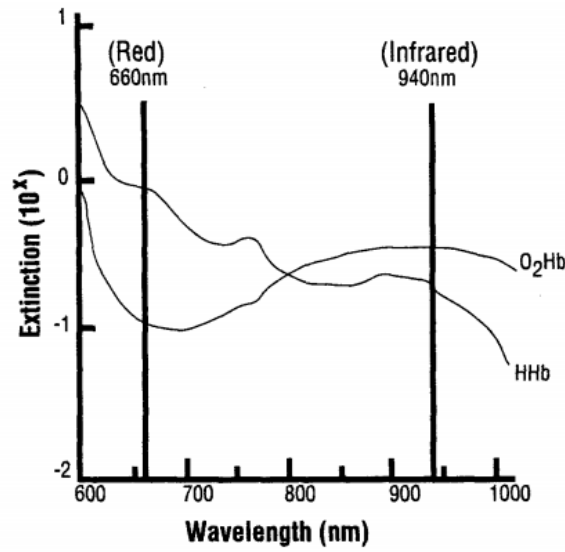


Figure 3.15: *Hb and HbO₂ absorption spectrum. Reproduced from [51].*

The Nonin WristOx₂ 3150 finger pulse oximeter uses transmission plethysmography. The time series of the near infra-red signal, the photoplethysmograph (PPG), is sampled at 75Hz and is available for download from the device. HR and SpO₂ are internally calculated, and also output at 75Hz. The 4P tablet application updates the values of HR and SpO₂ every 0.33s (approximately every four beats) as the median of the values returned from the pulse oximeter since the last update, rounded to the nearest whole number. The data available for analysis are therefore a PPG time series (75Hz) and a median HR and SpO₂ time series (3Hz) for each recording.

3.4.2 Previous methods for artefact suppression

Motion artefacts are the most common form of inaccuracies in PPG signals. The pulsatile component of the PPG is about 0.1% of the total signal amplitude, so even small movements will disturb the signal [80]. There are two main approaches for dealing with motion artefacts in the PPG waveform: the signal is either reconstructed to eliminate the effects of motion artefacts, or the artefacts are detected and removed from further processing. Techniques that suppress the motion artefacts often require a reference signal, such as an accelerometer, to provide input to an adaptive filtering scheme [81]. Independent Component Analysis (ICA) methods require both the near infra-red and red PPG waveforms [82]. Singular Value Decomposition (SVD) [83] and the Discrete Wavelet Transform (DWT) [80] have also been used to reconstruct the PPG signal without artefacts. Methods for detection of artefacts include Gaussian filters with cross-correlation [84], Dynamic Time Warping (DTW) [85], scanning for abnormal pulse periods [86], and detecting artefacts based on Hjorth parameters [87].

3.4.3 Signal quality index for PPG

Motion artefacts in the pulse oximeter compromise the accuracy of the calculated HR and SpO₂. The Nonin WristOx₂ 3150 finger pulse oximeter returns a series of flags that indicate when the PPG recording is not reliable or considered to be of poor quality. These flags are useful in discarding poor quality sections. However, they do not identify all types of motion artefact, so further detection is required. As the pulse oximeter recording may only be 60 seconds long (4P systems 1 and 2, and all home monitoring), it is important to identify motion artefacts and correct or remove them before calculating the value of HR and SpO₂ for that reading.

We developed a Signal Quality Index (SQI) using Dynamic Time Warping (DTW) to identify and remove sections of poor quality PPG data. Given our application of

obstetric monitoring, where vital sign observations are not usually time-pressured, we preferred a technique that removes poor-quality sections. In prospective applications, recordings would therefore be repeated until a good-quality signal is obtained.

The SQI was formed in three steps:

1. SQI(Invalid): Reject sections of invalid data
2. SQI(Flags): Reject intermittent poor quality sections marked with flags
3. SQI(DTW): Reject beats determined as poor quality by DTW

The first two indices are based on metrics returned by the Nonin pulse oximeter device directly. The third index supplements these by mapping each beat individually to a template, in order to identify poor-quality beats missed by the device. Each index indicates whether the segment is good quality ($SQI = 0$) or poor quality ($SQI = 1$).

Where poor-quality PPG sections are identified, the corresponding sections of the HR and SpO_2 time series are also marked as poor quality and excluded from analysis. The final HR and SpO_2 values are the mean of good-quality sections in the HR and SpO_2 time series respectively. If less than 30 seconds of good-quality signal remains, the entire signal is discarded and no vital sign execution is performed.

1. SQI(Invalid): Sections for which HR and SpO_2 could not be calculated by the pulse oximeter device were marked as invalid ($SQI(Invalid) = 1$). These were characterised by error values ($HR = 511$ bpm; $SpO_2 = 127\%$).

2. SQI(Flags): Sections with intermittent segments of poor-quality data were indicated by flags generated by the pulse oximeter device. If any flag occurred, the segment was marked as poor quality ($SQI(Flags) = 1$). The Nonin flags and $SQI(Flags)$ are described in more detail in Appendix B.2.

3. SQI(DTW): An SQI based on DTW ($SQI(DTW)$) was implemented to supplement the identification of poor-quality PPG segments by the metrics calculated by the pulse oximeter device ($SQI(Invalid)$ and $SQI(Flags)$). We use DTW to fit the individual beats to a template, specific for each PPG signal. If the beat has a poor correlation with the

template, even after DTW, we consider it poor quality. The SQI(DTW) algorithm is described in Appendix B.3.

3.4.4 SQI applied to 4P data

The SQI strategy developed here was effective in identifying poor quality sections of the PPG signal. An example PPG trace is shown in Figure 3.16 where poor quality sections are highlighted. All beats that appear poor quality (irregular) have been successfully highlighted.

In the 4P study, 10,000 vital sign diaries were submitted in the antenatal and postnatal stages of the study, and of these, 9,793 (97.9%) had PPG signals recorded. 101 PPG recordings were less than 30 seconds in duration (mostly recorded by women in the postnatal period), so no further signal quality assessment was performed on these. After each stage of the SQI algorithm, signals were excluded from further analysis if less than 30 seconds of good-quality signal remained. Thus, 30 recordings were excluded after SQI(Invalid); 81 excluded after SQI(Flags); and a further 200 excluded after SQI(DTW). Of the 311 poor quality recordings excluded by the SQI stages, 118 (37.9%) were recorded by the women at home. 9,381 (93.8%) of PPG recordings were good quality (antenatal, $n = 3,581$; postnatal, $n = 5800$). The flow chart in Figure 3.17 indicates how many recordings were excluded at each step in the SQI process.

The SpO₂ and HR values associated with discarded PPG recordings are shown in Figures 3.18 and 3.19 respectively. After each SQI assessment, the vital sign values are recalculated from the remaining good-quality sections. Note that before step 1 of the SQI, the SpO₂ and HR displayed are the original medians returned by the Nonin device and are therefore rounded to the nearest whole number (and, where recordings were empty or files not received, a value can still be displayed). Our subsequent calculations return the mean. The SQI strategy excluded 26 of the 46 low (<90%) SpO₂ values.

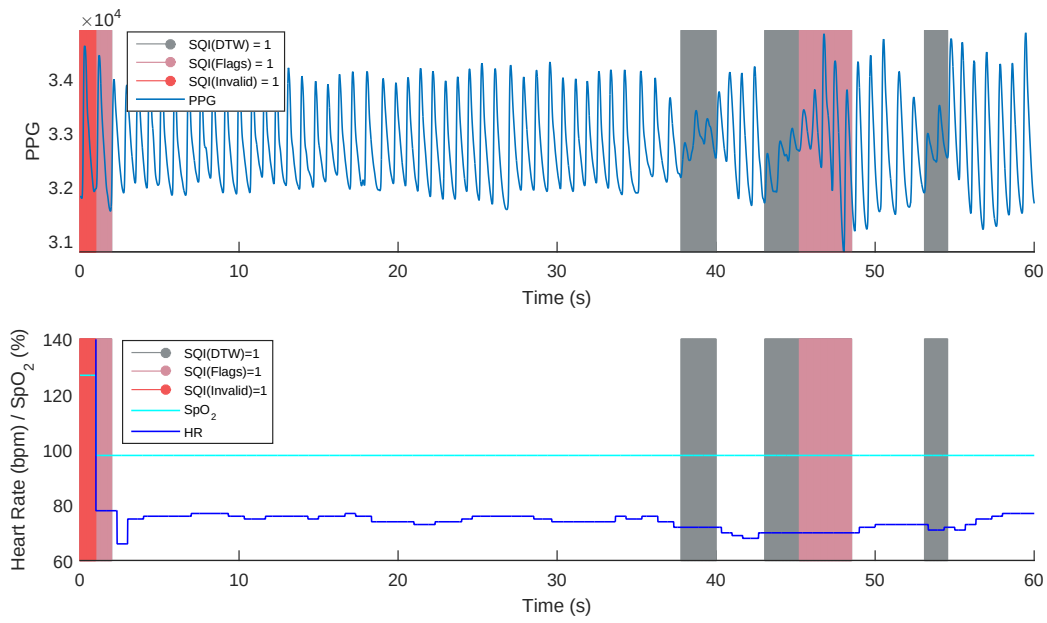


Figure 3.16: Example of SQI highlighting poor quality sections in a PPG recording.

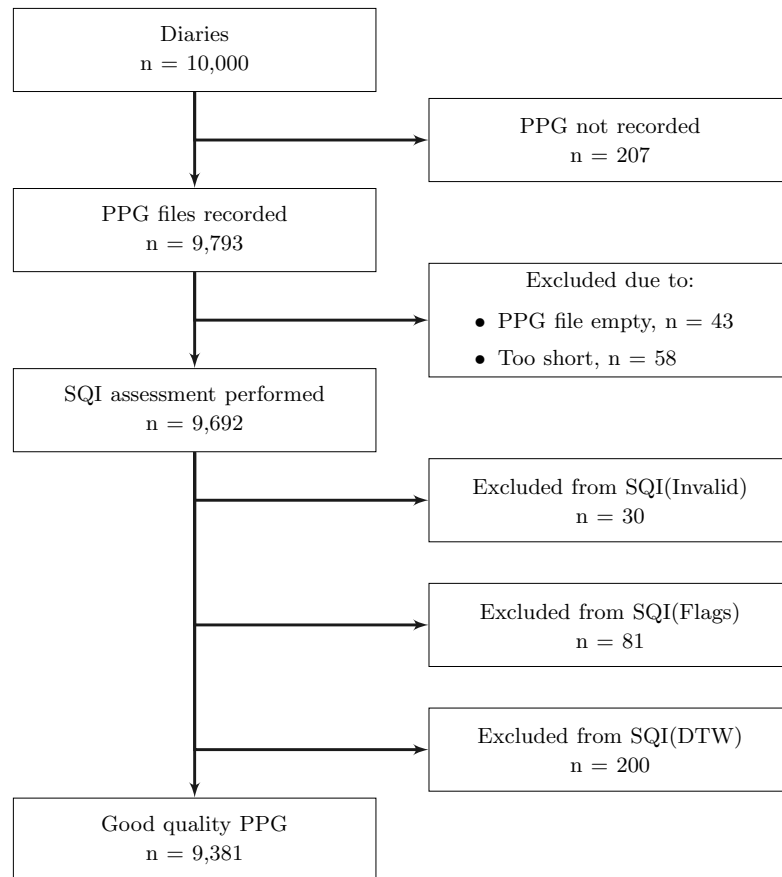


Figure 3.17: Flowchart for exclusion of poor-quality PPG recordings using the SQI algorithm.

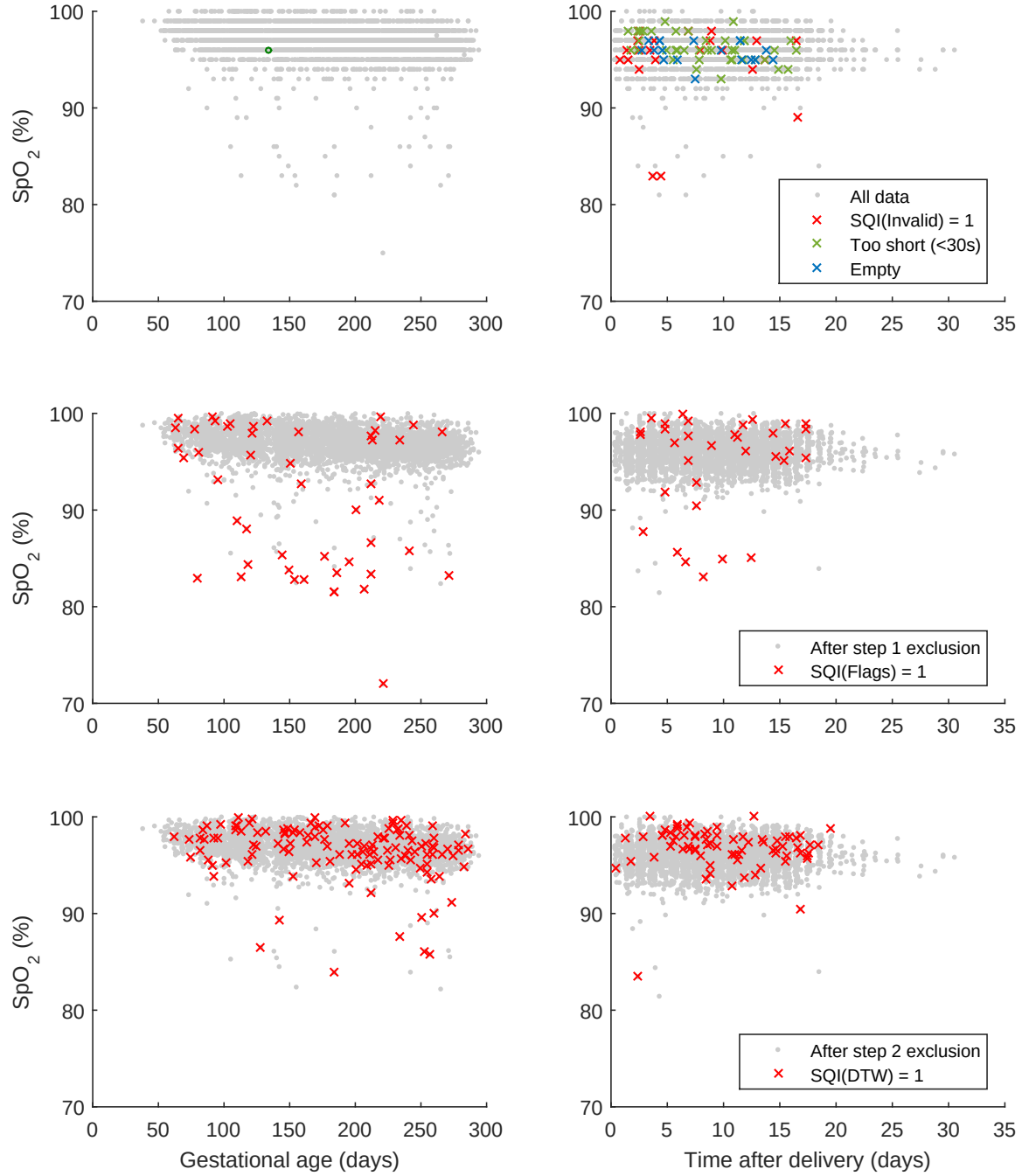


Figure 3.18: Antenatal (left column) and postnatal (right column) SpO_2 values excluded due to poor quality. After each stage of the SQI algorithm (top, middle and bottom plots), SpO_2 is recalculated from the good-quality sections of the recording.

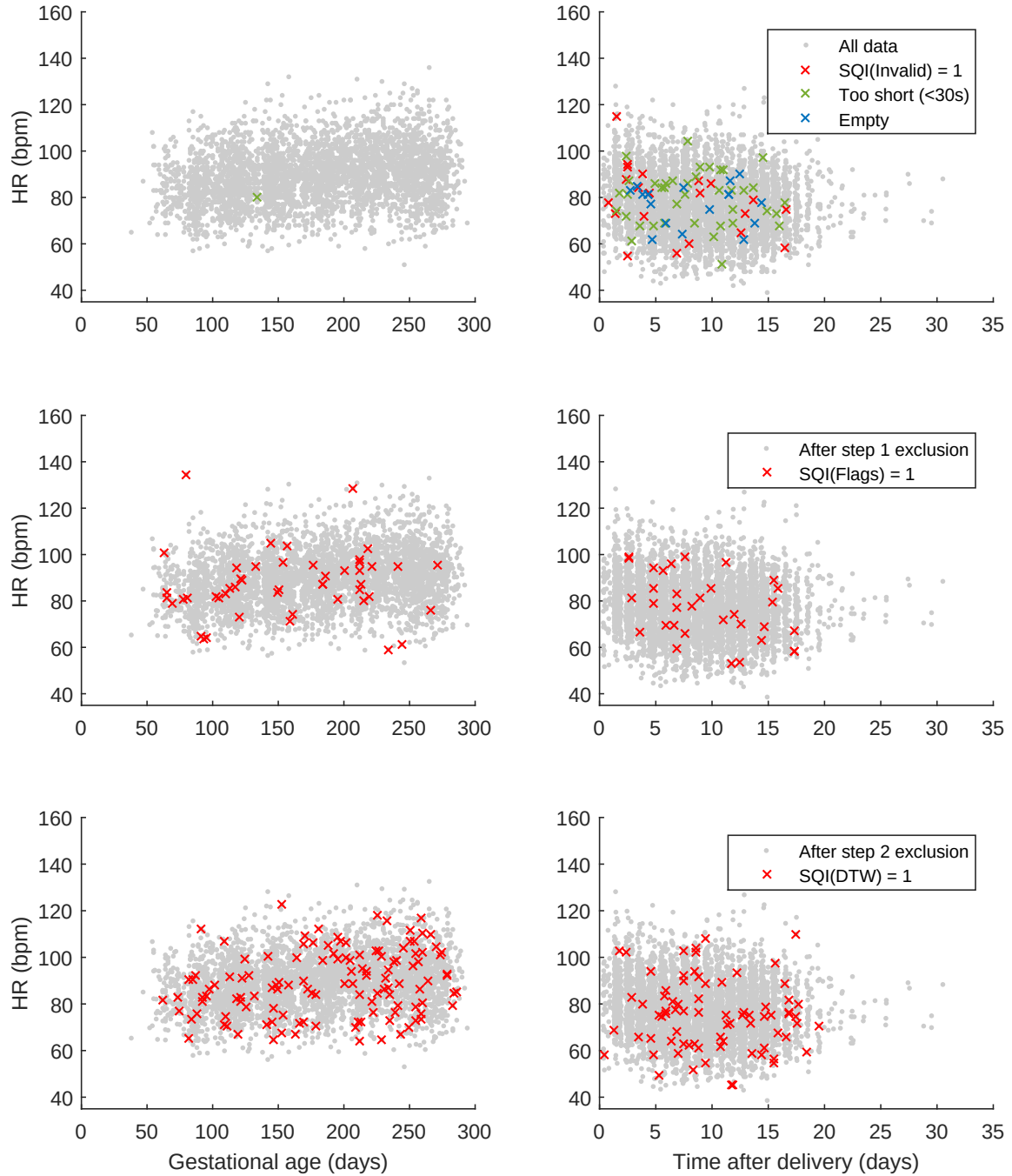


Figure 3.19: Antenatal (left column) and postnatal (right column) HR values excluded due to poor quality. After each stage of the SQI algorithm (top, middle and bottom plots), HR is recalculated from the good-quality sections of the recording.

3.5 Conclusion

The 4P study aimed to collect information on maternal vital signs and other pregnancy-related information from women experiencing a normal pregnancy. The recruitment target of 1,000 women was met in December 2016, and now the final stages of follow up and data cleaning are underway. The 4P data collected from women who had delivered by July 2016 will be used in the following chapters of this thesis.

An essential check for any dataset is to ensure good-quality data. In this chapter we examined the PPG waveforms from the pulse oximeter from which HR and SpO₂ are calculated. These waveforms are particularly sensitive to motion artefacts. We demonstrated how dynamic time warping could be used for beat-to-beat comparison to identify, and subsequently remove, artefactual beats in the PPG waveform. In the next chapter, we use the good-quality PPG waveforms and the smartphone accelerometer waveforms to derive a robust method for estimation of RR using signal-processing techniques.

Respiratory rate estimation

Respiratory rate (RR) is one of the best physiological indicators of patients at risk. An increase in ventilation reflects a lack of oxygen and is often the first sign of distress for respiratory diseases or complications such as hypoxia (oxygen deficiency), pneumonia (inflammation of the alveoli), or asthma (inflammation of the airways of the lungs). Relative changes in RR are much greater than changes in HR or SBP, and is therefore an important indicator of physiological deterioration [53, 88].

However, RR is the only vital sign that is not routinely measured electronically. The WHO recommends manual observation of respirations for 60 seconds [89]. In standard maternity practice, the midwife observes chest wall movement for 15 or 30 seconds, then multiplies to get an estimate of RR in respirations per minute (rpm). Several studies have demonstrated the inaccuracy of this practice, such as Lovett *et al.* [8] who compared respiratory rates from triage nurses in Emergency Departments (counting respiration for 15 seconds) to the gold standard measurement (counting respiration for 60 seconds) and found a low correlation between the two methods [90]. The burden of manual measurement may explain the poor accuracy and indeed the cases where RR measurement is altogether missed out [42].

This chapter focuses on the derivation of RR from two non-invasive devices, the pulse oximeter (introduced in Section 3.4) and a smartphone accelerometer, and on the

assessment of their suitability for robust RR measurement during pregnancy using data from the 4P study. We suggest that a robust measurement is one that is more precise than the current standard manual measurement, and one that performs consistently across the entire range of respiratory rate values.

4.1 Mechanics of breathing

There are several effects of respiration that can be measured. These include: airflow through the nasal passages during respiration, often measured externally at the nose and mouth; vibrations of vessels in the chest as air passes through; physical movement in response to the expansion and contraction of the lungs during breathing; and changes in heart rate frequency due to an autonomic response from stretch receptors in the lungs. In this section we describe the mechanical effects and autonomic responses to breathing, as they can be measured using non-invasive instrumentation.

4.1.1 Movement detection

The most visible change during breathing is the mechanical expansion and contraction of the thorax (rib cage) as the lungs physically fill with air and then deflate. This movement is what nurses observe when counting respirations, and can be measured by any device that quantifies movement or displacement, such as an accelerometer, strain gauge, or impedance method.

Movement of the thorax also affects the intrathoracic pressure which in turn alters peripheral blood volume, and therefore pulse strength [91]. These latter changes can be monitored with a pulse oximeter device as amplitude variation in the PPG waveform.

4.1.2 Heart rate variability

The variability of HR is widely used to investigate neural regulation of the heart [92]. The heart rhythm is mediated by the sinus node in the left atrium of the heart, which

is innervated by both the sympathetic and parasympathetic autonomic nerve divisions. The activation of either sympathetic or parasympathetic nerves is accompanied by an inhibition in the other: parasympathetic activation acts rapidly to slow HR, while sympathetic activation acts slowly to cause an increase in HR and heart contractility [93]. Since these mechanisms occur at different frequencies, it is possible to identify and quantify the sympathovagal balance in the frequency domain.

Heart rate variability (HRV) is the analysis of the variations of the intervals between consecutive normal heart beats. The human HRV frequency spectrum contains three components: a high frequency (HF), low frequency (LF) and very low frequency (VLF) [93, 94]. Figure 4.1 shows an example HRV spectrum, derived from the pulse oximeter recording of a 4P woman at 31 weeks gestation.

It is generally accepted that the HF component (~ 0.25 Hz) reflects RR due to Respiratory Sinus Arrhythmia (RSA). RSA is the cyclical variation in HR due to respiration: during inspiration, stretch receptors within the lungs send inhibitory projections to the cardiac vagal neurons (parasympathetic), resulting in an elevation of HR; and the reverse occurs during expiration [91]. RR can therefore be estimated from the HR component of the HRV spectrum. The LF component (~ 0.1 Hz) is thought to be either due to cardiac sympathetic activity, or the regulation of cardiac activity through baroreflexes [93, 95]. The VLF component is not well characterised, thought to be either from changes due to thermoregulatory mechanisms or peripheral chemoreceptors [96].

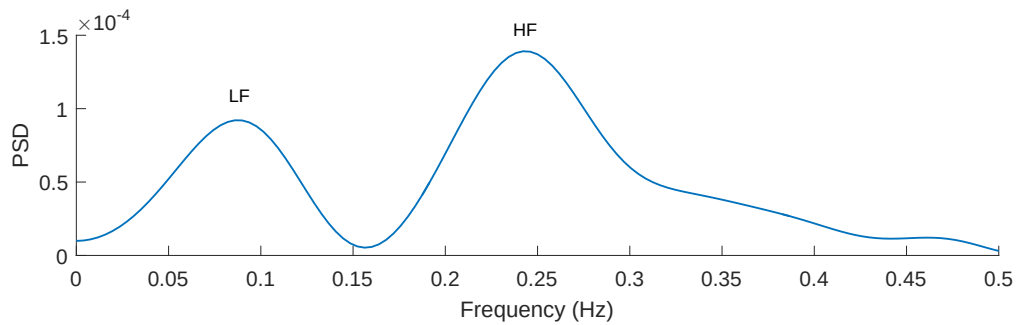


Figure 4.1: An example of the HRV spectrum with the high frequency (HF, corresponding to RR), and low frequency (LF) components indicated.

4.2 Related work

Methods of automatically monitoring respiration are usually either devices that measure the flow of air (eg. facemask or nasal cannula), or observe the movement or tissue composition of the abdomen or chest (eg. chest band). Many of these methods are obtrusive in nature, require time to set up and are poor in accuracy, which has led to a search for simpler, low-cost devices that would be suitable for remote monitoring. In this section we investigate previous attempts to derive RR from the pulse oximeter and the accelerometer, two devices which are non-invasive and readily available.

4.2.1 Respiratory rate from accelerometers

Accelerometers measure respiratory effort by using spatial information. Although accelerometers are starting to be utilised to derive RR and HR, this remains primarily an area of research. The earliest studies used uni-axial (single direction) or bi-axial (in two directions) accelerometers [97]. Tri-axial accelerometers (measured in three directions, usually orthogonal $x - y - z$ axes) are now standard, typically fastened to an elastic band around the subject's chest (for example, see Figure 4.2a). Other locations include a belt around the abdomen (Figure 4.2b) [98], an adhesive patch on the chest [99], or as a wrist-worn device [100].

Before estimating RR, a single respiratory axis needs to be chosen or extracted from the three-orthogonal axes available. Simpler studies choose one axis for analysis [101, 102, 100], while others project the three axes onto a single respiratory axis using Principal Component Analysis (PCA) [103, 98] or previous observation of a maximum signal [104]. The axes can also be summed due to their orthogonal nature [105]. RR derivation techniques are often straightforward, using peak detection or frequency domain analysis. Related to the accelerometer are gyroscopes, often coupled together in the hardware of mobile devices. While accelerometers measure acceleration in a certain direction,

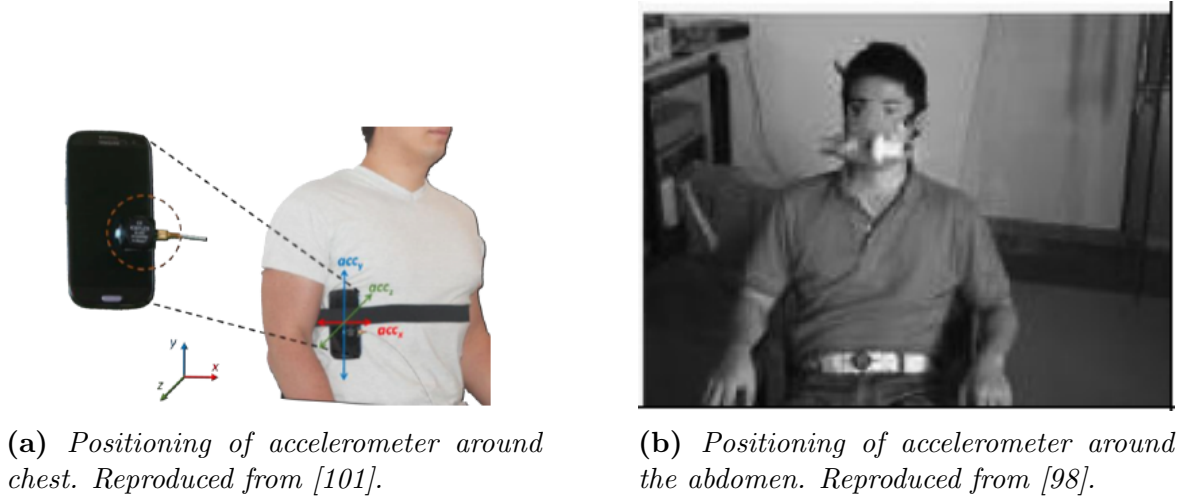


Figure 4.2: Examples of accelerometer positioning to measure RR.

gyroscopes measure rotation around an axis (also typically orthogonal $x - y - z$ axes) from which RR can also be derived.

Accelerometers in-built into everyday mobile devices are now being used for RR estimation. Recently, Estrada *et al.* (2015) [101] strapped a smartphone (Samsung Galaxy S3) to the chest and compared the z -axis accelerometer trace with a commercial uni-axial accelerometer aligned in the same direction (Figure 4.2a). RR was extracted using peak detection of the time-domain respiratory signal. There was high correlation between respiratory rates derived from the smartphone and reference accelerometers (limits of agreement: -1.44-1.46 rpm), indicating that the hardware in mobile devices is sufficiently robust for such a healthcare application. RR has also been derived from a wrist worn smartwatch (Samsung Galaxy Gear) [100]. This proof-of-concept study used subjects either in a controlled experimental environment (sitting, standing or supine positions), or overnight while asleep. It is therefore not surprising that RR estimates obtained from Fourier analysis were within 1 rpm of the gold standard chest belt. They also extracted RR from the in-built gyroscope, which was more accurate than RR extracted from the accelerometer. A particular challenge with wrist-worn devices is elimination of motion artefacts since the hands are rarely still. Most accelerometer-derived respiratory rates

discard high artefactual periods, which for wrist-worn devices, may inhibit their use in regular ambulatory care.

Surprisingly few studies report utilising the related gyroscope signal. Bates *et al.* [99] derived the rotation angle from accelerometers rather than using gyroscope hardware due to high power concerns. Hernandez *et al.* [100] used three accelerometer axes and three gyroscope axes, yet extracted the final respiratory waveform as the channel with the most periodic signal, not utilising any additional information provided by the gyroscope.

All of the studies mentioned here unfortunately suffer from small sample sizes, often 15 or less subjects; however they demonstrate how RR can be successfully derived from accelerometer and/or gyroscope waveforms.

4.2.2 Respiratory rate from the pulse oximeter

RR was first derived from the pulse oximeter in the early 1990s, originally inspired by the work of those researchers who derived respiratory activity from the electrocardiogram (ECG), a measure of the electrical activity of the heart. The delay in incorporating RR into commercial devices is partly due to the availability of more reliable RR measurement methods in clinical environments, such as spirometry or capnometry, and sensitivity to motion artefacts [86]. However, RR from the pulse oximeter remains a practical solution for remote or resource-poor locations where no other measurement techniques are readily available.

RR derivation from the PPG waveform first requires the extraction of a *respiratory signal*. Breathing causes frequency-modulation of HR by RSA as discussed in Section 4.1.2, and two types of amplitude-modulation caused by changes in intrathoracic pressure which affects peripheral blood volume (measured by baseline wander) and subsequently causes changes in pulse strength (measured by height of the PPG waveform). The extraction of these three respiratory signals from the PPG waveform (frequency-modulation, FM; baseline wander, BW; amplitude-modulation, AM) are illustrated in

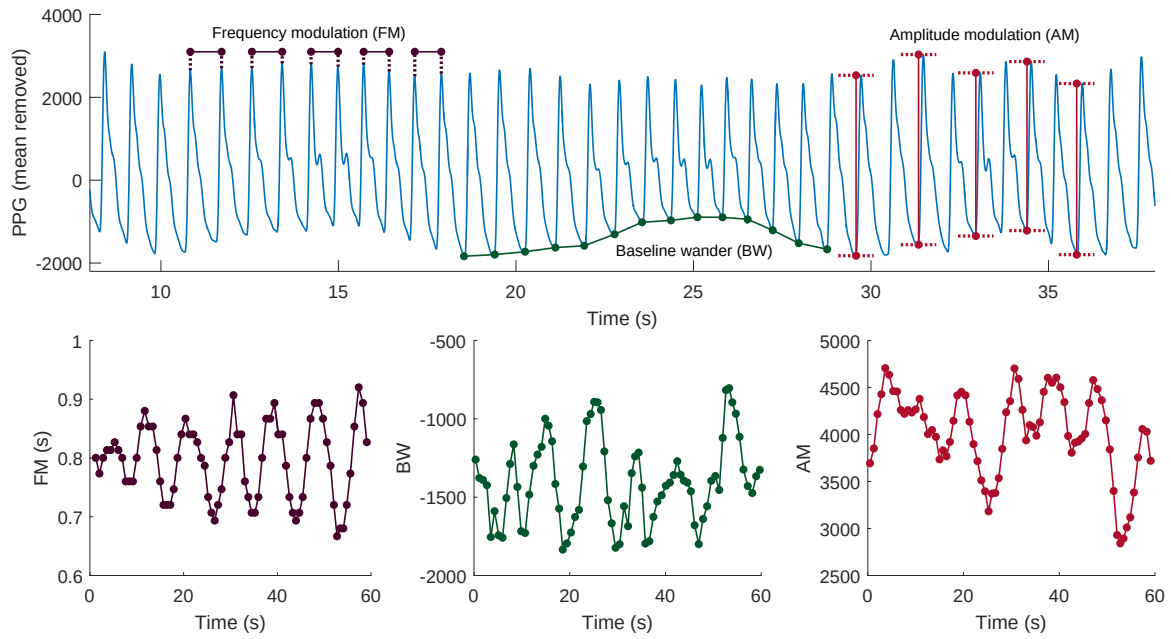


Figure 4.3: Illustration of how the three respiratory signals (FM, BW and AM) are extracted from the PPG waveform.

Figure 4.3 [106]. Most techniques to extract RR from the PPG utilise the amplitude-modulation effects of respiration [107, 108, 109, 110, 111, 112, 113, 114]. Fewer studies have used frequency-modulation [115, 116], however it is becoming more common to combine amplitude- and frequency-modulation techniques since the same signal processing techniques can be applied to both types of extracted respiratory signal [117, 86, 118].

The earliest methods to derive RR from the PPG waveform used digital filtering to identify amplitude changes. A mixture of band-pass and low-pass filters were used, usually with subsequent peak detection or manual annotation to estimate RR from the resulting respiratory waveform [114, 113, 119]. These methods performed adequately for feasibility studies, but digital filtering alone was superseded by more sophisticated techniques.

The wavelet transform decomposes a signal into a series of characterising signals at different frequencies. This has been a popular time-frequency analysis tool to estimate RR from PPG. Leonard *et al.* [120] used the Continuous Wavelet Transform with Secondary Wavelet Feature Decoupling to select the respiratory components. Their

method is now fully automated and results from 12 subjects breathing spontaneously in the seated position, gave errors of less than 0.5 rpm. Further studies with wavelets have also shown promising results [121, 117]. The main criticism of the wavelet transform in RR estimation is the difficulty caused by motion artefacts as their frequency content often overlays with RR frequencies.

Auto-Regressive (AR) models are also attractive because they only require a few cycles or fraction of a cycle to obtain a reliable frequency estimate. AR models have frequently been used to derive RR from PPG, with varying success. Fleming *et al.* [107] used an 11th-order AR model on 5-minute PPG segments from the MIMIC database¹ which performed favourably when compared to digital filtering and wavelet-based methods. In 2009, Chon and Lee [115] employed the same AR model and compared it with their own variable-frequency complex demodulation (VFCDM) method for estimating RR. VFCDM gave more accurate RR estimations than the AR model, however they did not use PPG records with spontaneous breathing. Chon and Lee have since revisited and further developed their AR model [110, 111], because an AR model can estimate high respiratory rates accurately and handle the dynamics of breathing transitions.

Other methods used to extract respiratory information from PPG include neural networks [109], PCA [122], the Wigner-Ville distribution [118] and Empirical Mode Decomposition [112].

4.2.3 Fusing respiratory rate estimates

Data fusion methods are approaches that combine multiple methods of deriving RR to select the most appropriate RR estimate. Usually this involves taking the mean or median of multiple respiratory estimates [118, 121].

In research with accelerometers, the dominant respiratory axis is usually extracted

¹A freely accessible critical care database where the PPG is recorded during spontaneous breathing. <https://mimic.physionet.org/>.

or chosen before estimation of RR. Thus only one rate is presented. In contrast, RR from the PPG waveform is usually calculated from each derived respiratory signal, such as the frequency-modulated or amplitude-modulated signal, before seeking to combine estimates to obtain a final, more accurate, RR. The underlying assumption is that the different modalities in which respiration affects the PPG are complementary. This is reasonable since amplitude and intensity variations are mechanical effects, and frequency-modulation is neurological.

Methods that take multiple features to obtain a single result are also instances of data fusion. One of the earliest examples of data fusion in RR is a neural network [109]. Here the authors extract 5 features from each beat in the PPG waveform: peak value, previous trough value, time since last peak, pulse amplitude, and intensity variation (from bandpass filtering). The features for 5 consecutive beats were presented to a neural network which determined whether the input followed inspiration, expiration or neither. A neural network allows a complex decision boundary without full knowledge of the statistical properties of the data. However, the lack of transparency and binary output also limits understanding.

The fusion method that has gained the most popularity is “smart fusion” proposed by Karlen *et al.* [86]. Three respiratory signals are derived from the PPG: frequency modulation, amplitude modulation and intensity variation (which we will refer to as *baseline wander*). The signals are resampled to 4 Hz and RR estimated as the largest magnitude frequency of the Fourier transform. Windows of data (32 seconds in length) are labelled “untrustworthy” if they contain artefacts or the three RR estimates have a standard deviation greater than 4 rpm. The final RR is fused only for trustworthy windows, calculated as the mean of the three respiratory estimates. Although smart fusion improved RR estimates, the main criticism of this data fusion approach is the high rate of data exclusion. 41.3% of windows were excluded due to being untrustworthy.

An improvement to smart fusion would be the incorporation of uncertainty about the estimates. Zhu *et al.* [123] propose a probabilistic Bayesian model, BCLA, to fuse RR estimates from different algorithms, particularly advantageous when the ground truth is unknown. Gamma and Gaussian distributions are included as priors and the model solved using Expectation-Maximisation where the hyperparameters are calculated as Maximum-a-priori. Results from BCLA were compared to those using “smart fusion” using the same dataset and pre-processing as described by Karlen *et al.* [86]. Respiratory estimates fused using BCLA were obtained for all data windows compared to “smart fusion” where 41.3% of windows were discarded, and had a lower mean absolute error. This study demonstrates how a probabilistic model can perform reliable fusion despite substantial noise.

4.2.4 Respiratory rate in pregnancy

To the best of our knowledge, the methods described here to derive RR from a pulse oximeter, accelerometer and/or gyroscope, have not been tested in pregnancy. There are two main reasons why pregnant women may not have been popular candidates for RR derivation. Firstly the physically growing chest and abdomen of the woman may inhibit the securing of devices used to monitor mechanical chest wall movement during respiration, such as accelerometers. Additionally, it is unknown how the cardiovascular adaptations during pregnancy affect respiratory-induced modulations of blood volume and HR. RR derived from the PPG waveform relies on the detection of these mechanical and neurological modulations, which may be reduced or emphasised during pregnancy.

Earlier work by Loerup [124] in our research group investigated RR in pregnancy, using data from system 0 of the 4P study (Section 3.2). Results were promising but suffered from the inaccuracy of the reference RR (based on a 15-second count of chest wall movement). The work described in this chapter builds on this earlier work.

4.3 Dataset

Respiratory measurements performed using system 3 of the 4P study were used to develop and validate an electronic RR algorithm. System 3 recordings were chosen because midwives counted respirations for 60 seconds, as recommended by the WHO. For each recording, the PPG waveform was available from the Nonin finger pulse oximeter, and tri-axial accelerometer and gyroscope waveforms were available from the Samsung Galaxy S3 smartphone placed on the woman's chest (Section 3.2.1). As described in Section 3.2.2, the pulse oximeter recording in the midwife version of the 4P tablet app went for the duration of a Manual HR count and the RR count. Only the section of the PPG waveform corresponding to the RR count was used for analysis (see Figure 4.4). The position of each breath was indicated midwife taps on the tablet keypad.

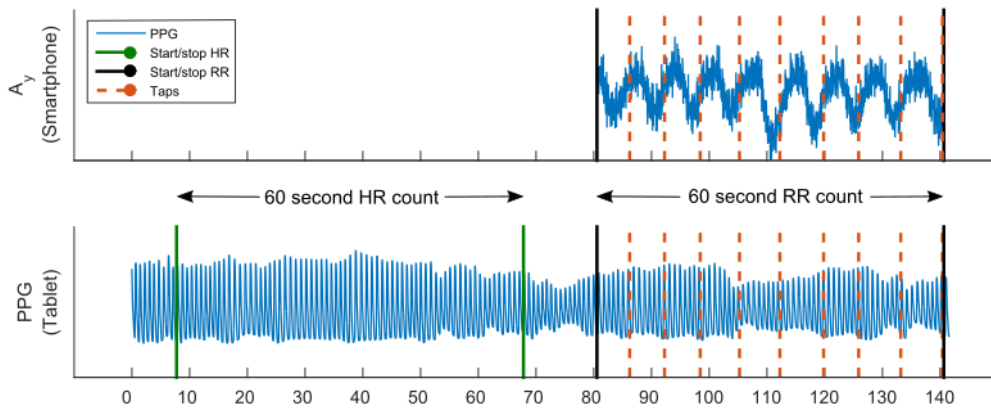


Figure 4.4: *Illustration of the alignment between the accelerometer/gyroscope waveforms from the smartphone (top) and PPG waveform from the pulse oximeter which is connected to the tablet (bottom). During the pulse oximeter recording, the midwives perform a Manual HR count before starting RR.*

From 7 July 2015 to 22 March 2016, there were 912 respiratory measurements recorded using system 3. 165 recordings were excluded because they were missing either the PPG waveform, accelerometer/gyroscope waveform, or the midwife taps indicating each breath. A further 8 recordings were excluded based on comments by the midwife about recording inaccuracies, and 42 more recordings excluded due to disagreements

between the RR recorded and the number of breaths indicated on the tablet. Duplicate recordings (two measurements on the same participant at the same appointment, $n = 256$) were not excluded. Therefore, a total of 697 recordings from 152 women were used for respiratory analysis (334 from Oxford; 2 from London; 361 from Newcastle). 595 of the measurements were recorded during the antenatal period, and 102 recorded during the postnatal period (see Figure 4.5).



Figure 4.5: *Distribution of recordings available for respiratory rate derivation in antenatal and postnatal periods.*

4.3.1 Reference respiratory rate

RR is typically recorded as a whole number of respirations per minute. As previously mentioned, during the 60-second RR measurement in system 3 of the 4P study, midwives tapped a keypad button on the tablet screen once per respiratory cycle. The midwives were not given any instruction about which stage of the breathing cycle to tap the tablet, but they were asked to be consistent in their choice. To validate the gold standard RR (counting respirations for 60 seconds), we compare the RR recorded by the midwives with a RR estimate derived from the tablet keypad taps.

Methodology

RR from keypad taps has recently been presented in the literature as a faster alternative to a 60-second count. Karlen *et al.* [125] present a free mobile application, *RRate*, which uses consecutive taps to estimate RR. The length of a breath (time between consecutive taps) must be compliant with those of the surrounding breaths, otherwise more breaths are required. For real-time monitoring, there is a trade-off between the number of breaths to be used for respiratory estimation (window length), and the accuracy of the final measurement. When compared to the official WHO timer, *RRate* measured RR with a similar accuracy but in less time [126].

We utilise a similar algorithm to calculate RR from the midwife taps (RR_{Taps}). Breath length, b_i , is defined as the difference between consecutive taps. The compliance of each breath, C_i (%), is measured by comparison of the breath length to the median breath length of that recording, b_m , described by:

$$C_i = \frac{|b_i - b_m|}{b_m} 100 \quad (4.1)$$

If the maximum compliance in the recording is larger than a pre-defined threshold, C_{th} , the breath length furthest from the median is excluded, and the median and compliance recalculated. This iterative process is continued until the compliance of all breaths within a recording are below the threshold. RR is then calculated in breaths per minute from the mean breath length. After the removal of outliers, we can assume that breath lengths are normally distributed due to human error and neurological modulation.

Results

We empirically chose a compliance threshold of 50%. This means, for example, a median breath length of 4 seconds (equivalent to 15 rpm) requires all breath lengths for that recording to be within ± 2 seconds of the median (equivalent to between 10 and 30 rpm),

otherwise they are discarded. A 50% compliance threshold was found to be sufficient to exclude unreasonable outliers. Values below this made little difference to the final RR since the majority of breath lengths were tightly clustered (median compliance of all breath lengths from all recordings before exclusions was 9.5%).

The Bland-Altman plot in Figure 4.6 shows the difference between RR_{Taps} and the 60-second original midwife count, RR_{60sec} . RR_{Taps} is an average of 0.8 rpm higher than RR_{60sec} . This is to be expected due to the quantised nature of RR reporting that causes manual measurements to be rounded down to the nearest breath. In Figure 4.7 the difference in measurements are grouped by RR_{60sec} . There is a large spread in RR_{Taps} for each group, such as at 15 rpm. The mean of each group is within 1 rpm of the midwife count, except for those groups at 12, 18 and 19 rpm.

The difference between RR_{Taps} and RR_{60sec} is greater than 2 rpm for 73 (10.5%) recordings. These are most often cases where taps have been missed at the start or middle of the recording, or the recording is poor quality. Consider the three example accelerometer traces in Figure 4.8 where the timings of the tablet keypad taps are indicated.

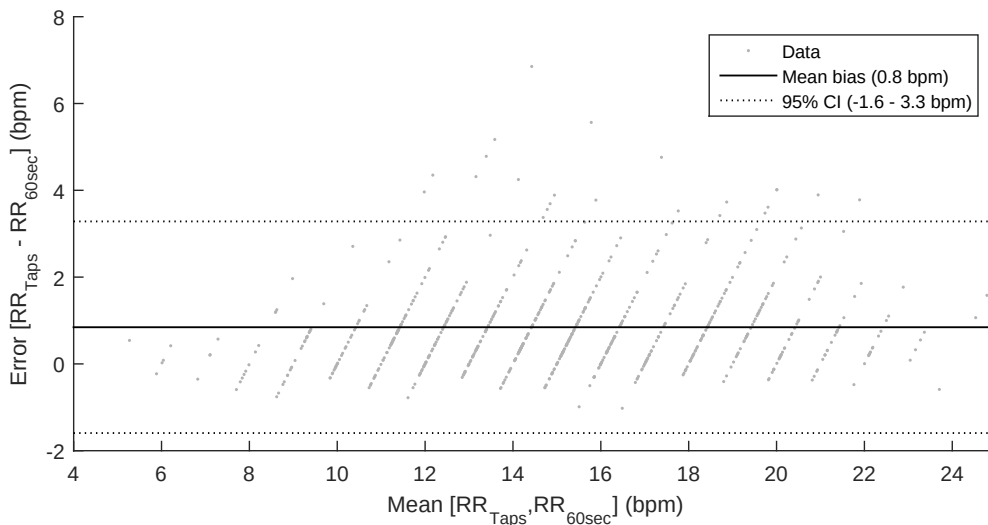


Figure 4.6: Bland-Altman plot of differences between respiratory rate estimates RR_{Taps} and RR_{60sec} . Note one data point not shown at $\{15.0, 18.7\}$.

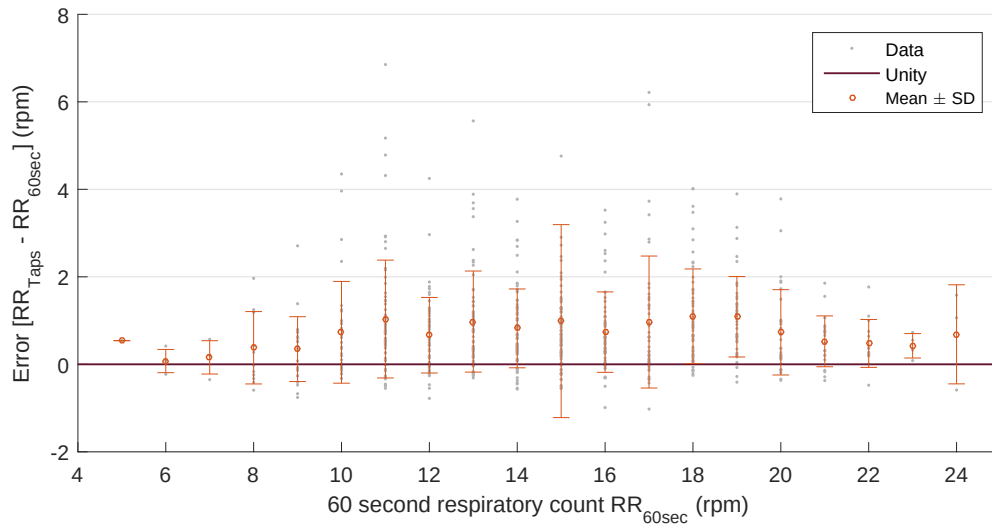


Figure 4.7: Difference between respiratory rate estimates RR_{Taps} and RR_{60sec} . Note one data point not shown at $\{15.0, 18.7\}$.

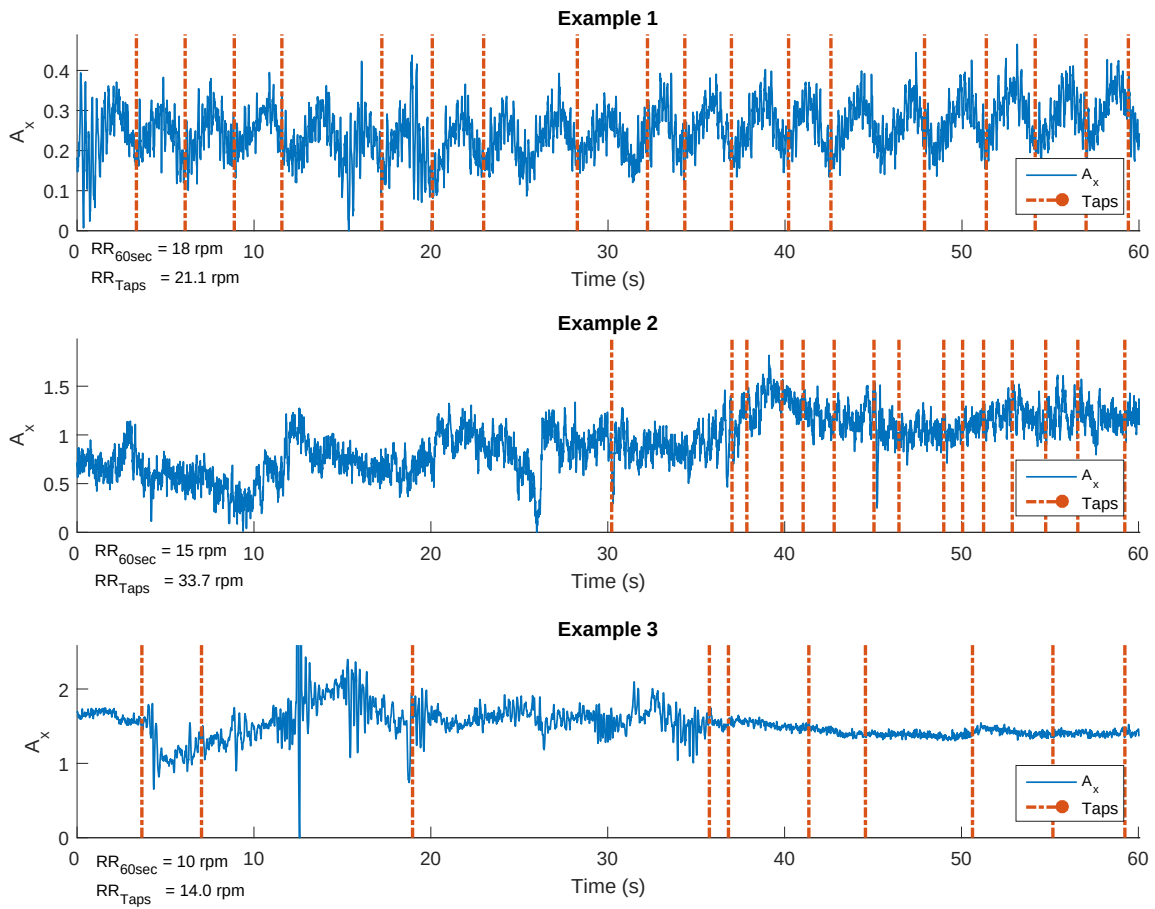


Figure 4.8: Three examples of inconsistent tapping patterns when recording respiratory rate. Respiratory rate recorded by the midwife (RR_{60sec}) is inconsistent with respiratory rate estimated from breath lengths (RR_{Taps}).

In the first example, visual examination suggests there are probably three missed taps, one at 15, 25 and 45 seconds. The midwife has recorded RR as 18 rpm, with 19 taps indicated over 60 seconds. Taking into account the probable missed taps, this RR is most likely incorrect, and a better estimate would be 21 rpm (22 taps). The RR_{Taps} agreed with this updated estimate (21.1 rpm), as the algorithm excluded the three longest breaths (where taps are most likely missed) from the RR calculation due to poor compliance. Had RR been calculated from the average breath lengths before the exclusion of outliers, the inaccuracy of the RR_{60sec} measurement would have been missed.

In the second example in Figure 4.8 the first tap is at 30 seconds, indicating the unlikely absence of breathing in the first half of the recording. From 37 to 59 seconds we observe 14 reasonably evenly-spaced taps (maximum breath compliance = 52.2%). Since the taps suggest 15 breaths occurred within the last 30 seconds, the midwife RR of 15 rpm is unlikely to be accurate. RR_{Taps} calculates the RR as 33.7 rpm which is high for a normal adult. The inconsistency between RR_{60sec} and RR_{Taps} suggest that this recording is not suitable for RR analysis.

The final example in Figure 4.8 presents a recording with few sparsely spaced taps. A total of 11 taps are indicated (10 rpm) with breath lengths ranging from 1.1 to 16.8 seconds (equivalent of between 3.6 and 54.0 rpm). RR accuracy is therefore questionable. The accelerometer trace is also noisy, suggesting the participant was talking or moving, further complicating the estimation of RR.

Reference respiratory rate distribution

We use the RR_{Taps} measurement as the reference RR to evaluate the RR estimates derived from the pulse oximeter and smartphone accelerometer waveforms. Those cases where RR_{Taps} does not agree to within 2 rpm of RR_{60sec} are excluded since this indicates some error with tapping or the RR reporting, and therefore the correct RR is not

known. RR_{Taps} thus provides a validated, non-quantised reference RR. The distribution of RR_{Taps} for the dataset considered in this chapter is shown in Figure 4.9. Respiratory rate ranges from 5.5 to 25.6 rpm with most measurements between 10 and 20 rpm.

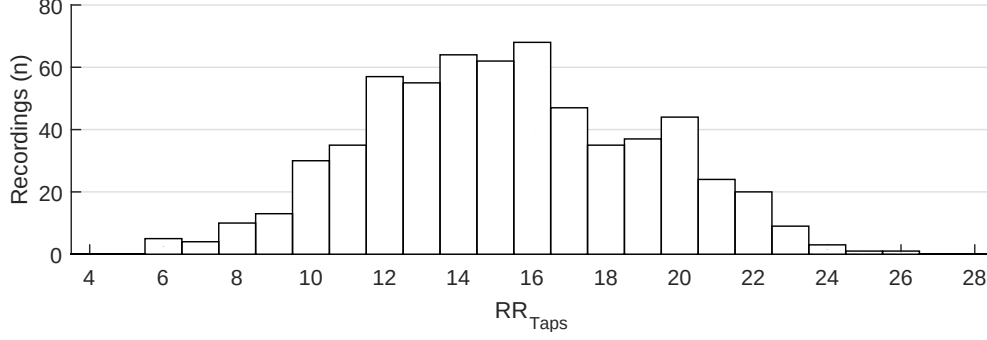


Figure 4.9: Histogram of validated respiratory rates

4.4 Respiratory rate algorithms

In this section we describe the methodology for deriving RR from the accelerometer and gyroscope waveforms, and the PPG waveform. RR can be directly obtained from the accelerometer and gyroscope signals after removing noise, since they directly measure mechanical movement during respiration. RR from the PPG waveform first requires extraction of respiratory information from either the HRV signal or the amplitude-modulation caused by indirect effects of chest wall movement (see Section 4.2.2).

4.4.1 The accelerometer and gyroscope waveforms

Accelerometer and gyroscope information is recorded from the in-built hardware in the Samsung Galaxy S3 smartphone, placed on a women's chest while she is slightly reclining (see Figure 3.4 in Section 3.2.1). We refer to the 3 orthogonal accelerometer axes as A_x , A_y , and A_z , and the 3 orthogonal gyroscope channels as G_x , G_y , and G_z . We use the waveforms of all six channels for respiratory analysis. The respiratory signal typically appears as baseline wander in accelerometer and gyroscope signals. HR information is often visible in the gyroscope signal, as shown in Figure 4.10.

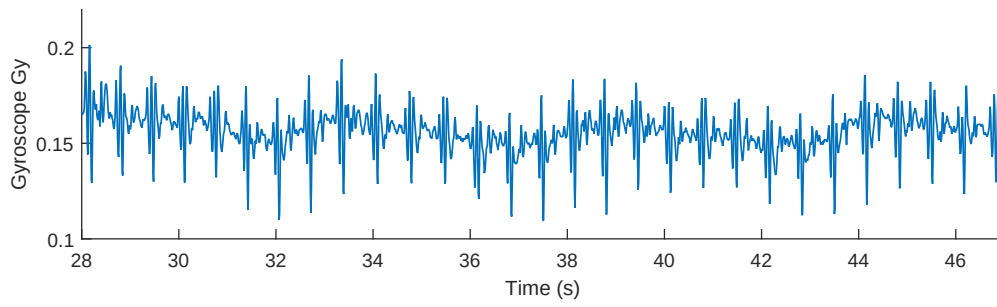


Figure 4.10: *Example of a gyroscope waveform (y-axis). Respiration is the underlying low-frequency waveform. The vertical spikes indicate heart rate.*

The Samsung Galaxy S3 samples the accelerometer at approximately 100 Hz, and the gyroscope at approximately 200 Hz. Both the accelerometer and gyroscope channels were resampled to 96 Hz before further processing. All signals were then low-pass filtered below the expected HR frequency and high-pass filtered to remove any baseline offset. All filters were Butterworth IIR filters, and each signal was filtered twice (forwards and backwards) to obtain a zero-phase shift².

Different filter limits were chosen for the accelerometer and gyroscope signals due to differences observed in the frequency spectra. The gyroscope channels contained more high-frequency noise (above 0.4 Hz) than the accelerometer channels, although in both the gyroscope and accelerometer signals it was apparent that the dominant respiratory frequencies were between 0.08 and 0.40 Hz (4.8 to 24.0 rpm) (Figure 4.11). To ensure sufficient attenuation of the high-frequency noise in the gyroscope signal, the low-pass cut-off for the gyroscope data was reduced to 0.3 Hz, but the equivalent cut-off frequency for the accelerometer data was maintained at 0.5 Hz. At the other end of the spectrum, we observed more low-frequency noise in the accelerometer channels compared to the gyroscope. The high-pass cut-off frequency for the accelerometer channels was therefore fixed at 0.12 Hz, and that for the gyroscope channels set to be 0.08 Hz. The filter responses are shown in Figure 4.12. The filtered signals were then downsampled to 4 Hz, to obtain the final respiratory signals from which RR is estimated.

²MATLAB function *filtfilt*

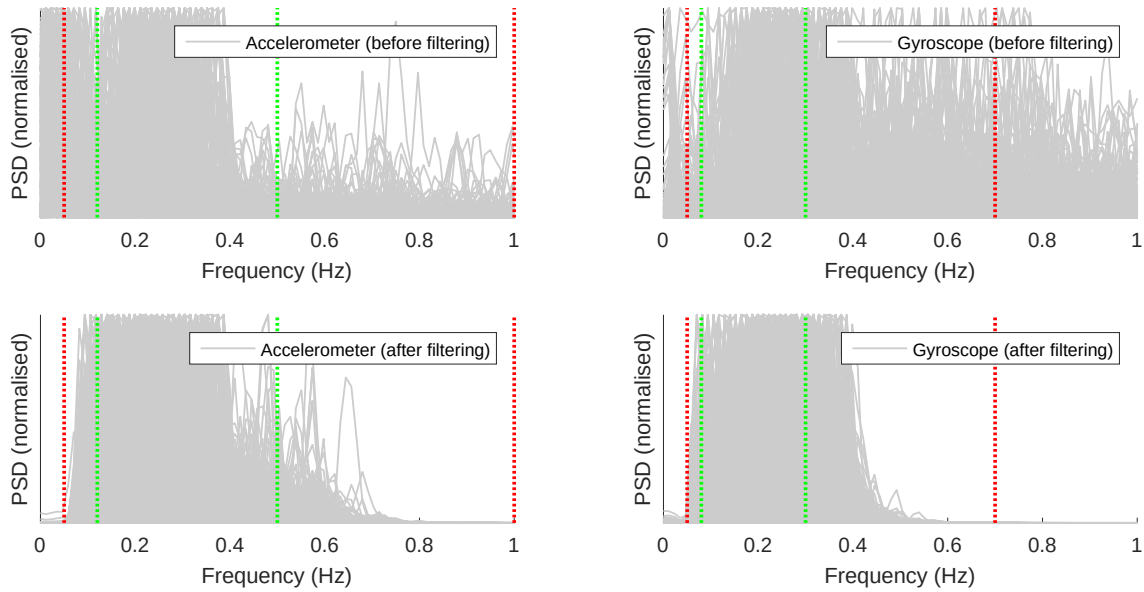


Figure 4.11: Superimposed normalised frequency spectra of a sample of the accelerometer (left column) and gyroscope (right column) recordings, before (top) and after (bottom) filtering. The filter limits are indicated by the vertical dashed lines (pass frequency in green; stop frequency in red).

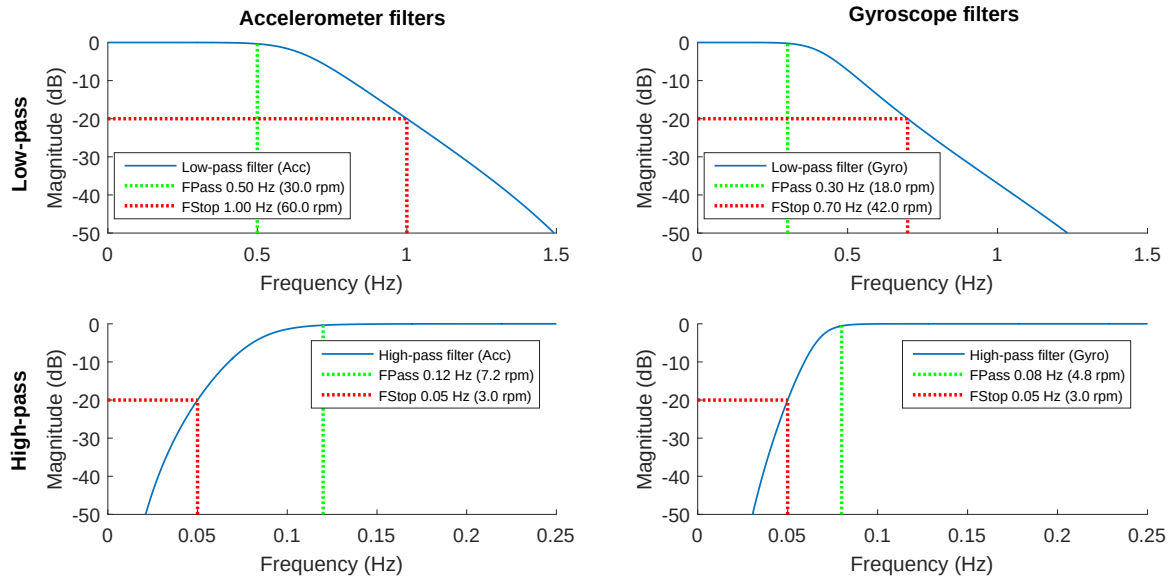


Figure 4.12: Filter responses for the low- (top) and high-pass (bottom) Butterworth IIR filters used to process the accelerometer (left column) and gyroscope (right column) signals. The attenuation in the stop-band is 20 dB in both the forward pass and the backward pass, giving an overall attenuation of 40 dB.

4.4.2 The PPG waveform

All peaks and troughs in the PPG waveform were identified using the peak detector described in Appendix B.1. The three respiratory signals, FM, BW and AM, were then extracted from the PPG (see Figure 4.3):

- **Frequency-modulation (FM):** The FM signal captures changes in HR frequency (HRV) by plotting the time between consecutive PPG peaks at the time of the first peak. FM therefore allows analysis of RR due to RSA, the HR variation during the respiratory cycle controlled by the parasympathetic nervous system (see Section 4.1.2).
- **Baseline wander (BW):** The troughs and corresponding time points were extracted for the BW signal. This captures the variation in perfusion baseline caused by chest movement during breathing.
- **Amplitude-modulation (AM):** The AM signal is the height of each PPG beat (peak - previous trough), plotted against the time point of the corresponding trough. This captures the changes in pulse strength caused by chest wall movement during breathing.

The signal quality index (SQI) tool described in Section 3.4 was applied to the respiratory window (60 seconds) of the PPG waveforms considered for respiratory analysis. Where the SQI tool highlighted poor-quality beats in the PPG waveform, the points derived from the corresponding peaks and troughs are removed from the FM, BW, and AM signals. New points were estimated using linear interpolation of the neighbouring points up to 5 seconds. If poor-quality sections were longer than 5 sections, the period from the start of the recording to the poor-quality section was removed (or from the poor-quality section to the end of the recording, whichever interval was shorter). If more than 30 seconds of a signal was excluded due to poor-quality, the entire recording was discarded with no attempt at RR estimation.

Poor-quality sections were treated in this way because the extraction of respiratory signals from the PPG relies on the position of peaks or troughs in the original PPG waveform. Any poor-quality beats where the peaks or troughs are not correctly positioned cause outliers in the derived respiration signals. Outliers then corrupt the frequency spectrum which make it more difficult to determine the correct RR. An example of this effect is shown in Figure 4.13. Outliers can also be introduced as a result of the detection of false peaks. Since the PPG SQI algorithm also begins with peak detection, false peaks would usually cause the corresponding beats to be marked as poor-quality.

As most frequency analysis techniques require a uniformly sampled signal, the respiratory signals were re-sampled to 4Hz using linear interpolation. Low frequencies below

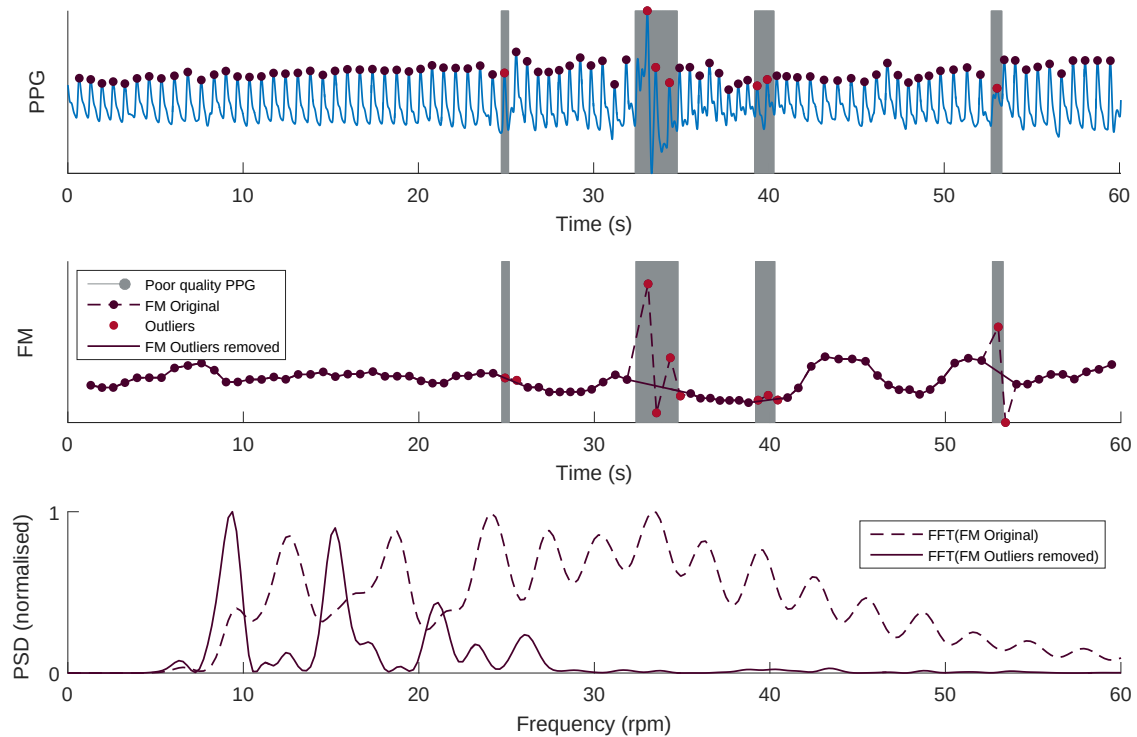


Figure 4.13: Points derived from poor-quality beats in the PPG waveform, are re-estimated by linear interpolation of the neighbouring points. **Top:** Original PPG waveform with poor quality sections highlighted (grey) and peaks identified from which the frequency modulated (FM) respiratory signal is derived. **Middle:** FM signal before (dashed line) and after (solid line) linear interpolation of poor quality sections. **Bottom:** Corresponding frequency spectra for the original FM signal (dashed line) and after exclusion of poor quality sections (solid line). High-frequency noise has been reduced after exclusion of poor-quality sections, and the mean peaks are now between 8 and 16 rpm which will change the respiratory rate estimate.

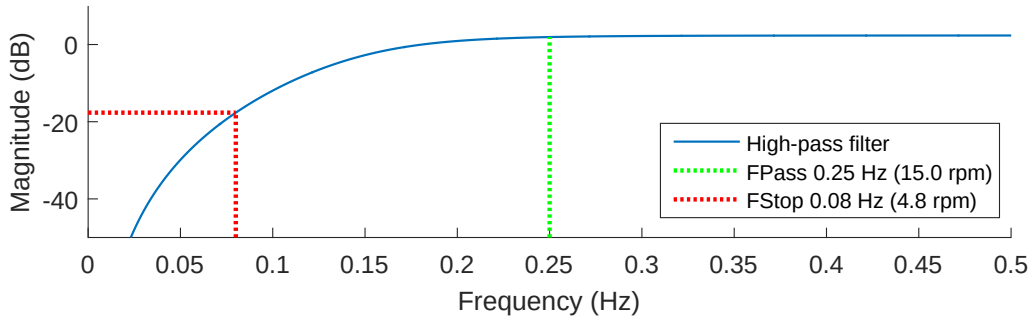


Figure 4.14: *Butterworth IIR high-pass filter used to attenuate low frequencies in the respiratory signals derived from the PPG waveform.*

the respiratory frequency (and therefore not of interest) were attenuated using a Butterworth IIR high pass filter with a cut-off frequency of 0.25 Hz (15.0 rpm) and 20 dB attenuation at 0.08 Hz (4.8 rpm) (Figure 4.14). Signals were filtered twice (forwards and backwards) so there was a zero-phase shift, and a total of 40 dB attenuation at 0.08 Hz.

The use of a high-pass filter was considered necessary for the PPG-derived respiratory waveforms because the low-frequency spectrum is thought to be altered by pre-processing steps embedded in commercial pulse oximeters [117]. The removal of low frequencies is also particularly important for the FM signal. HRV causes two major peaks in the frequency spectrum (see Section 4.1.2). The HF peak corresponds to the RR, and so the LF and VLF peaks need to be attenuated by the high-pass filter. However, when the RR was low (<10 rpm), the high-pass filter partially attenuated the RR frequency and so a different high-pass filter with a lower cut-off (0.15 Hz, 9 rpm) was used instead. The signals thought to contain a low RR were identified using the respiratory quality index described in the next section.

4.4.3 Respiratory Quality Index

We used a simple Respiratory Quality Index (RQI) using autocorrelation to obtain information about the quality of the respiratory signals derived from the PPG, accelerometer, and gyroscope waveforms.

Autocorrelation describes how strongly a signal is correlated with itself. The corre-

lation coefficient is computed as the sample correlation, r_k , between pairs of samples k units apart, and is normalised by the overall mean, $\bar{y}$, and overall sum of squares, c_0 , as described by:

$$r_k = \frac{c_k}{c_0}, \text{ where } c_k = \sum_{t=k+1}^n (y_t - \bar{y})(y_{t-k} - \bar{y}), \quad k = 1, 2, \dots \quad (4.2)$$

The sample correlation is a number between -1 and 1, where a 1 indicates perfect correlation, as would be expected for a sinusoid when the delay is equal to one period. We define $RQI(acf)$ as the highest correlation coefficient between 0.05 and 0.55 Hz (3 to 33 rpm) which corresponds to a time delay of between 1.82 and 20.0 seconds. For example, the correlation coefficients of a typical respiratory waveform is shown in Figure 4.15 where the $RQI(acf)$ is 0.66.

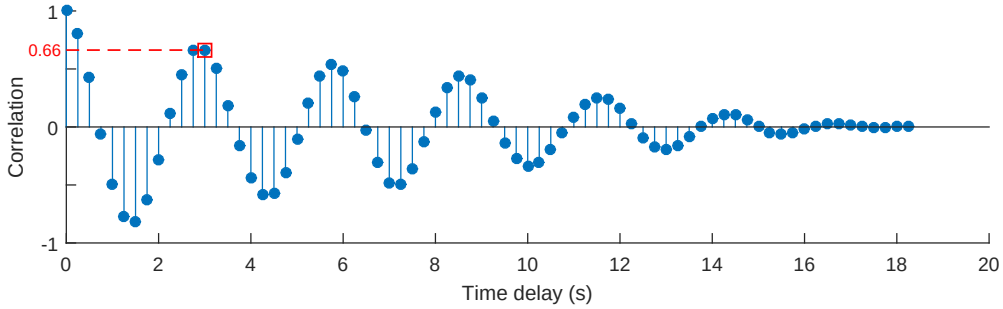


Figure 4.15: Illustration of calculation of $RQI(acf)$ to assess signal quality of respiratory signals derived from the PPG waveform and accelerometer.

$RQI(acf)$ for re-filtering PPG-derived respiratory signals

We hypothesised that when the RR is low, the current proposed filter (Figure 4.14) would attenuate some or all of the respiratory frequency from the PPG-derived signals (FM, BW, and AM). The resulting signal would not contain a sinusoidal respiratory signal, and therefore have a low $RQI(acf)$. A different high-pass filter would be required that did not attenuate the respiratory frequency.

If $RQI(acf)$ was less than 0.4 after the high-pass filtering of a PPG-derived signal, we re-filtered the signal using a high-pass filter with a transition band of 0.05 to 0.15

Hz. The filter response of the new filter is compared with the original filter in Figure 4.16. If the new filter improved the $RQI(acf)$ we kept the resulting signal and continued with RR estimation, otherwise the original filter was used. A flowchart of this process is illustrated in Figure 4.17, and an example of the difference this makes for signals with a low RR is in Figure 4.18.

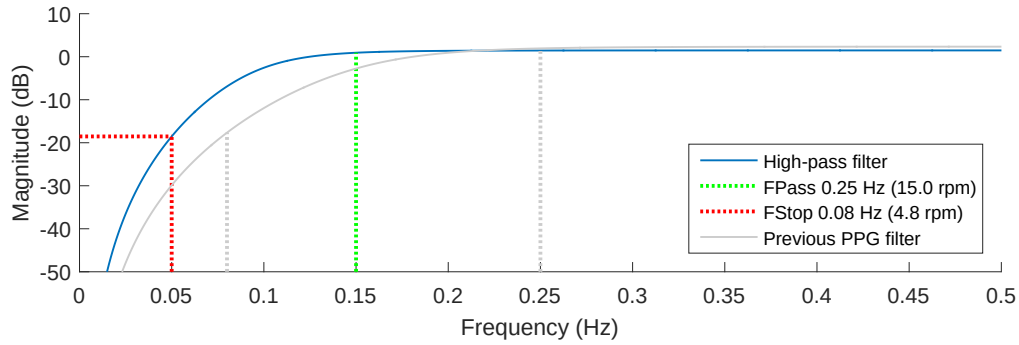


Figure 4.16: Modified Butterworth IIR high-pass filter for respiratory signals with poor signal quality. The original filter response is shown in grey.

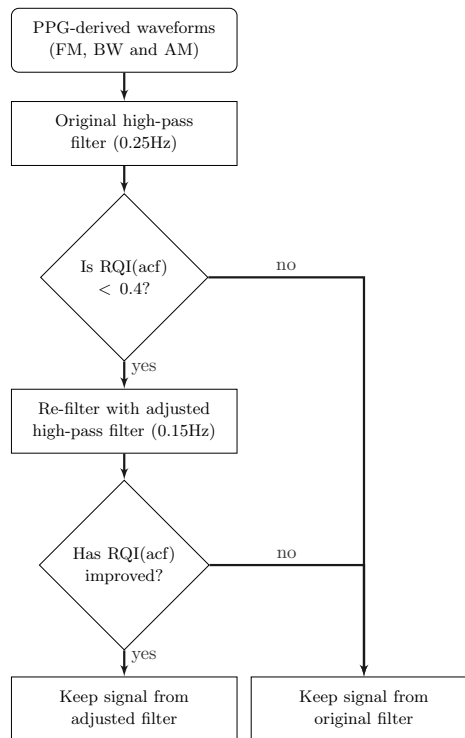


Figure 4.17: Flow chart of how respiratory quality index (RQI) is utilised to provide more appropriate low pass filtering to the PPG-derived respiratory waveforms.

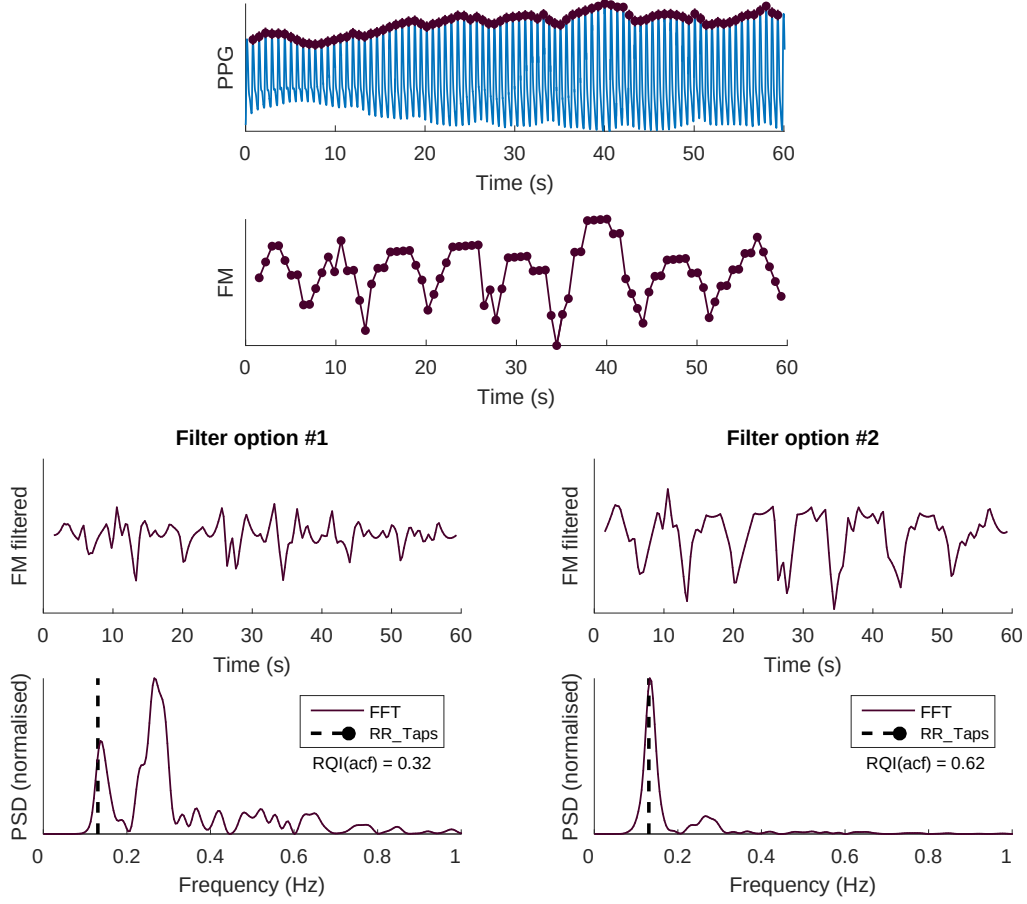


Figure 4.18: Re-filtering signals with a low $RQI(acf)$ can improve signal quality. In this example $RQI(acf)$ from the respiratory signal of the first filter is 0.32 compared to 0.62 from the second filter. Breaths are more visible in the second case. Note respiratory frequency (RR_{Taps}) is low (<10 rpm).

4.4.4 Principal component analysis

Principal component analysis (PCA) is a non-parametric tool for extracting relevant information from complex or confusing data sets [127]. The principal components are orthogonal axes orientated such that maximum variance is captured. PCA is often used for dimensionality-reduction.

Researchers in RR derivation have long recognised that some respiratory mechanisms are more dominant in some individuals than others, so deriving multiple respiratory waveforms can enhance respiratory accuracy [117]. This is particularly true for the PPG waveform where indirect effects of respiration are measured, which depend on age,

health, weight, and other factors. The smartphone axes are not necessarily aligned with physical movement during respiration. Smartphone orientation on the chest may also be different between individuals due to bust size or mechanisms used for breathing.

Previous research suggests that a dominant respiratory axis can be extracted from the orthogonal accelerometer and gyroscope axes using PCA [103]. Jin *et al.* found that the respiratory signal yielded from the dominant principal component led to more accurate RR estimation compared to choosing a the x , y , or z axis. We therefore extracted the principal component of the filtered accelerometer and gyroscope channels (referred to as A_p and G_p respectively), by taking the eigenvector (principal component) with the largest associated eigenvalue. Figure 4.19 illustrates the positions of the principal components in $x-y-z$ space for an example accelerometer recording. RR was estimated from A_p and G_p in addition to the six individual accelerometer and gyroscope channels.

We also applied PCA to the three respiratory signals derived from the PPG waveform (FM, BW and AM), since they all represent the same respiratory signal. We refer to the principal component of the PPG-derived signals as PC.

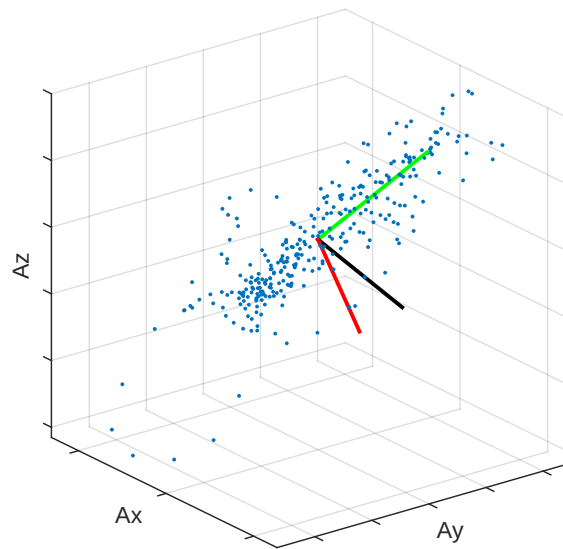


Figure 4.19: The principal components (black, red, green) of the three accelerometer axes. Here, the green principal component has the largest eigenvalue (i.e. it explains most of the variance in the data), so data is projected along this “respiratory” axis to obtain A_p .

4.4.5 Respiratory rate calculation

We estimated RR using three techniques: the Fourier transform (frequency domain), an Auto-Regressive (AR) model (frequency domain), and peak detection (time domain). Furthermore, we obtained three RR estimates from the AR model, utilising different properties. These five algorithms will be referred to as the “respiratory methods”.

For frequency based techniques, we defined a respiratory range, *RR range*, from 0.05 to 0.55 Hz (3 to 33 rpm) within which we expected the respiratory frequency to be located. This includes the normal adult breathing range of 12 to 20 rpm [88, 44] and the top limit exceeds the maximum expected RR of 24 rpm in the reference data (see Figure 4.9).

Fourier transform:

The Fourier transform decomposes a time-series signal into a sum of sinusoids at varying frequencies. We use the Fast Fourier Transform (FFT) with a Hamming window to obtain the frequency spectrum. RR is estimated as the frequency with the largest magnitude within *RR range*. Estimates from the Fourier transform will be referred to as FFT:max.

AR model:

The AR model is a set of auto-correlation functions that predicts each value in a series as a linearly-weighted sum of the preceding M values of the same series plus an error term, $e[n]$, described by:

$$x[n] = \sum_{k=1}^M a_k x[n-k] + e[n] \quad (4.3)$$

M is the model order, which is usually much smaller than the length of the sequence. The AR model can be formulated as a system of linear equations, where the optimal weights are those that minimise the error. We use Burg’s method to solve the AR equations [128]. The AR model can also be written in the form of a filter and therefore

used for frequency analysis. The transfer function is:

$$H(z) = \frac{1}{\sum_{k=1}^p a_k z^{-k}} = \frac{z^M}{(z - z_1)(z - z_2) \dots (z - z_M)} \quad (4.4)$$

The denominator can be factorised into M terms, each of which defines a root z_k corresponding to a pole of $H(z)$. Since $H(z)$ has no finite zeros, the AR model is an all-pole model. The poles occur in complex-conjugate pairs, and define spectral peaks in the power spectrum of the signal. Large-magnitude poles correspond to large-magnitude spectral peaks. The phase angle θ of a pole, gives us the frequency f of the corresponding spectral peak, described by:

$$\theta = 2\pi f \Delta t \quad (4.5)$$

AR modelling is commonly used to extract information in biological systems [129]. We used the AR model in three ways to estimate RR from the respiratory waveforms. We initially set the model order to $M = 5$, and then determined the best model order using autocorrelation. The waveforms were first downsampled to 2 Hz to improve frequency resolution of the pole-zero plot.

1. AR spectrum (AR:spec): The frequency spectrum from an AR model order 5 is used to estimate the RR. We selected the frequency with the largest magnitude power density within the *RR range*.
2. AR poles (AR:poles): RR is determined directly from the poles of an AR model order 5. We follow the method suggested by Fleming *et al.* [107]. Candidate poles must have an amplitude which is at least 95% of the largest pole magnitude. First, a narrow RR range from 0.05 to 0.43 Hz (3 to 26 rpm) is searched for a candidate pole. If no pole is identified, the frequency range of interest is extended to RR range. If still no pole is selected, RR is not estimated. If there are multiple candidate poles at any stage, the pole with the smallest angle (corresponding to

the lowest frequency) is selected. Harmonic poles with a larger magnitude than the true RR pole sometimes occur, although this is more important for higher model orders.

3. Best AR model (BestAR:spec): The most appropriate AR model order for the respiratory waveform is selected by cross correlation [130]. AR models with orders between 5 and 11 are considered, from which the frequencies at all poles and frequency spectrum peaks are noted as possible RRs. A pseudo-sine wave signal is generated at the possible RR frequencies and cross correlated with the original respiratory waveform. The frequency for which the cross correlation coefficient has the largest magnitude is chosen as the RR. Thus this method is a combination of AR poles and AR spectrum peaks, for all model orders.

Peak detection:

Peaks are identified in the time domain of the respiratory waveforms using the peak detector described in Appendix B.1. The RR is calculated as the inverse of the mean peak-peak intervals. Outliers are not rejected. Estimates using peak detection will be referred to as PK:mean.

4.4.6 Evaluation of methods

Respiratory rate estimates from each waveform and algorithm were compared to the reference RR (derived from tablet taps as described in Section 4.3.1). Algorithm performance was assessed by calculating the Absolute Error (AE) in respirations per minute for each recording, defined by:

$$AE_i = |RR_i - RR_{iTaps}| \quad (4.6)$$

where RR_i is the estimated respiratory rate and RR_{iTaps} is the reference respiratory rate for recording i . The Mean Absolute Error (MAE) across all n recordings was calculated

as:

$$MAE = \frac{1}{n} \sum_{i=1}^n AE_i \quad (4.7)$$

MAE describes the average error and is a common metric used to evaluate algorithm performance. Multiple recordings from the same subject were considered to be independent.

The grouped Kruskal-Wallis test, a non-parametric version of the classical one-way analysis of variance (ANOVA) that uses the chi-square statistic (χ^2), was used to determine any significant difference in the AE distributions between methods used to derive RR. Posthoc pair-wise Mann-Whitney U test were performed for further investigation. The Bonferroni corrected p -values are reported [131].

The best respiratory method per waveform was selected based on the lowest median AE. The grouped Kruskal-Wallis test and pair-wise Mann-Whitney U tests were again used to compare the AE distributions of each waveform.

The effect of waveform quality on RR estimation was investigated by removing waveforms with a low RQI value from the analysis. Waveforms with an $RQI(acf) < 0.4$ were removed, and the AEs were recalculated using only the high quality waveforms. Statistical analysis was repeated on the new AE distributions. The correlation of RQI with respiratory rate and heart rate was assessed by linear regression.

RR estimates from different devices were also compared. A representative waveform from each device was selected for comparison (e.g. PC from the pulse oximeter, and G_p from the smartphone). The absolute difference between RR estimates from the two waveforms was calculated by respiratory method; first for all recordings, and then for high quality recordings only. The absolute difference distributions were compared using the statistical tests described above.

4.4.7 Summary

The methodology for RR estimation from PPG and accelerometer/gyroscope signals is summarised in the flow diagram in Figure 4.20. PPG signals are first checked using the SQI tool described in Section 3.4 and discarded if less than 30 seconds of good-quality signal remains. Three respiratory signals (FM, BW, and AM) are derived from the PPG waveform, capturing neurological and mechanical effects of breathing. Erroneous beat-to-beat intervals from poor-quality sections are removed and linear interpolation used to replace them. All signals are high-pass filtered to eliminate noise, and then combined using PCA to obtain a fourth respiratory waveform, PC. The smartphone accelerometer recordings contain three accelerometer axes (A_x , A_y and A_z) and three gyroscope channels (G_x , G_y and G_z). These are low- and high-pass filtered, and then combined using PCA to obtain A_p and G_p .

Twelve respiratory waveforms are therefore considered for RR estimation. Signal quality is assessed using autocorrelation, RQI(acf). PPG-derived waveforms with a low RQI ($\text{RQI}(\text{acf}) < 0.4$) are re-filtered with a different high-pass filter that has a lower cut-off frequency.

RR is estimated from the respiratory waveforms using five methods: (1) FFT frequency spectrum; (2) AR model order 5 frequency spectrum; (3) poles from AR model order 5; (4) the frequency spectrum and poles from AR model orders 5 to 11; and (5) peak detection in the time domain. For frequency based methods (Fourier and AR models), the frequency range of interest, *RR range*, is defined as 0.05 to 0.55 Hz (3 to 33 rpm). Respiratory estimates were evaluated by comparing the AE distributions using non-parametric statistical tests.

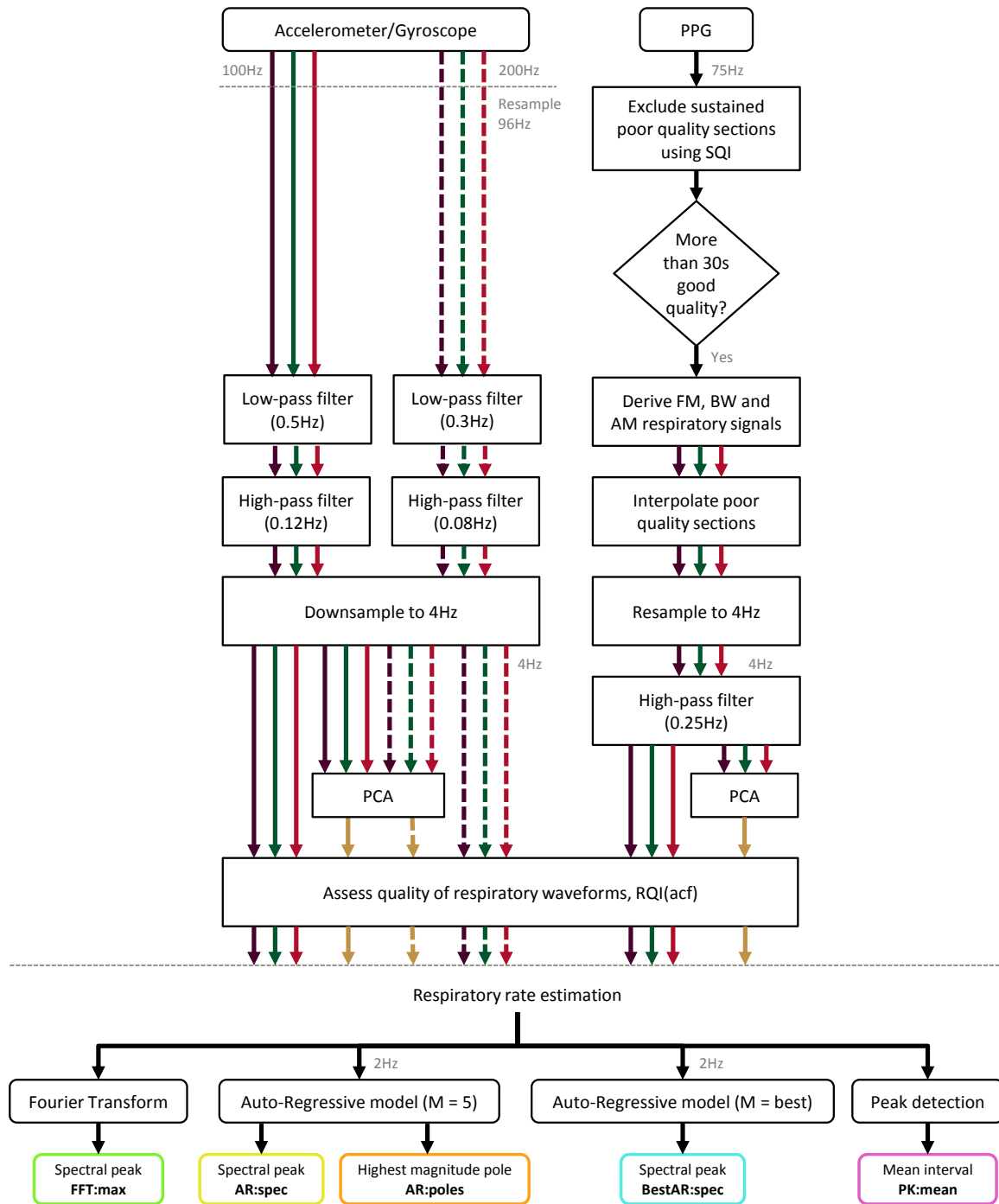


Figure 4.20: Summary of respiratory rate estimation from PPG and accelerometer/gyroscope waveforms.

4.5 Results

697 pairs of PPG and accelerometer/gyroscope recordings, from 152 women, were considered for RR analysis. 73 recordings were excluded because the RR calculated from tablet taps (RR_{Taps}) was not consistent to within 2 rpm of the midwife count and thus the true RR could not be determined. A further 57 sets of recordings were eliminated due to poor-quality of the PPG waveform. Respiratory estimates were therefore calculated for 567/697 (81.4%) of available recordings.

Figure 4.21 shows the Absolute Error (AE) distributions of each method for the twelve respiratory waveforms. The median, 25th centile, and 75th centile, are shown in Table 4.1.

The grouped Kruskal-Wallis test showed a significant difference between the methods ($\chi^2 = 122.3, p < 0.001$). The post-hoc multicomparison tests according to the Bonferroni method revealed that the performance of FFT:max was significantly better than AR:spec ($p < 0.001$), AR:poles ($p < 0.001$), and PK:mean ($p = 0.012$). No statistical significant difference was observed between FFT:max and BestAR:spec ($p = 1.000$). BestAR:spec also had a significantly lower median AE than AR:spec ($p < 0.001$), AR:poles ($p < 0.001$), and PK:mean ($p = 0.001$). However, although the median AE of both FFT:max and BestAR:spec were low, the MAE was observed to be relatively large (often greater than the 75th centile) which indicates a highly skewed distribution. We hypothesise that many outliers in these distributions will be due to a poor-quality respiratory signal, and that the MAE reduces by eliminating waveforms with a poor quality index. The relationship between waveform quality and error will be examined in Section 4.5.1.

The best method for each waveform, chosen by the lowest median AE, is highlighted in blue in Table 4.1. For the majority of waveforms (8/12), BestAR:spec was selected as the best method. PK:mean was selected for 3/12 waveforms, and FFT:max selected for 1/12 waveforms. The AE distributions for the best method per waveform are shown in

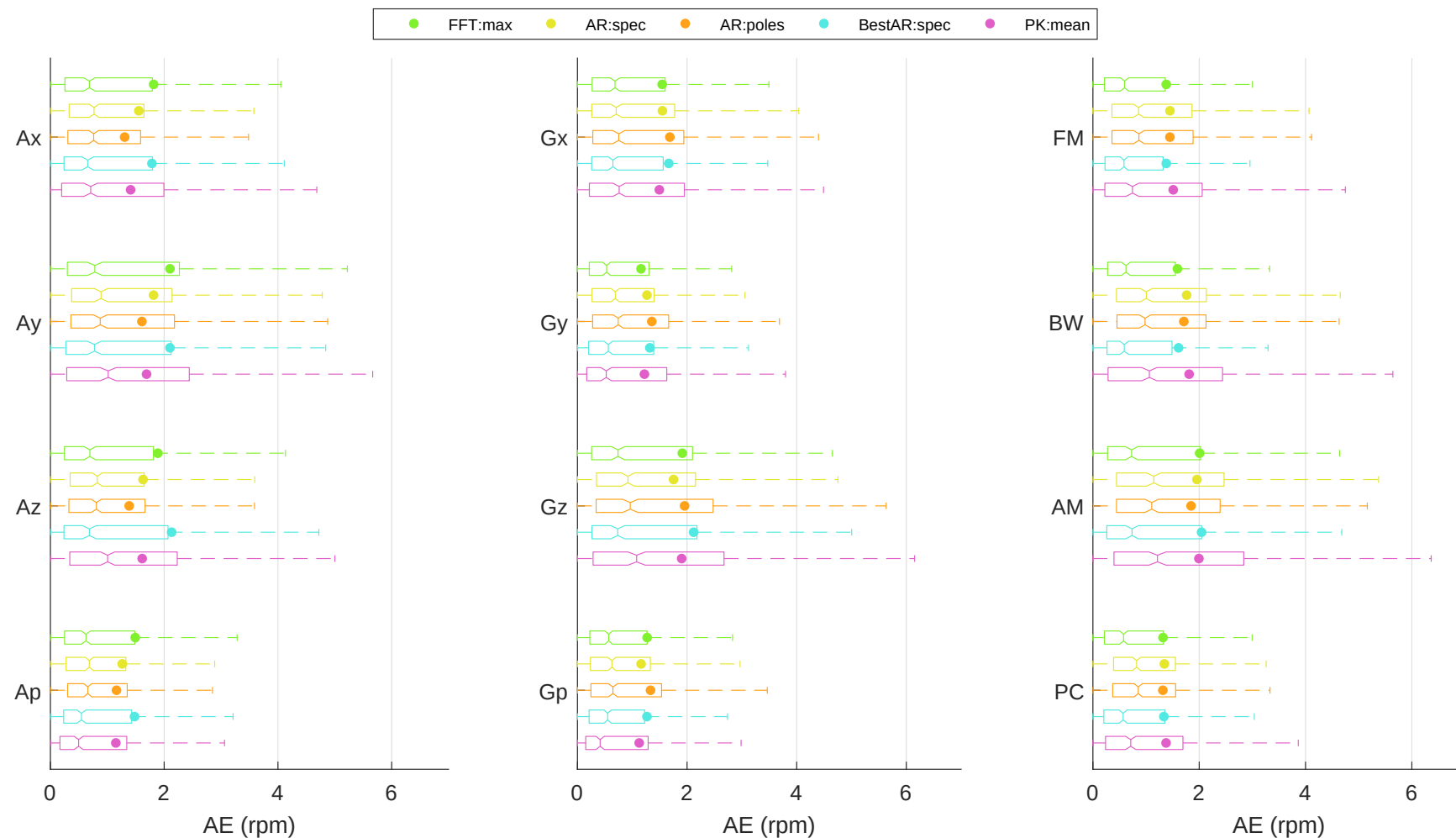


Figure 4.21: Comparison of the Absolute Error (AE) for all respiratory methods (colours) and respiratory waveforms (y-axis). Error is defined as the reference respiratory rate (from tablet taps) subtracted from the estimated respiratory rate. The boxplot shows distributions of AEs, with lower quartile, median and upper quartile values displayed as left, middle and right vertical lines of the boxes. Whiskers are used to represent the most extreme values within 1.5 times the interquartile range from the central box. Circles represent the mean values.

	FFT:max	AR:spec	AR:poles	BestAR:spec	PK:mean	<i>p-value</i>
A_x	0.68 (0.25, 1.79)	0.77 (0.33, 1.64)	0.76 (0.31, 1.59)	0.66 (0.24, 1.80)	0.71 (0.20, 2.00)	<i>0.003</i>
A_y	0.78 (0.30, 2.27)	0.89 (0.37, 2.14)	0.88 (0.36, 2.18)	0.78 (0.27, 2.12)	1.02 (0.29, 2.44)	<i>0.010</i>
A_z	0.69 (0.25, 1.81)	0.82 (0.35, 1.65)	0.81 (0.32, 1.66)	0.69 (0.24, 2.07)	1.01 (0.34, 2.23)	<i>0.011</i>
A_p	0.63 (0.25, 1.48)	0.68 (0.28, 1.33)	0.66 (0.30, 1.35)	0.55 (0.23, 1.43)	0.50 (0.17, 1.34)	<i>0.023</i>
<i>p-value</i>	<i><0.001</i>	<i><0.001</i>	<i><0.001</i>	<i><0.001</i>	<i><0.001</i>	
G_x	0.69 (0.27, 1.60)	0.71 (0.26, 1.77)	0.75 (0.28, 1.94)	0.65 (0.26, 1.56)	0.76 (0.22, 1.95)	<i>0.013</i>
G_y	0.53 (0.22, 1.31)	0.69 (0.27, 1.40)	0.74 (0.28, 1.66)	0.56 (0.20, 1.39)	0.53 (0.17, 1.63)	<i>0.004</i>
G_z	0.74 (0.26, 2.10)	0.92 (0.35, 2.15)	0.96 (0.34, 2.48)	0.74 (0.26, 2.18)	1.08 (0.29, 2.67)	<i><0.001</i>
G_p	0.57 (0.23, 1.27)	0.63 (0.24, 1.33)	0.65 (0.25, 1.54)	0.55 (0.21, 1.23)	0.41 (0.15, 1.29)	<i>0.193</i>
<i>p-value</i>	<i>0.001</i>	<i><0.001</i>	<i><0.001</i>	<i>0.004</i>	<i><0.001</i>	
FM	0.60 (0.22, 1.36)	0.86 (0.36, 1.86)	0.87 (0.36, 1.88)	0.58 (0.23, 1.32)	0.74 (0.23, 2.06)	<i>0.026</i>
BW	0.62 (0.28, 1.55)	1.00 (0.44, 2.13)	0.98 (0.45, 2.13)	0.59 (0.27, 1.49)	1.06 (0.29, 2.44)	<i>0.002</i>
AM	0.73 (0.28, 2.03)	1.15 (0.44, 2.47)	1.11 (0.44, 2.39)	0.74 (0.26, 2.05)	1.22 (0.40, 2.84)	<i><0.001</i>
PC	0.58 (0.22, 1.33)	0.82 (0.39, 1.55)	0.85 (0.37, 1.56)	0.57 (0.21, 1.35)	0.71 (0.24, 1.69)	<i><0.001</i>
<i>p-value</i>	<i>0.020</i>	<i>0.001</i>	<i><0.001</i>	<i>0.004</i>	<i><0.001</i>	

Table 4.1: Median absolute error (25th centile, 75th centile). The *p*-values reported are the results of grouped Kruskal-Wallis statistical tests for difference in the median of the AE distribution across methods (rows) and groups of waveforms (columns). The values in blue indicate the best method chosen per waveform, based on the lowest median AE value.

Figure 4.22. The grouped Kruskal-Wallis test showed a significant difference between the median AEs of the best method per waveform ($\chi^2 = 91.1$, $p < 0.001$). Posthoc pairwise tests revealed that the performance of the A_p waveform was significantly different from the A_x ($p = 0.011$), A_y ($p < 0.001$), and A_z ($p = 0.003$) waveforms. Similarly, median AE from the G_p waveform was significantly lower than median AE from G_x ($p = 0.001$) and G_z ($p < 0.001$), although not significantly different from G_y ($p = 1.000$). No significant difference was observed between the AE distributions of A_p , G_y , and G_p . For PPG-derived waveforms, the only significant difference in distributions was between PC and AM ($p = 0.028$), where PC had a lower median than AM.

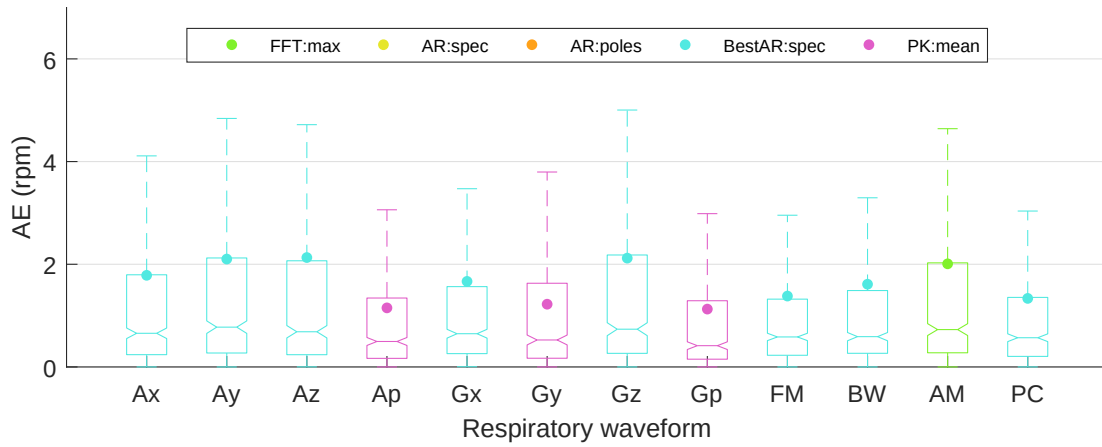


Figure 4.22: Comparison of the Absolute Error (AE) for the best respiratory method per respiratory waveform. The AE distributions for all methods are shown in Figure 4.21. The boxplot shows distributions of AEs, with lower quartile, median and upper quartile values displayed as left, middle and right vertical lines of the boxes. Whiskers are used to represent the most extreme values within 1.5 times the interquartile range from the central box. Circles represent the mean values.

The largest median AE from the smartphone-derived waveforms was 0.78 rpm from A_y , followed by 0.74 rpm for the G_z waveform. As previously discussed, the smartphone axes are not necessarily aligned with physical movement during respiration and the smartphone orientation on the chest may be different between individuals due to bust size or mechanisms used for breathing. It is therefore intuitive that some waveforms will contain more respiratory information than others. The A_y and G_z signals are likely to contain less respiratory information because of the orientation of the smartphone on the

women's chest, contributing to the poor quality of the waveform³. Figure 4.23 shows the three gyroscope axes (x , y , and z) of an example recording, where the respiratory signal is visible in the G_x and G_y waveforms, but not in G_z .

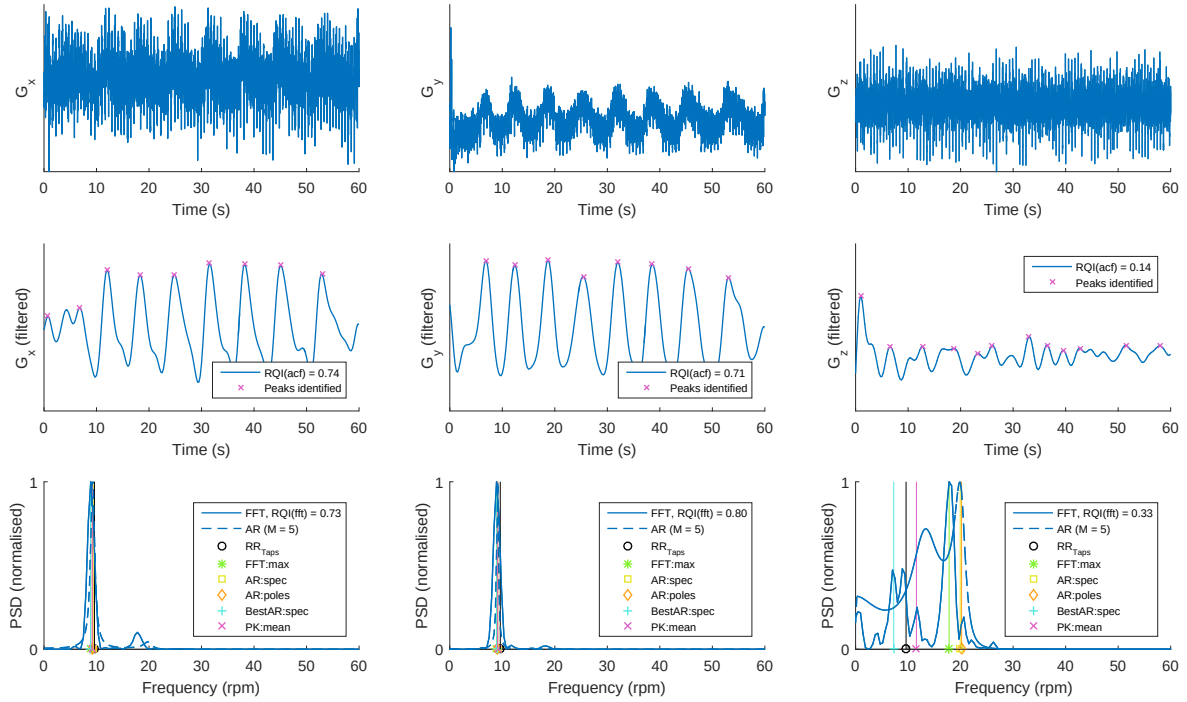


Figure 4.23: Example of the difference in respiratory information between the three gyroscope channels. A periodic respiratory signal is clearly visible from G_x (left column) and G_y (middle column) and thus the respiratory methods converge to a common respiratory estimate. However, G_z (right column) contains more noise and the respiratory signal is less prominent. The frequency spectrum for G_z shows no consensus on the true respiratory rate.

The largest median AE from the PPG-derived respiratory waveforms was 0.73 rpm from AM, calculated from the amplitude-modulation of the PPG signal. This is likely to be caused by attenuation of the amplitude modulation by the filters used in commercial pulse oximeters, such as the Nonin device used in the 4P study.

The AE distributions examined so far have described the variance in RR estimates, an indication of method *precision*. An ideal method will have both a high precision and high *accuracy*. The mean (95%CI) bias for the best method for each waveform

³The orientation of the smartphone on the woman's chest is shown in Figure 3.4 in Section 3.2.1. The x - and y -axes are in the plane of the phone, and the z -axis is perpendicular, coming out from the chest.

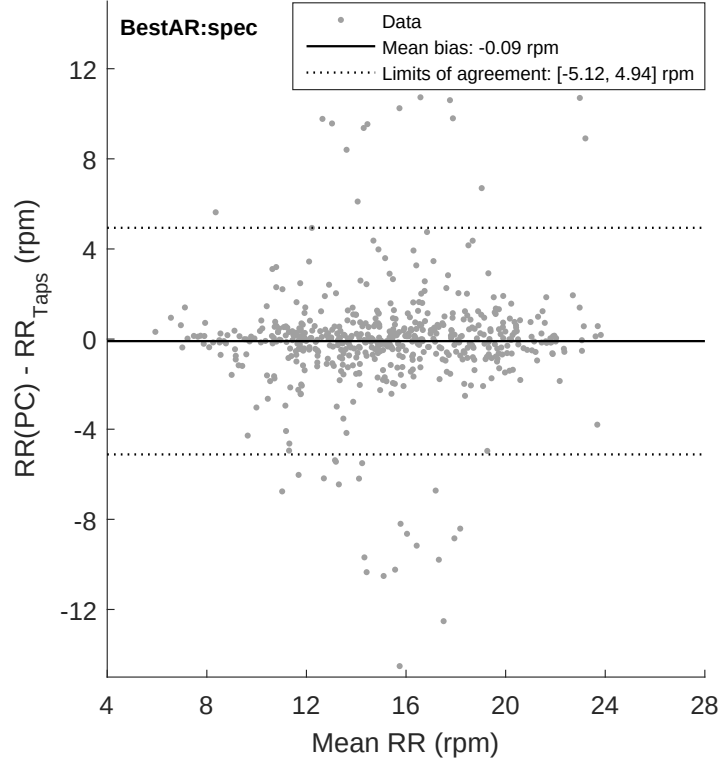


Figure 4.24: Bland-Altman plot comparing the reference respiratory rate, RR_{Taps} , with respiratory estimates from the PC waveform using the BestAR:spec method.

ranged from -0.96 (-1.12, -0.81) rpm to 0.48 (0.23, 0.73) rpm. The mean bias was not significantly different from zero for 5/12 waveforms. Most distributions had a negative mean bias (11/12), which indicates that on average the algorithms underestimate RR. The relationship of error with RR can be examined using a Bland-Altman plot where the mean RR estimate is plotted against the difference between the two estimates [132]. Figure 4.24 shows an example Bland-Altman plot for the PC waveform.

4.5.1 Respiratory quality

The autocorrelation quality index, $RQI(acf)$, described in Section 4.4.3, can be used to assess the quality of each respiratory waveform. Figure 4.25 shows the spread of $RQI(acf)$ for each waveform, where 1 is a perfect sinusoid and 0 indicates poor quality. The highest $RQI(acf)$ recorded was 0.92 for a G_y waveform, and the lowest value was 0.02 for a A_z waveform.

The $RQI(acf)$ values of the principal component waveforms are, generally, higher than the $RQI(acf)$ values for the waveforms from which they were derived. Pair-wise Mann-Whitney U tests showed that the median $RQI(acf)$ of the A_p waveforms was significantly higher than that of A_x ($p = 0.015$), A_y ($p < 0.001$), and A_z ($p < 0.001$). Similarly, the median $RQI(acf)$ of the G_p waveforms was significantly higher than that of G_x ($p < 0.001$), G_y ($p = 0.031$), and G_z ($p < 0.001$), and the PC median $RQI(acf)$ was significantly higher than that of FM ($p = 0.005$), BW ($p < 0.001$), and AM ($p < 0.001$).

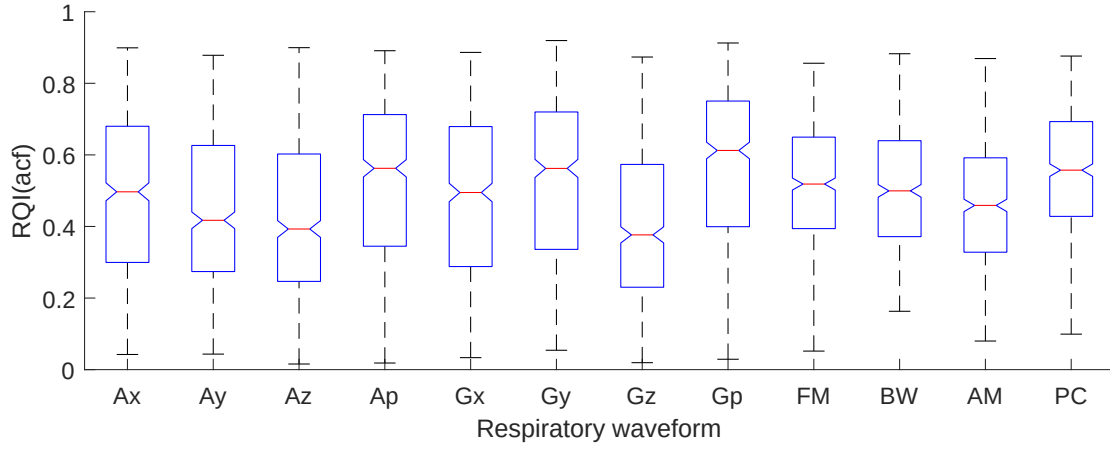


Figure 4.25: Distribution of respiratory quality index, $RQI(acf)$, for the twelve respiratory signals. The box plots indicate the median and interquartile range.

We previously hypothesised that respiratory estimates from poor-quality signals (indicated by a low $RQI(acf)$) would be more likely to have a high error. Consider the respiratory estimates from PC in Figure 4.26, where the colour indicates the value of $RQI(acf)$. Most estimates far from the reference RR are coloured red, which indicates that the respiratory waveform is of poor-quality. When these poor-quality waveforms are excluded, the MAE decreases, as shown in Figure 4.27. When all recordings with $RQI(acf) < 0.4$ ($n = 126$, 22.2%) are excluded, the MAE for all methods decreases to below 1 rpm.

Note also that the MAE of FFT:max and BestAR:spec methods decreases faster than the MAE for the other methods until convergence occurs when only high-quality record-

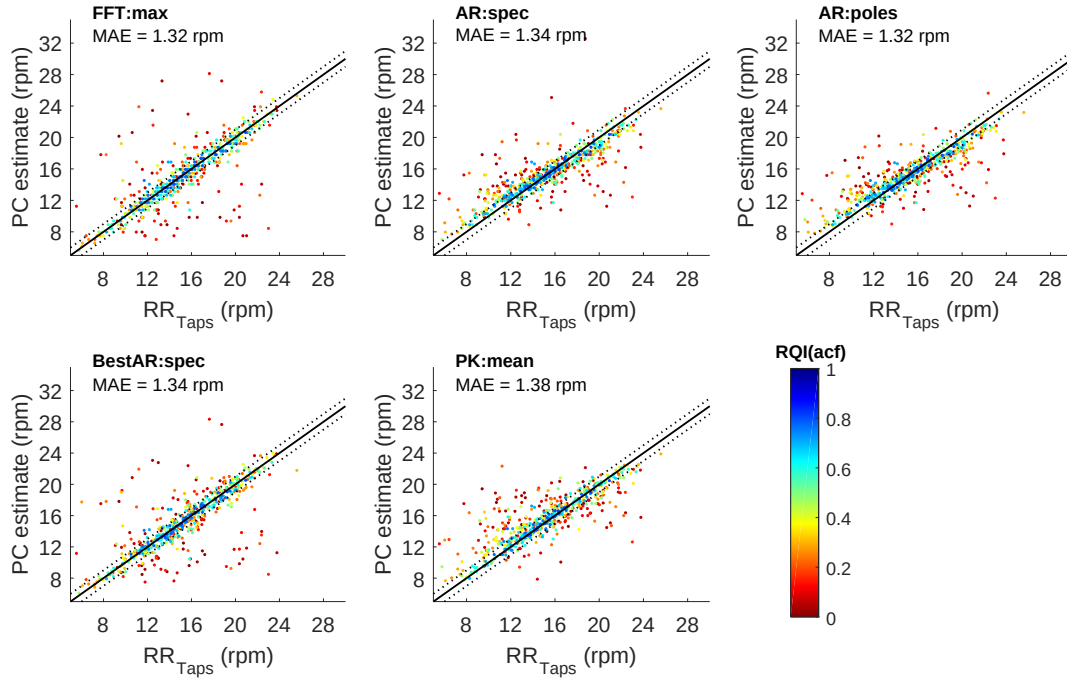


Figure 4.26: Respiratory rate estimates from the PC waveform (principal component of the three PPG-derived waveforms: FM, BW, and AM), against the reference respiratory rate, RR_{Taps} . The colour indicates the quality of the PC waveform, $RQI(acf)$, where red indicates poor-quality waveforms and blue indicates good-quality waveforms.

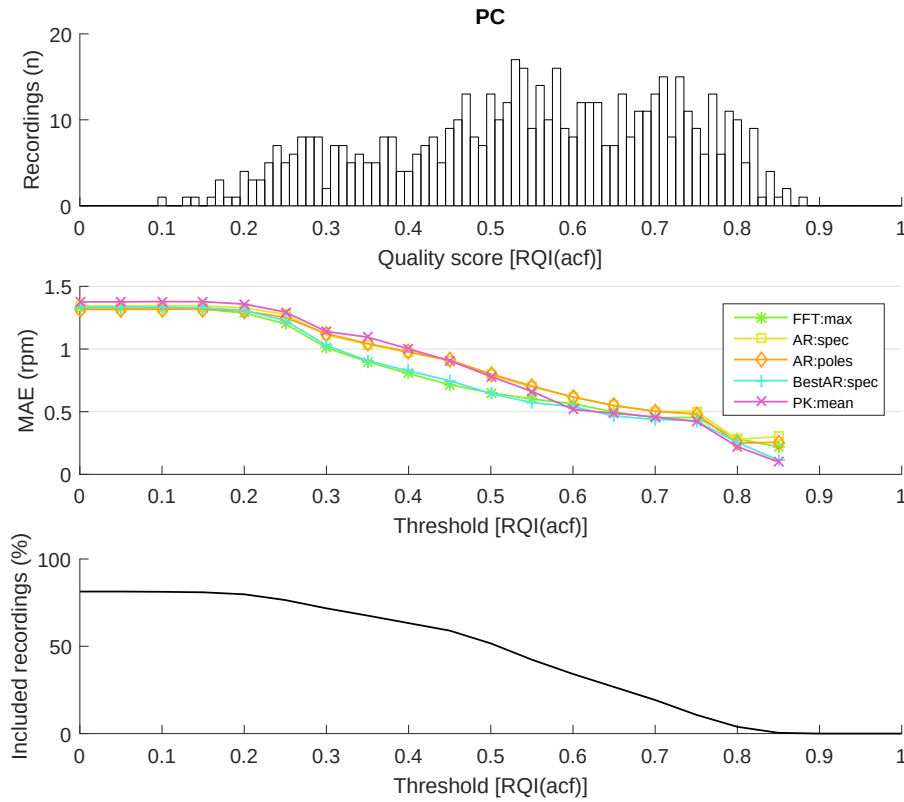


Figure 4.27: As low-quality recordings are excluded, MAE decreases.

ings remain. The AE distributions for these methods are highly skewed (as previously shown in Figure 4.21) due to outliers. The rapid decrease in MAE when poor-quality waveforms are excluded suggests that outliers in the AE distribution are associated with poor-quality waveforms. It therefore follows that FFT:max and BestAR:spec are more robust than other methods at identifying the true RR for *medium-quality* recordings.

The robustness of BestAR:spec in comparison to the other two AR models, AR:spec and AR:poles, for medium-quality recordings can be explained by the choice of model order. The former method chose the model order based on the sinusoidal nature of the resulting waveform, whereas the latter two methods had a fixed model order ($M = 5$). The frequency spectra of medium-quality waveforms were often noisy, so the flexibility of model order often allowed BestAR:spec to more appropriately describe the frequency spectrum and therefore select the correct respiratory frequency. Consider the example in Figure 4.28, in which the PC respiratory waveform has an RQI(acf) of 0.51. There are two peaks in the frequency-spectrum, at 11 rpm and 21 rpm. AR model order 5 has limited poles available, so fits to a single peak at 13.6 rpm. In contrast, BestAR:spec selects model order 10, and correctly identifies the two spectral peaks (the fundamental and second harmonic). This pattern is similar for other medium-quality recordings, particularly when the true RR is low. High-order AR models can lead to “pole splitting”, whereby the dominant spectral peak is split between more than one pole. Hence lower order models are preferred, and are suitable for high-quality recordings. The recordings where BestAR:spec selected a model order of 5 (the lowest option), were of significantly higher quality than the recordings when higher model orders were selected ($p < 0.001$).

We set the RQI(acf) threshold at 0.4. Waveforms with an RQI(acf) above the threshold were considered to be of good-quality, and those below or equal to the threshold were classified to be poor-quality waveforms. Figure 4.29 shows the updated boxplots of the AE distributions from respiratory rate estimates derived only from good-quality wave-

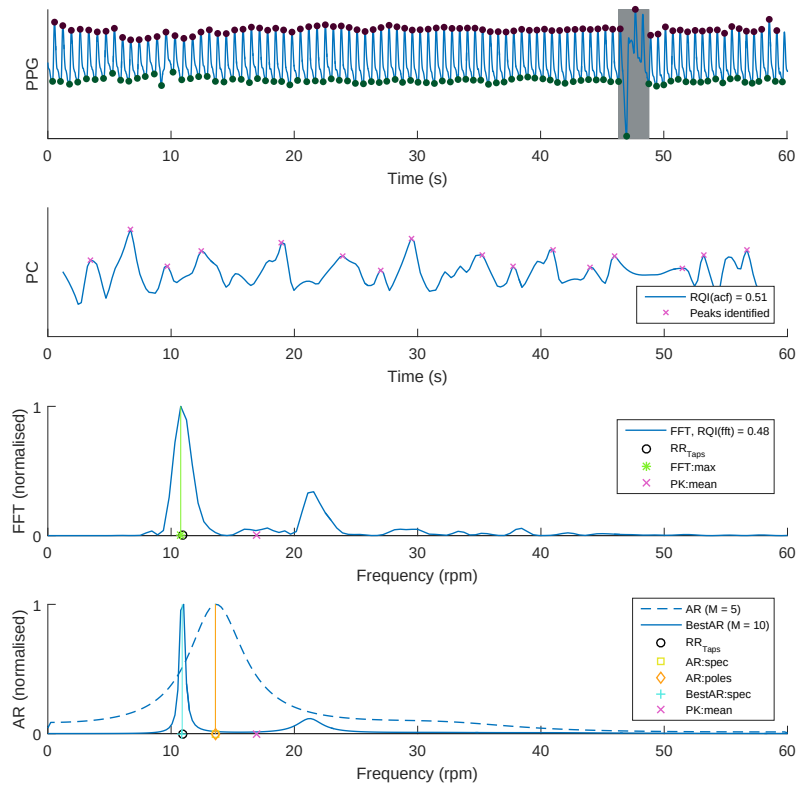


Figure 4.28: Example of a medium-quality ($RQI(acf) = 0.51$) PC respiratory signal (derived from the PPG) where only FFT:max and BestAR:spec identify the true respiratory rate.

forms.

The Kruskal-Wallis test again showed a significant difference between distributions from each method ($\chi^2 = 202.4$, $p < 0.001$). The time-domain peak-detection algorithm, PK:mean, performed particularly well on good quality signals and had the lowest median AE for 7/8 of the accelerometer and gyroscope waveforms. Pairwise tests across accelerometer and gyroscope distributions only (PPG-derived waveforms excluded) showed that the PK:mean distribution was significantly different from that for every other method ($p < 0.001$). BestAR:spec gave the lowest median AE for the PPG-derived waveforms.

The results of $RQI(acf)$ regressed on RR and HR showed that generally the quality of the gyroscope waveforms was positively correlated with the RR value, and that the quality of the PPG-derived waveforms was negatively correlated with the HR value for the recording. For every 10 rpm, $RQI(acf)$ increased a mean (95% CI) of 0.10 (0.05,

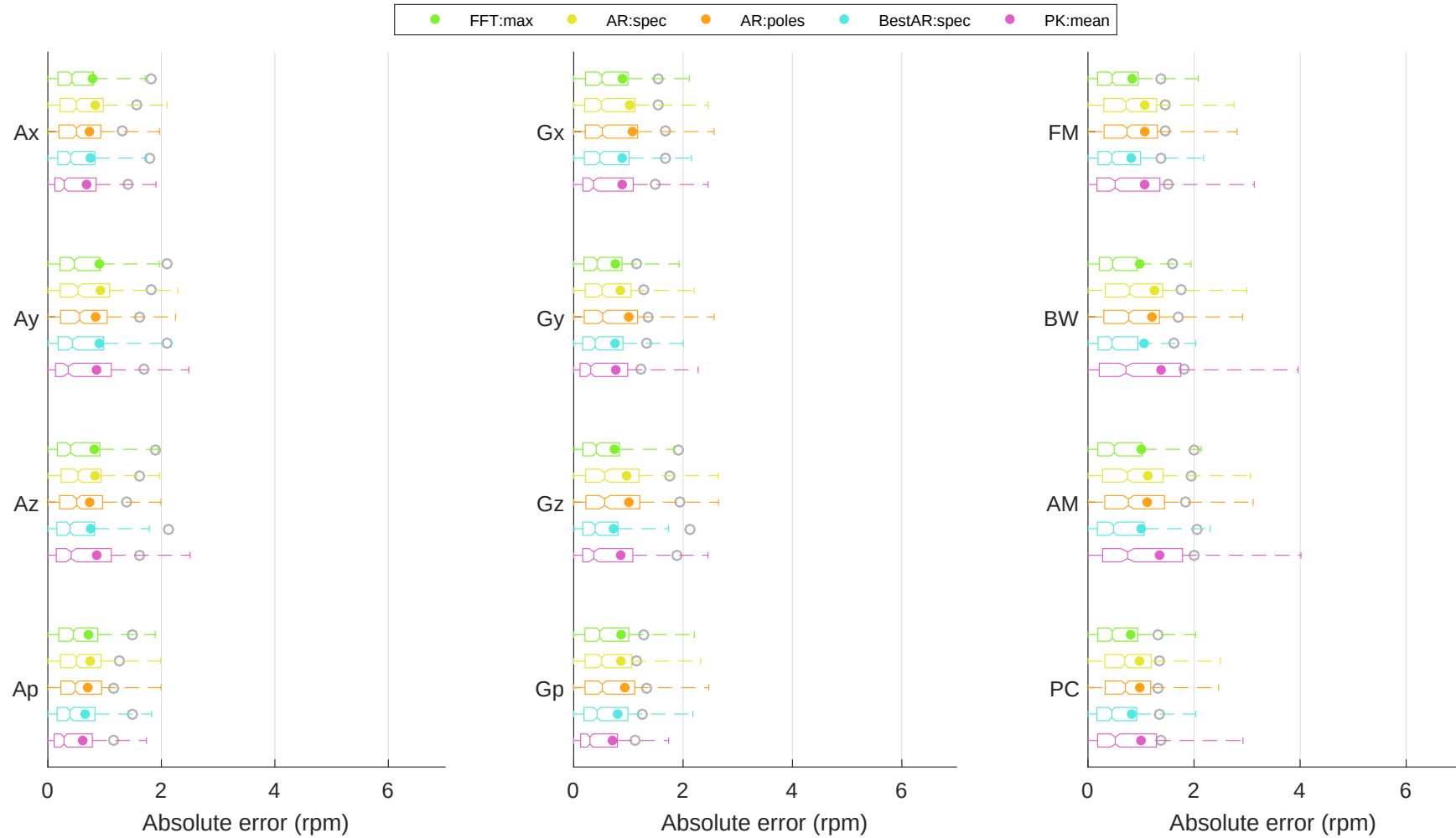


Figure 4.29: Comparison of the Absolute Error (AE) for all respiratory methods (colours) and good quality ($RQI(acf) > 0.4$) respiratory waveforms (y-axis). Error is defined as the reference respiratory rate (from tablet taps) subtracted from the estimated respiratory rate. The boxplot shows distributions of AEs, with lower quartile, median and upper quartile values displayed as left, middle and right vertical lines of the boxes. Whiskers are used to represent the most extreme values within 1.5 times the interquartile range from the central box. Circles represent the mean values. The mean values from all waveforms are shown in grey.

0.14) units for G_x ($p < 0.001$), 0.10 (0.06, 0.16) for G_y ($p < 0.001$), and 0.12 (0.08, 0.17) for G_p ($p < 0.001$). There was no significant correlation between G_z and RQI ($p = 0.178$). For every 10 bpm, RQI(acf) decreased a mean (95% CI) of 0.04 (0.03, 0.05) units for FM ($p < 0.001$), 0.04 (0.02, 0.05) for AM ($p < 0.001$), and 0.02 (0.01, 0.04) for PC ($p < 0.001$). The quality of the BW waveforms was not significantly correlated with HR ($p = 0.264$); however it was positively correlated with RR ($p = 0.013$) with a mean (95% CI) increase in RQI(acf) of 0.04 (0.01, 0.08) units for every 10 rpm. The quality of the accelerometer waveforms was not correlated with RR or HR.

4.5.2 Comparison of devices

Although care was taken to ensure the accuracy of the reference RR, errors may still have persisted. In this section we compare representative waveforms from the pulse oximeter (PC) and smartphone accelerometer (G_p) to each other, rather than the reference RR. As the pulse oximeter and smartphone measure respiration by different mechanisms (neurological feedback and mechanical movement respectively), if the respiratory estimates from these two devices agree, even if inconsistent with the reference RR, we hypothesise that the mean estimate will be the true respiratory rate.

Figure 4.30 shows the absolute error distributions for PC and G_p for all waveforms and for good-quality waveforms only. The pair of waveforms were only considered to be of good quality when both the PC and G_p waveforms were good-quality ($n = 356/567$, 62.8%).

The Kruskal-Wallis test showed that there was a significant difference between the difference distributions when all waveforms are considered ($\chi^2 = 234.5$, $p < 0.001$). The median difference between PC and G_p for the FFT:max method is 0; that is for more than 50% of the recordings the respiratory estimates from PC and G_p are the same. This is more likely for the FFT than other methods as a results of the quantised nature of the output from the Fourier Transform, where RR estimates are computed

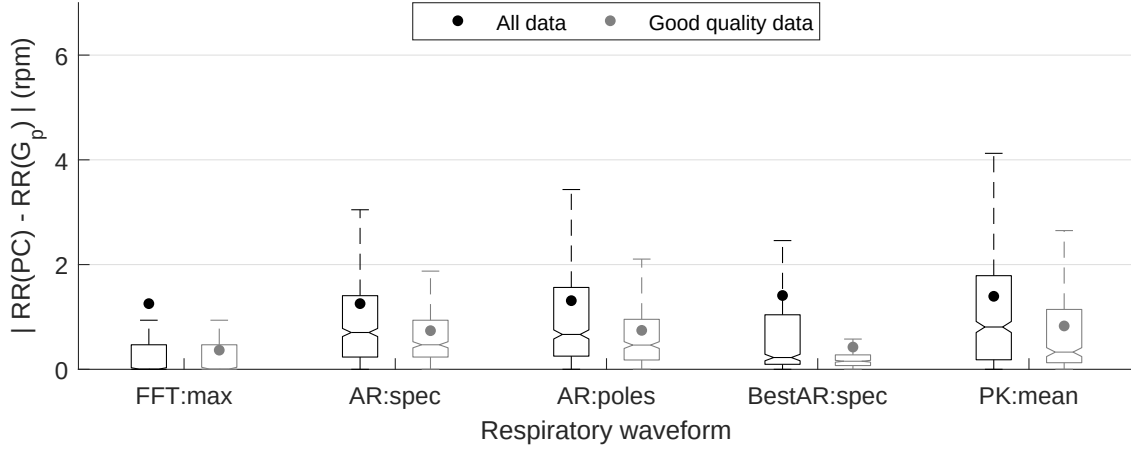


Figure 4.30: Distribution of the absolute difference between respiratory rate estimates from PC (pulse oximeter) and G_p (smartphone accelerometer). The effect of excluding poor quality signals ($RQI(acf) < 0.4$) on the error is shown in grey. The circle represents the MAE.

to the nearest 0.47 rpm⁴. The median AEs from other methods were all significantly larger than FFT:max ($p < 0.001$). When only good-quality waveforms were considered, there was still a significant difference between methods ($\chi^2 = 334.9$, $p < 0.001$). The median AE decreased significantly for all methods ($p < 0.001$), with the largest median decrease from 0.81 rpm to 0.47 rpm. The range of MAEs decreased from between 1.25 and 1.41 rpm to between 0.37 and 0.82 rpm. The MAE changes the most for the FFT:max and BestAR:spec methods, as with the changes observed when estimates were compared to the reference RR, again indicating that large disagreements between respiratory estimates are due to the poor quality of the respiratory waveform.

4.6 Discussion

The methods described in this chapter have sought to explore the robustness of respiratory rate estimates from pulse oximeter and smartphone accelerometer waveforms when monitoring pregnant women. The algorithms were applied to 697 pairs of PPG and accelerometer/gyroscope recordings from 152 pregnant or recently pregnant women during Phase 2 of the 4P study.

⁴A minimum of 512 frequency bins is used in the FFT. For a signal sampled at 4 Hz, each frequency bin is 0.008 Hz or 0.47 rpm.

Respiratory waveforms could be successfully extracted from both the pulse oximeter and smartphone accelerometer. In particular, there were no problems extracting respiration information from the pulse oximeter which relies on the respiratory-induced modulations of blood volume and HR, even during the pregnant or recently pregnant state of the women. Results were similar to those reported in the literature. The lowest MAE from our dataset was 1.13 rpm from the G_p waveform. When poor-quality waveforms were excluded, MAE reduced further to a low of 0.61 rpm from the A_p waveform. Previous studies focusing on respiratory rate extracted from the PPG waveform reported MAEs of 2 rpm [133], 2.8 rpm [86] and 3.4 rpm [108], for example. Estrada *et al.* [101] derived respiratory rate from an accelerometer with a MAE of just 0.5 rpm, although their study only evaluated data from 5 participants. In most of these studies the reference RR was derived from a capnograph, a continuous measure of expired CO_2 , such as in the publicly available Capnobase dataset⁵. Respiratory rate from capnography is likely to be more reliable than the tapping method used in our dataset.

Our results showed that the gyroscope generally performed better than the accelerometer. The lowest error from an individual signal was from the G_y waveform, significantly lower than the error of any individual accelerometer waveform. Other studies have reported similar findings [100]. Yoon *et al.* [134] found errors decreased when gyroscope information was combined with accelerometer information, compared to the use of accelerometer data alone. Initial work on the 4P system 0 data set showed the opposite, whereby the respiratory signal was more prominent in accelerometer channels than in the gyroscope data [124]. Yet this data set suffered from lack of detail about when the midwife count started, and an inaccurate reference respiratory rate (15-second count).

While the respiratory rate estimates from the individual gyroscope channels had lower errors than those obtained from the individual accelerometer channels, the esti-

⁵<http://www.capnobase.org/>

mates derived for the principal components, A_p and G_p , had the lowest errors and there was no significant difference between them. Furthermore, the quality of each principal component waveform was significantly higher than its individual components (i.e. the values of $RQI(acf)$ from the A_p waveform were higher than those for the A_x , A_y , and A_z waveforms). We recommend using the principal component for respiratory rate estimation.

Of the five methods used to derive respiratory rate (the “respiratory methods” - see the lowest set of boxes on Figure 4.20), FFT:max and BestAR:spec reported the lowest errors. PK:mean, the only time-domain based method considered, performed well only when poor-quality signals were excluded. These results contrast with those from a recent review of all published methods for estimating respiratory rate from the PPG and ECG evaluated on a public database of healthy participants [106]. They found that time-domain methods (such as breath detection) performed better than frequency-based methods (such as Fourier, autocorrelation, or the Welch periodogram). We found that for medium-quality recordings the peaks of the time-domain signal were noisy and thus sometimes false peaks were identified. Methods that relied on the frequency-domain were more robust. The review [106] focused on the use of the ECG and PPG to derive a respiratory signal, with analysis methods relying on induced changes in HR and blood volume. These methods may not generalise well to the accelerometer and gyroscope signals in our dataset which directly measure chest movement during respiration.

Although our results showed significant differences in absolute error between waveforms and respiratory methods, these differences are small compared to the values of expected respiratory rates. Furthermore, when poor-quality signals were removed, the MAE for all types of waveform decreased to below 1 rpm and there were even smaller differences between methods. Mean bias was also less than 1 rpm. However, we recognise that a bias is easier to correct for than low precision, and thus bias should not be

the focus when selecting a reliable method for extracting respiratory rate. The current gold-standard method for respiratory rate measurement is manual observation of chest wall movement for 60 seconds [89]. Usually reported to the nearest whole number of breaths, this measurement already has a ± 1 rpm error. Furthermore, in time-pressured environments, clinicians usually only count chest wall movement for 15 seconds which produces a respiratory estimate with a ± 4 rpm margin of error [8]. We argue that a reliable respiratory rate method should be of greater precision than the current manual methods. If waveform quality can be taken into account at the time of measurement, all methods and waveforms meet this requirement. We therefore suggest that a single method should be chosen (such as BestAR:spec) and more pragmatic choices, such as availability and cost, used to select the most appropriate recording device. Alternatively, respiratory rate estimates from both types of devices and all methods investigated in this thesis could be fused using the techniques reviewed in Section 4.2.3.

One limitation of the analysis presented in this chapter is the relatively “normal” range of respiratory rates considered. The reference respiratory rate values ranged from 5.5 to 25.6 rpm. A robust respiratory rate algorithm should be reliable over the entire range of respiratory rate values, particularly as clinical decisions are often made when extreme physiology is exhibited. Within the range of data available, there was evidence that respiratory rate affected the quality of gyroscope waveforms, but not accelerometer or PPG-derived waveforms. A high quality respiratory waveform is the first step to obtaining an accurate respiratory rate estimate. As respiratory rate increases it comes closer to the HR. Low-pass filtering to exclude high-frequency signals such as HR risks attenuating the respiratory rate signal. We anticipate that the quality of the respiratory waveforms from the gyroscope will be compromised more than those from the accelerometer when respiratory rates are high for a number of reasons. The gyroscope measures rotation about an axis rather than displacement, and therefore motion due to

respiration will be relatively low. This means the respiratory rate frequency is of similar magnitude to the HR frequency and other high-frequency noise. Furthermore, when breathing is rapid, it is often also shallow which will further reduce the magnitude of the respiratory signal [135]. Filtering of the gyroscope signal, although successful for the range of respiratory rates available in the 4P dataset, is likely to be more challenging when respiratory rates are abnormally high. A two-step filtering algorithm, similar to that used to re-filter PPG-derived waveforms for suspected *low* respiratory rates, may be suitable. We do not expect low respiratory rates to affect the accuracy of estimation from accelerometer and gyroscope signals.

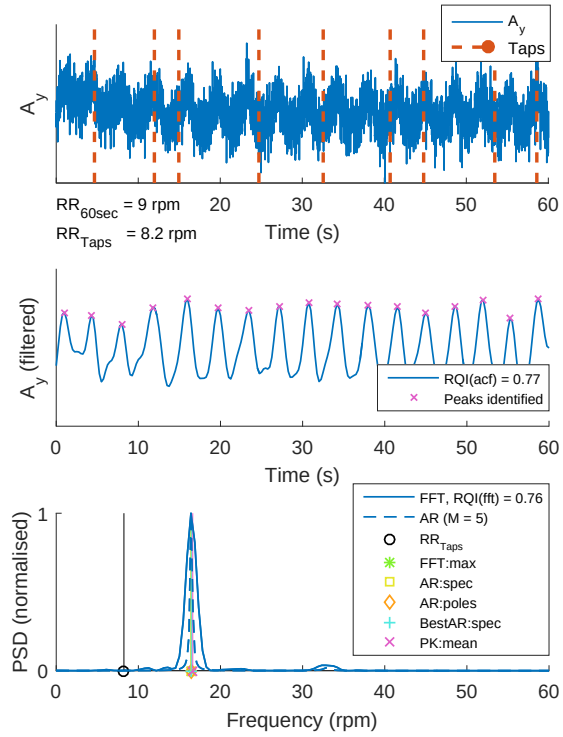
PPG-derived respiratory waveforms rely on the detection of HR beats in order to construct the respiratory waveform, and so are more affected by the HR frequency than the respiratory frequency. Results from regressing HR with waveform quality showed that as HR increased, the quality of the FM and AM PPG-derived waveforms decreased. This change may be due to the quality of the PPG signal itself and the challenge of correctly identifying all the peaks. Further investigation would be needed to establish a clear cause for the correlation. As the accelerometer, gyroscope, and PPG-derived waveforms appear to have different advantages over the range of respiratory rates, fusion of the waveforms may be a way to increase precision, particularly at the physiological extremes. PCA has already been shown to increase waveform quality. Ultimately however, further testing on a dataset with high respiratory rates would be necessary for the full evaluation of the algorithms presented here.

We identified three main sources of error that could have contributed to poor respiratory rate estimation: an incorrect reference respiratory rate, external noise sources outside our control, and noise introduced during pre-processing. Examples of each type of error are shown in Figure 4.31. Firstly, despite careful checking of the midwife respiratory count, there were occasions when it was apparent that the reference respiratory

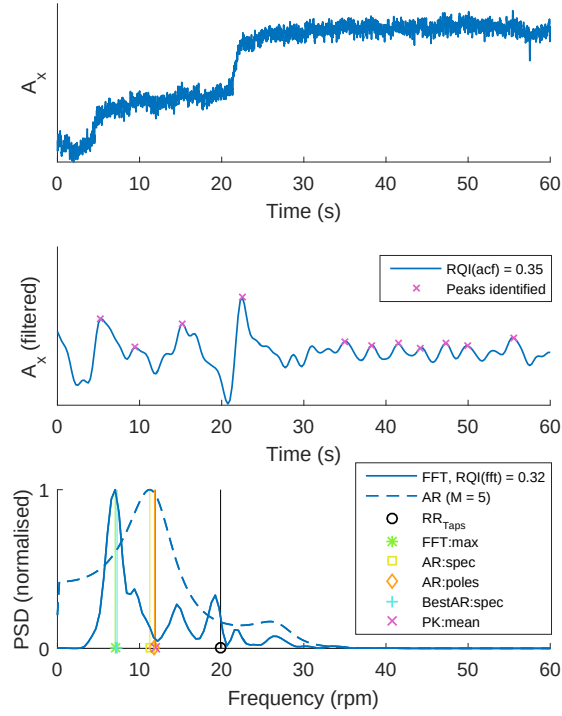
rate, RR_{Taps} , was not correctly recorded. This is a limitation of the available dataset. An example of an incorrect reference respiratory rate is shown in Figure 4.31a. The top plot shows the original accelerometer signal, A_y , with the position of taps indicated by vertical bars. The oscillations of respiration are clear in the accelerometer signal, suggesting several missed taps (at approximately 9, 20, 29, 27, 49, and 57 seconds). Since these missed beats are regular, the respiratory rate calculated from the taps, $RR_{Taps} = 8.2$ rpm, was in close agreement with the midwife reported respiratory rate, $RR_{60sec} = 9$ rpm. However, the filtered accelerometer signal in the middle plot reveals the true position of breaths. Time-domain peak detection and the Fourier transform of the respiratory waveform both indicate a dominant frequency of 16.5 rpm, well above the reported respiratory rate of 9 rpm.

A second source of error is noise introduced by external sources, such as motion artefacts and recalibration of the smartphone accelerometer. Motion artefacts introduce high magnitude spikes into the time-series waveform of the PPG and accelerometer. Corrupted segments of the PPG are removed using the SQI, and low-pass filtering attenuates artefacts from the accelerometer and gyroscope waveforms, however some artefacts still remain.

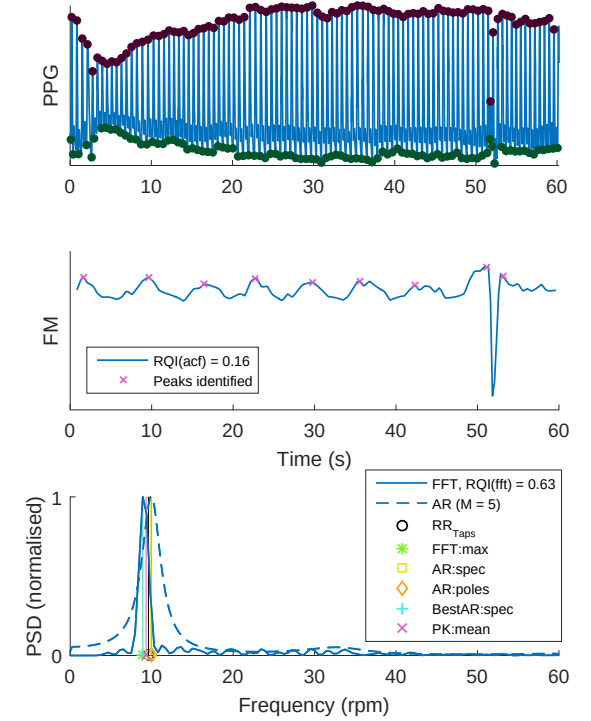
Recalibration of the smartphone accelerometer appears as a sudden baseline shift in the accelerometer channels, and as a high magnitude artefact in the gyroscope channels. The accelerometer suffers more because low-frequency noise is introduced from the baseline shift which overlaps with the respiratory frequency. As the baseline remains steady in the gyroscope, the derivation of respiratory rate is not usually interrupted. Consider the A_x trace in Figure 4.31b. Recalibration, as indicated by a vertical shift, occurs at 4 and 22 seconds. The filtered accelerometer signal is noisy for the first 25 seconds, which causes a large low-frequency peak at 8 rpm in the frequency spectrum. This dominates the peak associated with the true respiratory rate of 20 rpm, thus mak-



(a) Incorrect RR_{Taps} , despite cross-checking with RR_{60sec} .



(b) True respiratory rate not identified due to recalibration shifts in the A_x trace at 4 and 22 seconds.



(c) $RQI(acf)$ suggests a poor quality signal but all respiratory methods converge.

Figure 4.31: Examples of sources of error in respiratory estimation from the pulse oximeter and smartphone accelerometer.

ing respiratory rate estimation erroneous. As far as we are aware, recalibration of the smartphone accelerometer is not controllable, but can be limited by reducing sudden motion.

Thirdly, noise can also be introduced during pre-processing. For the PPG, peak detection is the most common place for errors to be introduced. As all PPG-derived respiratory waveforms rely on peak detection, false identification of peaks or missed peaks can introduce artefacts. For example, in the PPG trace shown in Figure 4.31c, an extra peak has been identified at 52 seconds. Consequently, the FM respiratory signal contains an artefact at 52 seconds. In this example, as only a single erroneous peak has been introduced in the 60-second segment, the final respiratory estimate is not affected.

In summary, respiratory rate can be successfully estimated from a pulse oximeter and smartphone accelerometer, with less than 1 rpm difference from the reference respiratory rate derived from tablet taps when poor-quality waveforms are excluded. These electronic devices are low-cost and readily available, and would allow remote monitoring of respiratory rate. Furthermore, analysis of the signals recorded with these devices leads to estimates with similar errors, so more pragmatic factors should be used to choose the most appropriate device. Measurement precision can be ensured by incorporating information about respiratory signal quality. Further analysis is required to assess the suitability of these algorithms for extreme respiratory rates. In this chapter we only considered respiratory rate recordings of 60-seconds in duration. The same methods could be applied to a 15-second count, which would likely provide a more accurate respiratory rate estimate than a midwife count for the same duration. In the next chapter we consider whether there are any changes in respiratory rate and other vital signs during pregnancy and after delivery.

Modelling vital sign trends

To be able to identify women who are unwell and at risk of developing complications during pregnancy, we require knowledge about normal physiology in pregnancy. The vital-sign data collected in the 4P study form an ideal dataset from which to understand the expected vital-sign trajectories during pregnancy and in the two weeks after delivery. In this chapter we build univariate models in order to develop reference ranges for each vital sign that can be used in obstetric care.

5.1 Reference ranges in obstetric care

Reference ranges are often used in obstetrics, such as in standardised growth charts that describe fetal and child growth [136], or normal changes of cervical length with gestational age [137]. For maternal physiology, reference ranges for antepartum SBP and DBP have previously been presented [72]. However, we have not been able to identify previous attempts at constructing reference ranges for the other vital signs, or for the immediate postpartum period.

Reference ranges differ from standard charts by their descriptive, rather than prescriptive, approach. *Reference ranges* describe observed changes in an unselected population. On the other hand, *standard charts* describe expected physiological changes under optimal conditions. The study population for these charts is therefore highly selected with strict inclusion and exclusion criteria (such as age, BMI and medical his-

	4P cohort n = 620	England 2015-2016 n = 648,107	p-value
<i>Maternal characteristics</i>			
Mean maternal age (years)	31.8	30.3	
White ethnicity (4P phase 2 only)	76.4 %	71.2 %	0.167
Nulliparous	41.3 %	33.4 %	<0.001
<i>Complications developed</i>			
Gestational hypertension	5.3 %	2.5 %	<0.001
Pre-eclampsia	1.8 %	2.1 %	0.348
Gestational diabetes	3.9 %	5.9 %	0.043
<i>Delivery statistics</i>			
Pre-term delivery	7.6 %	8.4 %	0.308
Spontaneous vaginal delivery	62.7 %	59.4 %	0.099
Cesarean-section	17.9 %	15.2 %	0.069
Median length of hospital stay after delivery (days)	1	1	

Table 5.1: Comparison of healthy women in the 4P study with the national English maternal population. The variables included in this table are those that were available from both datasets in the same format. p-values are not displayed for maternal age and length of hospital stay as the variance was not known for the national English population.

tory). The 4P study was designed with minimal exclusion criteria such that the 4P population would be representative of the pregnant population of England. To check this assumption, the two populations are compared in Table 5.1, using data from the 2015-2016 maternal births in England¹. Furthermore, the distribution of maternal age for the two populations is presented in Figure 5.1 in 5-year ranges. Both groups have the largest proportion of women in the 30-34 year old range. The 4P cohort have a narrower range of maternal ages, evident in the smaller proportion of women under 24 years old. This is likely due to the different socio-economic profile of Oxford compared to the national population.

There is a significant difference in the proportion of nulliparous women ($p < 0.001$); 41.3% of the 4P cohort are nulliparous compared to 33.4% in the national population.

¹Data on hospital maternity activity for the period 1 April 2015 to 31 March 2016 is available from NHS England at: <http://content.digital.nhs.uk/catalogue/PUB22384/hosp-epis-stat-mat-tab-2015-16-tab.xls>

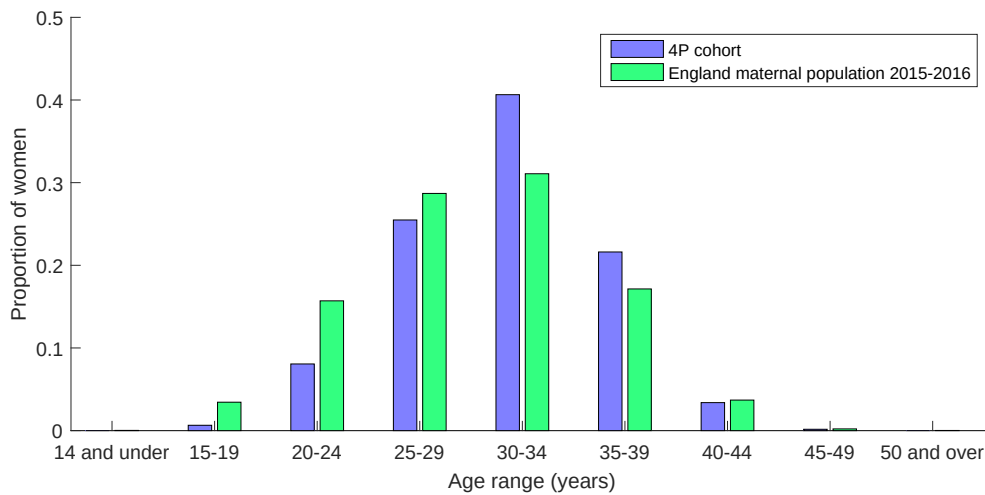


Figure 5.1: *Distribution of maternal age in the 4P cohort compared with the English population.*

The effect of parity on vital signs will be discussed in Section 5.5.1 where it will be shown that the differences in vital signs between nulliparous and multiparous women are small compared to the changes in vital signs associated with pregnancy, and are therefore unlikely to be clinically significant. The difference in proportion of women in the 4P study who developed gestational diabetes was also statistically significant. There is no evidence in the literature that we identified to state that gestational diabetes has a direct effect on vital signs. However, gestational diabetes is correlated with a high BMI [138]. We discuss the effect of BMI on vital signs in Section 5.5.2.

There is also a significant difference in the proportion of women reported to have developed gestational hypertension. We note that the context in which NHS England provide pregnancy-related complication information is that relevant to delivery. Another 1.2% of deliveries in the national data were recorded as “unspecified maternal hypertension”. Furthermore, other studies have reported rates of gestational hypertension between 5 and 10% [14, 31, 72, 32]. We conclude that the two populations can be considered to be broadly similar, and thus anticipate that the centiles resulting from analysis of the 4P data can be *references ranges* representative of the pregnant population in England.

There are a range of statistical methods available for constructing growth charts. A key consideration when using longitudinal data is how the correlation of measurements within a subject is handled. Many methods assume that data points are independently and identically distributed (*i.i.d*), which is equivalent to treating the data as if it were collected in a cross-sectional design, with equal correlation between all points. Longitudinal data is not *i.i.d*, as multiple measurements within a subject are more related to each other than to measurements from another subject. Other modelling approaches build separate models for each subject and then join them together to form a population model. The concern with these methods is that the variance of the parameters associated with groups of only one or two measurements will be close to zero [139].

We identified multilevel models as the most common group of methods for constructing growth charts. These models do not assume *i.i.d* data but instead use a number of levels in a hierarchical structure to model the structure within a clustered or longitudinally collected data set. Each level has a separate variance parameter such that correlation within a subject is modelled in addition to correlation between subjects. The multilevel model handles varying group sizes by weighting the group parameters according to the number of measurements.

Other popular modelling techniques include Gaussian Processes [140], weighted kernel regression [141], and the L-M-S (lambda-mu-sigma) method [142]. These methods all assume *i.i.d* data. Another member of our research group compared these three methods and multilevel models for the development of antenatal vital-sign reference ranges using a subset of the 4P dataset [124]. They found few differences between the outer centiles of the resulting reference ranges. Wade and Kurmanavicius [143] also found few differences between modelling methods when developing fetal growth centiles. However, they concluded that incorporation of the correlation structure (as in multilevel models) is preferable to transforming the data into cross-sectional data (as in models that as-

sume *i.i.d* data), as this could severely affect the precision and accuracy of estimated centiles. We therefore focus on multilevel models as the most popular method for centile construction, and most appropriate for longitudinally collected data.

Within the multilevel structure, various functions can be chosen to model the relationship between the dependent variable (e.g. SBP) and independent variable (e.g. time). For example, the systematic review discussed in Section 2.3.2 identified six studies that used multilevel models to develop reference ranges for SBP and DBP during pregnancy, presented in Table 5.2. These examples model time using fractional polynomials [144], restricted cubic splines [72] and standard polynomials [69, 145].

In this chapter we develop multilevel models from the 4P data to construct vital-sign reference ranges during pregnancy and for the immediate postpartum period.

Year	Author	Aim	Number of women	Time function	Software	Covariates included
2010	Bakker <i>et al.</i> [144] Generation R study (Netherlands)	Investigate the effect of smoking during pregnancy on BP	7106	Fractional polynomials	SAS	• Maternal age (cont.) • Parity (cat.) • Height (cont.) • Weight (cont.) • Education (cat.) • Ethnicity (cat.) • Folic acid supplementation use (cat.) • Alcohol intake (cat.) • Maternal distress (cont.)
2012	Grindheim <i>et al.</i> [61] (Norway)	Describe longitudinal changes in BP and HR during healthy pregnancies	63	Not specified	SPSS	• Parity (cat.) • BMI (cat.) • Weight change (cont.)
2013	Ishikuro <i>et al.</i> [71] BOSHI study (Japan)	Investigate whether home BP and clinic BP measurements are different for nulli- and multi-parous women	690	Not specified	SAS	• Maternal age (cont.) • Parity (cat.) • BMI (cat.) • Alcohol intake (cat.) • Smoker (cat.) • Minimum outside temperature (cont.)
2013	Magriples <i>et al.</i> [69] (Connecticut, U.S.)	Investigate the effect of race, BMI and weight gain on BP in pregnancy	244	Quadratic polynomial	Not specified	• BMI (cat.) • Weight change (cont.) • Prenatal distress (cat.) • Race (cat.)
2015	Macdonald-Wallis <i>et al.</i> [72] AVON study (England)	Develop reference ranges for SBP and DBP during pregnancy	7504	Restricted cubic splines	STATA	• Parity (cat.) • BMI (cat.) • Smoker (cat.)
2015	Kac <i>et al.</i> [145] (Brazil)	Investigate influence of inflammatory biomarkers on SBP and DBP during pregnancy	255	Quadratic polynomial	STATA	• Maternal age (cont.) • Parity (cat.) • BMI (cat.) • Education (cat.) • Self-reported skin colour (cat.) • Smoker (cat.) • Leisure-time physical activity (cat.) • Biomarkers (cat.)

Cat.: categorical/discrete variable; Cont.: continuous variable.

Table 5.2: *Multilevel models for BP during pregnancy identified during the systematic review described in Section 2.3.2.*

5.2 Univariate model building

5.2.1 Dataset

We use antenatal and postnatal vital-sign data collected from the 620 *healthy* women in the 4P study². All women were included even if they only contributed one set of vital sign measurements. For postnatal models, only data collected between days 0 and 15 (inclusive) after delivery are used to develop the model, since relatively few data are available from 16 days after delivery³. The number of measurements for each vital sign used to construct vital sign reference ranges are presented in Table 5.3. We do not distinguish between measurements performed by the midwife and by the woman.

	Antenatal (n)	Postnatal (n)
SBP & DBP	3653	5974
HR	3451	5646
Temp	3159	5836
SpO ₂	3451	5666
RR	3503	901

Table 5.3: Number of measurements for each vital sign used to construct univariate antenatal and postnatal reference ranges. Data was collected in the 4P study (Section 3.3.3). Postnatal trends are only informed by data collected between day 0 and day 15 after delivery, inclusive.

5.2.2 Description of multilevel models

Multilevel models are also called *hierarchical* models because they model hierarchical structures in a dataset using a series of levels. For longitudinal data, such as that collected in the 4P study, the hierarchical structure is in the form of repeated measurement occasions (level 1, subscript j), on a single subject (level 2, subscript i). The model is

²The characteristics of these women are described in Section 3.3. Only 523 women contributed postnatal vital-sign measurements due to women not continuing with the study.

³Less than 200 measurement sets are available from 16 days after delivery compared to around 400 measurements for days 1 to 14. See Figure 3.12 in Section 3.3.3.

defined by population parameters, $\boldsymbol{\beta}$, which describe the shape of the underlying population mean, and subject-specific parameters, $\mathbf{b}_i$, which capture variation in the individual subjects that is not explained by the population model. The multilevel model thus describes observation y_{ij} , the j^{th} observation for the i^{th} subject, by Equation 5.1⁴. For the 4P data, the response variable y_{ij} is a physiological variable (e.g. blood pressure). The number of total measurements $j = 1, \dots, n_i$ may vary between the subjects.

$$y_{ij} = X_{ij}\boldsymbol{\beta} + Z_{ij}\mathbf{b}_i + \epsilon_{ij} \quad (5.1)$$

The shape of the response curve is typically defined as a function of time, x_{ij} , such as a polynomial. Additional covariates may be added at the subject level to explain some of the variation between individuals. The most common covariates included in previous antepartum blood pressure models are BMI, maternal age, parity, and smoking status (Table 5.2). The $\mathbf{X}$ and $\mathbf{Z}$ matrices contain the population (level 2) and subject-specific (level 1) covariates, respectively.

At level 1 of the model, the data points within a subject, $\mathbf{y}_i$, are assumed to be normally distributed about a mean within-subject trajectory, with variance specified by the within-subject (level 1) covariance matrix $\mathbf{R}$. The distribution of data points within subject i is therefore described by:

$$\mathbf{y}_i \sim \mathcal{N}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \mathbf{R}) \quad (5.2)$$

Subject-specific parameters, $\mathbf{b}_i$, are defined as the distance from the population parameter, $\boldsymbol{\beta}$. We assume the subject-specific parameters are normally distributed about

⁴We use the notation commonly adopted in the multilevel literature. Subscript i references all measurements for subject i , which thus selects n_i rows of a matrix or vector. For example, $\mathbf{y}_i$ is a vector of length n_i , a subset of the vector $\mathbf{y}$ which contains all measurements on all subjects (length $\sum_{i=1}^N n_i$). X_{ij} is a row vector, the j^{th} row of the matrix $\mathbf{X}_i$, which is n_i rows subsetted from the matrix $\mathbf{X}$.

a mean of zero, with correlations specified by the between-subject (level 2) covariance matrix, $\mathbf{D}$:

$$\mathbf{b} \sim \mathcal{N}(0, \mathbf{D}) \quad (5.3)$$

At level 2, the distribution of all data points is therefore described by Equation 5.4. Here the variance is made up of two parts, $\mathbf{ZDZ}^T$ and $\mathbf{R}$, which correspond to the between-subject variation and within-subject variation respectively. As the subject-specific covariate matrix, $\mathbf{Z}$, is usually a function of the independent variable (e.g. time), the variance can vary with time.

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{ZDZ}^T + \mathbf{R}) \quad (5.4)$$

The diagram in Figure 5.2 summarises this discussion of multilevel models. Firstly, multilevel models are suitable for hierarchical data where observations are clustered, such as multiple measurements from a single subject in a longitudinal study (panel (a)). Observations can be collected at any time point since a multilevel model allows for a continuous representation of time. Missing data are implicitly imputed under the missing at random assumption. Secondly, panel (b) illustrates the Gaussian distribution assumptions of the linear multilevel model, and further emphasises the hierarchical structure. The variation of observations within a group is assumed to be Gaussian. The group means (level 1) are described by the subject-specific parameters, $\mathbf{b}_i$. These are defined as the variation from the population mean, $\boldsymbol{\beta}$ (level 2). The Gaussian distribution assumptions mean that the final population and subject-specific trajectories can be described by equations (mean of the distributions in Equations 5.4 and 5.2, respectively), and centiles can be calculated using direct formulae (mean trajectory plus a proportion of the variance). Exact calculation of centiles was considered a key criterion for choosing a statistical model for the recent WHO child growth curve project [146].

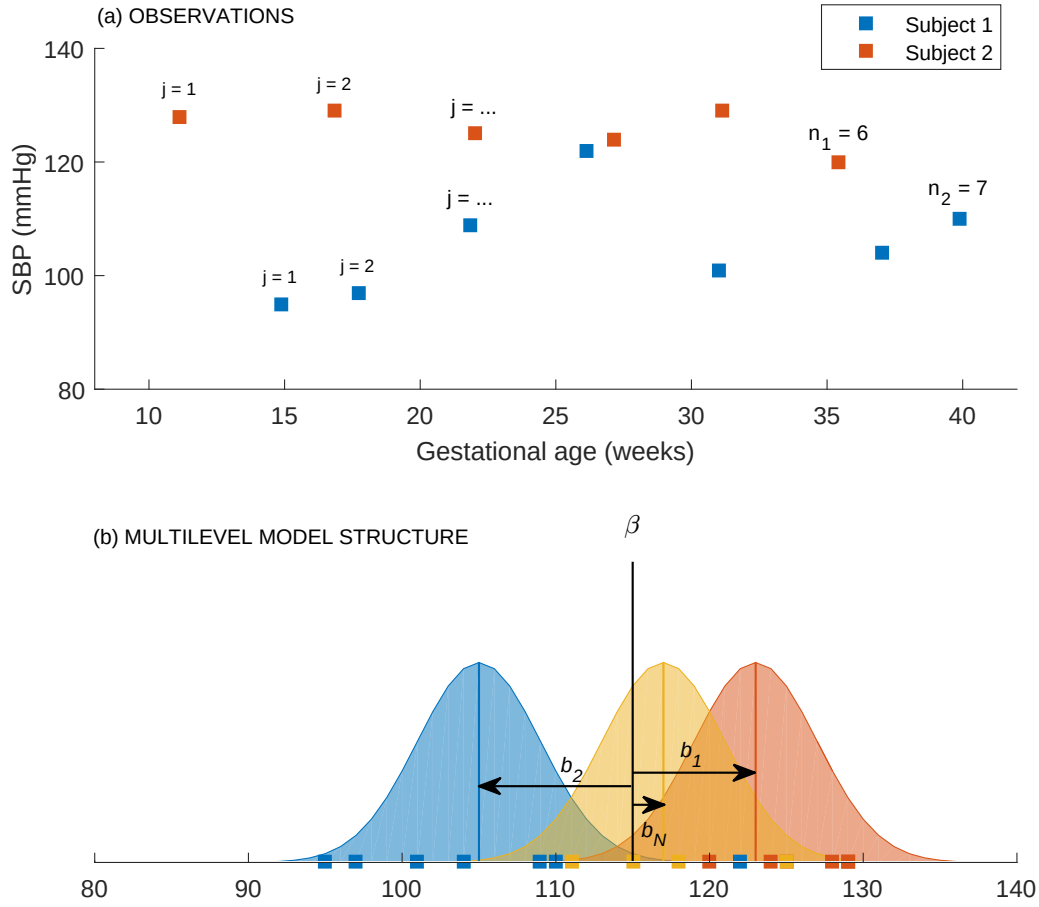


Figure 5.2: Illustration of how the multilevel model represents the hierarchical structure in a dataset. The top panel shows example raw data points from two subjects. The spread of each subject's measurements is modelled as a Gaussian distribution where the mean is some deviation b_i from the population mean (bottom panel).

5.2.3 Functional forms of time considered

We considered three functional forms of time: (a) standard polynomials, (b) fractional polynomials, and (c) restricted cubic splines (as implemented by Macdonald-Wallis *et al.* [72]). Model (c) was only considered for *antenatal* trends. All of these forms are linear with respect to the parameters.

Model (a) Standard polynomials: Standard polynomials of increasing order are the most straightforward way to build complexity when modelling the relationship between two variables. We considered polynomials of order 0 to 4 to represent vital-sign trajec-

tories over time. Models were built using a bottom-up approach, first adding population parameters (polynomial order P) and then subject-specific parameters (polynomial of order Q). We required the polynomial order of the population model to be greater than or equal to the subject-specific polynomial to reduce overfitting (i.e. $P \geq Q$). An example of the model building process for the standard polynomial model (a) is presented in Appendix C.7.1. The model is described by:

$$y_{ij} = \sum_{p=0}^P \beta_p x_{ij}^p + \sum_{q=0}^Q b_{iq} x_{ij}^q \quad (5.5)$$

Model (b) Fractional polynomials: Fractional polynomials extend the polynomial family of curves by allowing non-integer powers and repetition of powers, to better represent the curved relationship between a response variable and continuous covariates. A formal method for implementing fractional polynomials was first set out by Royston *et al.* in 1994 [147]. The polynomial powers are chosen from a restricted set of eight powers, $(-2, -1, -0.5, 0, 0.5, 1, 2, 3)$, where a power of zero is the natural logarithm function $\ln(x_{ij})$. Typically, a single power (p_1) or a pair of powers (p_1, p_2) is chosen to represent the curve. Where repeated powers are chosen ($p_1 = p_2$), the second is multiplied by $\ln(x_{ij})$.

We implement fractional polynomials at both the population and subject-specific levels. The fractional polynomial model is described by:

$$y_{ij} = \beta_0 + \beta_1 x_{ij}^{(p_1)} + \beta_2 x_{ij}^{(p_2)} + b_{i0} + b_{i1} x_{ij}^{(p_1)} + b_{i2} x_{ij}^{(p_2)} \quad (5.6)$$

The most appropriate powers were determined using the R package `\{mfp\}` which uses backwards elimination. All possible fractional polynomial powers are initially included in the model. The significance of each power is evaluated in turn by comparison of the full model with a fractional polynomial model excluding the power being evaluated. If

the difference is not significant ($p \geq 0.05$) the power is excluded. Once two powers remain, the fractional polynomial model is compared with a straight line, and then with a fractional polynomial model with only one power. The best model is selected.

Model (c) Restricted cubic splines: Restricted cubic splines are another method for modelling a curved relationship between a response variable and continuous covariate. A cubic spline defines cubic functions between a set of fixed knots. The slope and second derivatives at the knot points are set to match between cubic splines. Restricted cubic splines additionally require that the outside segments are linear to stabilise either end of the fit. Macdonald-Wallis *et al.* [72] proposed cubic splines to describe the changes in SBP and DBP during pregnancy. We implement their restricted cubic spline model for all antenatal vital signs. The restricted cubic splines are implemented at both the population and subject-specific levels. Knots are set at 11, 18, 30, 36 and 40 weeks of gestation, which yields four equations. Let $sp1, sp2, sp3$ and $sp4$ describe the four splines. The final restricted cubic spline curve is therefore described by:

$$y_{ij} = \beta_0 + \beta_1 sp1_{ij} + \beta_2 sp2_{ij} + \beta_3 sp3_{ij} + \beta_4 sp4_{ij} + b_{i0} + b_{i1} sp1_{ij} + b_{i2} sp2_{ij} + b_{i3} sp3_{ij} + b_{i4} sp4_{ij} \quad (5.7)$$

We use these cubic spline models to investigate differences in parity, discussed in Section 5.5.1.

For models (a) to (c), we assume the residuals are uncorrelated and normally distributed with a constant within-subject variance σ^2 . Therefore the within-subject matrix $\mathbf{R} = \sigma^2 \mathbf{I}$ for all subjects. The between-subject covariance matrix, $\mathbf{D}$, was assigned an unstructured matrix (i.e. without any constraints apart from symmetry and positive definiteness).

5.2.4 Methods for prediction

The models were formulated using a Bayesian approach with priors defined in Appendix C.1. Model parameters were estimated using sampling methods, described in Appendix C.2. Specifically, sampling for all parameters was performed using Stan software (Appendix C.3), with four Markov chains of 2000 samples each, where the first 500 samples were discarded as warm-up iterations. Prediction was performed from 10 to 40 weeks of gestation for antenatal models, and from 0 to 15 days after delivery for postnatal models.

5.2.5 Assessing model fit

For each model, we required that the model parameters had *converged*, to indicate the model was appropriate for the data. Convergence was assessed by checking that the Markov chains were drawing from the same distribution, and that there were enough independent samples. The specific measures calculated are described in Appendix C.4.

Model fit was assessed by visual inspection of the mean trajectory to the raw data and empirical centiles. Empirical centiles are calculated from the raw data each week of gestation, and each half day after delivery. Subject-specific trajectories were also examined to check the appropriate subject-specific polynomial order had been chosen for model (a). To assess fit at the outer centiles, the 3rd and 97th centiles were compared which represents 95% of the population. Formally, model fit was calculated using the Watanabe-Akaike Information Criterion (WAIC) which provides a quantitative measure with which to compare models of different complexity [148]. WAIC is defined in Appendix C.5.

An example of the model fit process using antenatal and postnatal SBP is described in Appendix C.7.2. The WAIC values for the final orders of model (a), (b) and (c) for each vital sign are presented and discussed in Appendix C.7.3.

5.3 Population trends

The standard polynomial multilevel model (a) was chosen as the most appropriate model to describe vital sign changes in the antenatal and postnatal periods of pregnancy (see Equation 5.5). The orders selected for the population polynomial (P) and subject-specific polynomial (Q) are shown in Table 5.4.

	Antenatal orders		Postnatal orders	
	P	Q	P	Q
SBP	2	1	3	2
DBP	2	1	3	2
HR	2	2	3	2
Temp	1	1	1	1
SpO ₂	1	1	1	1
RR	1	1	1	1

Table 5.4: Selected polynomial model orders for the population level (P) and subject-specific level (Q) for the standard polynomial multilevel (model (a)) fitted to the 4P antenatal and postnatal data. Comparison of fit with the fractional polynomial model (b) and restricted cubic spline model (c) is presented in Appendix C.7.3.

Population trends for each vital sign were derived by performing inference at the population level (level 2) of the multilevel model (Equation 5.4). We present the 50th centile (mean trajectory) and outer 3rd and 97th centiles in Figure 5.3 (SBP, DBP and HR) and Figure 5.4 (Temp, SpO₂ and RR). The 3rd and 97th outer centiles are commonly used to assess heteroscedascity in statistical models [74, 124]. The equations for each vital-sign trajectory are presented in Appendix C.8.

SBP and DBP changed significantly during pregnancy. Mean (95% CI) SBP was 116.2 (115.1, 117.3) mmHg at 10 weeks gestation, decreasing to a minimum of 113.7 (113.0, 114.4) mmHg at 21 weeks gestation; a mean change of 2.4 (1.6, 3.3) mmHg ($p < 0.001$). Mean (95% CI) SBP at 40 weeks gestation was 121.0 (119.9, 122.0) mmHg; a mean change of 7.2 (6.2, 8.2) mmHg from 21 to 40 weeks ($p < 0.001$). Mean (95% CI) DBP was 71.8 (71.0, 72.5) mmHg at 10 weeks gestation, decreasing to its lowest value

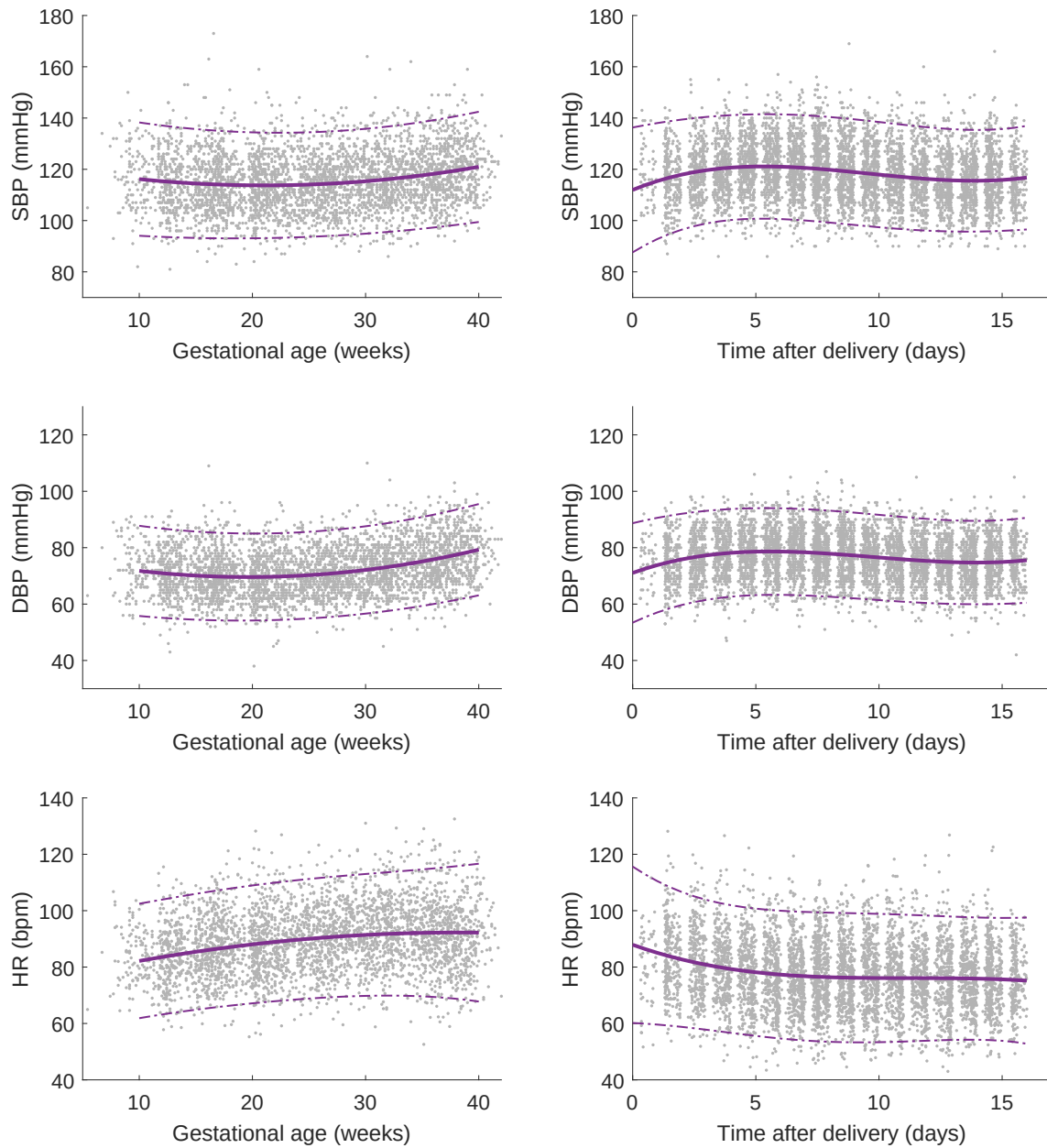


Figure 5.3: Multilevel model, with a polynomial time function, fitted to SBP (top), DBP (middle) and HR (bottom) data from the 4P study in the antenatal (left column) and postnatal (right column) periods. The 3rd, 50th and 97th centiles are shown on top of the raw data.

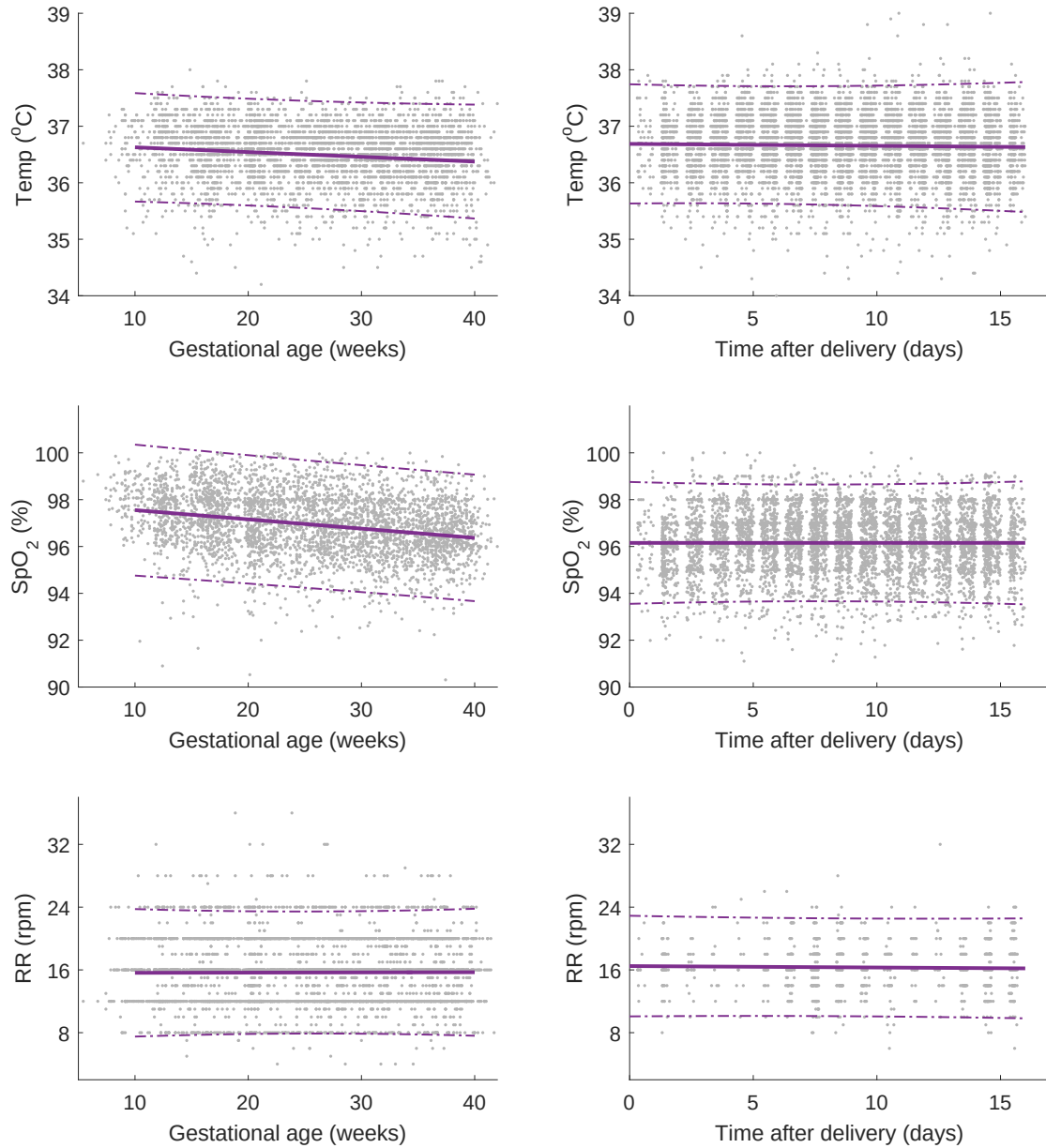


Figure 5.4: Multilevel model, with a polynomial time function, fitted to Temp (top), SpO₂ (middle) and RR (bottom) data from the 4P study in the antenatal (left column) and postnatal (right column) periods. The 3rd, 50th and 97th centiles are shown on top of the raw data.

of 69.6 (69.0, 70.1) mmHg at 20 weeks gestation; a mean change of 2.2 (1.6, 2.7) mmHg ($p<0.001$). At 40 weeks gestation, mean (95% CI) DBP was 79.3 (78.5, 80.0) mmHg, a mean change of 9.7 (9.0, 10.4) mmHg ($p<0.001$) from 20 to 40 weeks gestation. For both SBP and DBP the inclusion of a quadratic term to describe a curved, rather than straight, trajectory was statistically significant ($p<0.001$). The variation in BP values was greater for SBP than for DBP. The mean (95% CI) difference between the 3rd and 97th centiles at 10 weeks gestation was 44.2 (41.9, 46.5) mmHg for SBP and 32.0 (30.3, 33.7) for DBP. The differences did not change significantly during pregnancy ($p=0.232$ for SBP; $p=0.364$ for DBP).

SBP and DBP changes in the immediate postpartum period were also statistically significant. Mean (95% CI) SBP immediately after delivery (day 0) was 112.0 (110.2, 113.7) mmHg, rising to a maximum of 121.1 (120.3, 121.9) mmHg 5.5 days after delivery; a mean change of 9.1 (7.4, 10.9) mmHg ($p<0.001$). 16 days after delivery, mean (95% CI) SBP was 116.7 (115.6, 117.9) mmHg. Mean (95% CI) DBP was 71.1 (69.8, 72.3) mmHg after delivery (day 0), rising to a maximum of 78.7 (78.1, 79.2) mmHg at 5.5 days after delivery; a mean change of 7.6 (6.3, 8.9) mmHg ($p<0.001$). 16 days after delivery, mean (95% CI) DBP was 75.5 (74.7, 76.4) mmHg.

HR increased significantly during pregnancy from a mean (95% CI) of 82.2 (81.2, 83.2) bpm at 10 weeks gestation to 92.2 (91.1, 93.4) bpm at 40 weeks gestation; a mean (95% CI) change of 10.1 (9.0, 11.1) bpm ($p<0.001$). After delivery, mean (95% CI) HR was 87.9 (85.9, 89.9) bpm and had decreased to 76.8 (75.9, 77.6) bpm by 7 days after delivery; a mean (95% CI) change of 11.2 (9.1, 13.3) bpm ($p<0.001$). The difference between the HR outer 3rd and 97th centiles increased significantly during pregnancy, from a mean (95% CI) of 40.6 (38.0, 43.2) bpm at 10 weeks gestation to 48.9 (45.5, 52.2) bpm at 40 weeks gestation; a mean (95% CI) increase of 8.3 (4.6, 12.0) bpm from 10 to 40 weeks gestation ($p<0.001$).

Temperature decreased during pregnancy from a mean (95% CI) of 36.63 (36.59, 36.67) °C at 10 weeks gestation to 36.37 (36.33, 36.42) °C at 40 weeks gestation; a mean (95% CI) change of 0.25 (0.20, 0.31) °C ($p < 0.001$). Immediately after delivery, mean (95% CI) temperature was 36.69 (36.65, 36.73) °C. Temperature did not change significantly in the two weeks after delivery. The distance between the outer 3rd and 97th temperature centiles was a mean (95% CI) of 1.93 (1.83, 2.03) °C in the antenatal period, and 2.09 (2.02, 2.16) °C in the postnatal period; a mean (95% CI) change of 0.16 (0.04, 0.28) °C ($p = 0.006$). The larger variation in the postnatal period may be caused by the readings taken by the women at home (non-expert users). Tympanic thermometers can be sensitive to positioning.

SpO₂ also decreased during pregnancy from a mean (95% CI) of 97.6 (97.5, 97.7) % at 10 weeks gestation to 96.4 (96.3, 96.5) % at 40 weeks gestation; a mean (95% CI) change of 1.2 (1.1, 1.4) % ($p < 0.001$). Immediately after delivery, mean (95% CI) SpO₂ was 96.2 (96.1, 96.3) % and did not change significantly in the two weeks after delivery.

RR did not change significantly during pregnancy or in the two weeks after delivery. At 10 weeks gestation the mean (95% CI) RR was 15.6 (15.3, 16.0) rpm with a mean change of 0.1 (-0.3, 0.5) rpm from 10 to 40 weeks gestation ($p = 0.387$). Immediately after delivery, RR was a mean (95% CI) of 16.5 (16.0, 16.9) rpm; a mean (95% CI) change of 0.8 (0.2, 1.3) rpm ($p = 0.002$) from the end of the antenatal period.

5.3.1 Significance of results

The vital sign trends revealed by the multilevel model show statistically significant changes for SBP, DBP, HR, Temp, and SpO₂ during pregnancy, and SBP, DBP, and HR in the two weeks after delivery.

A mid-pregnancy dip in SBP and DBP has been reported by other studies [61, 71, 72, 69, 58], as has the increase in HR during pregnancy [149, 62]. These changes were summarised and data synthesised in the systematic review presented in Section 2.3.2.

Figure 5.5 compares the BP and HR trends determined from the 4P data with the modelling performed in the systematic review. Both datasets show the same trajectory shape; SBP and DBP rise in the second half of pregnancy, and HR increases gradually throughout pregnancy. The mean vital sign values of the 4P data model are significantly higher than the systematic review data model. The mean (95% CI) difference is 3.8 (2.1, 5.4) mmHg ($p < 0.001$) between the SBP models, 6.9 (5.4, 8.3) mmHg ($p < 0.001$) between the DBP models, and 5.8 (2.2, 9.4) bpm between the HR models. The mid-pregnancy dip occurs later in the 4P data models (21 weeks for SBP; 20 weeks for DBP) compared to the systematic review data models (15 weeks for SBP; 19 weeks for DBP). The studies pooled in the systematic review often had small sample sizes or were performed 20 to 30 years ago, contributing to large heterogeneity in the data. Changes in measurement

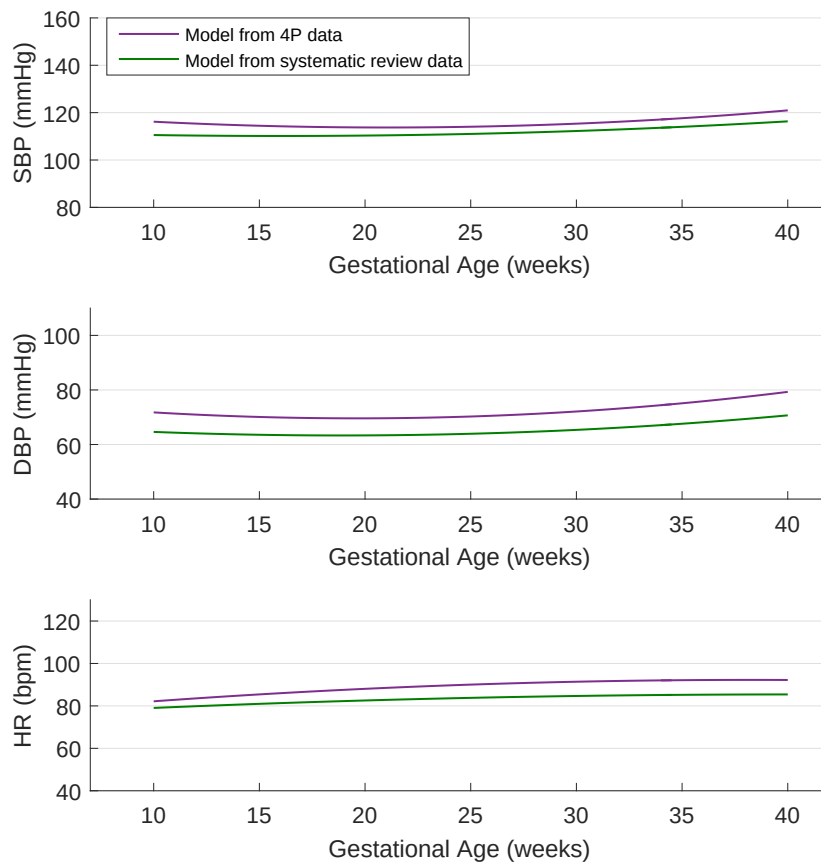


Figure 5.5: Comparison of BP and HR trends derived from the 4P data (model (a), purple) and systematic review data (green), see Section 2.3.2.

technique, changes in population demographics, and the availability of individual patient data, most likely account for the differences observed when compared to the 4P dataset [5].

The difference in Δ_{97-3} , where Δ_{97-3} is the difference between the 97th and 3rd centiles, between antenatal SBP, $[\Delta_{97-3}]_{\text{SBP}}$, and antenatal DBP, $[\Delta_{97-3}]_{\text{DBP}}$, has been previously noted by Macdonald-Wallis *et al.* [72] with analysis of data collected during the AVON study. The AVON study collected data from around 10,000 women during 1991 and 1992, and is currently the biggest published study of maternal BP in the UK.

The postnatal SBP and DBP trajectories from the 4P data both show a transient rise in the 3 to 6 days after delivery, to values similar to those at 40 weeks gestation. This change has previously not been quantified, as no conclusive studies of SBP and DBP during this time period have been reported.

The trends of temperature and SpO_2 contribute a new understanding of maternal physiology. Few studies have reported temperature changes during pregnancy [63], and changes in SpO_2 have not previously been reported. Similarly, postnatal vital sign changes are not widely known [65].

Although these models indicate statistically significant vital sign changes during pregnancy at a population level, this does not necessarily translate into clinical significance. Measurement accuracy and daily or seasonal vital sign variation places small, yet statistically significant, vital sign changes into perspective. For example, automatic BP devices are validated to within 5 mmHg [76, 150], and BP naturally varies by 10 to 20 mmHg in a 24-hour period [59]. The 2.4 mmHg change in mean SBP between 10 and 21 weeks of gestation therefore may not affect clinical practice.

Yet other changes are clinically significant. Obstetric EWS systems are used to monitor vital signs of a pregnant or recently pregnant woman that has been admitted to hospital, and indicates when one or more abnormal vital sign values are recorded. In

current obstetric EWS systems, changes of 10 mmHg for BP and 10 bpm for HR cause a change in the score assigned to that vital sign [36, 151, 152]. In the second half of pregnancy and in the immediate postpartum period, normal changes in SBP and DBP account for over half of these ranges. HR alone changes by more than 10 bpm during pregnancy. Thus it seems likely that taking account of expected values for that gestation would prompt changes in clinical care. The translation of population vital sign changes into a clinical tool will be discussed in Chapter 6.

5.4 Subject-specific trends

Inference in the multilevel model can also be performed at the subject-specific level (level 1; Equation 5.2). The subject-specific parameters, $\mathbf{b}_i$, describe the trajectory for subject i . As previously described, these parameters are assumed to be normally distributed with mean 0 and variance described by the within-subject covariance matrix, $\mathbf{D}$ (Equation 5.3). We can therefore assign a likelihood to each b_i which indicates how close the parameter lies to the mean 0, and therefore how similar the i^{th} subject's trajectory is to the population mean trajectory (see Appendix C.2). A high likelihood indicates the subject-specific parameter lies near the mean and therefore the trajectory will closely follow the population mean. A low likelihood indicates that the subject-specific parameter lies far from the mean and therefore the trajectory will be different from the population mean.

Figure 5.6 presents the distribution of antenatal SBP subject-specific parameters and the underlying probability density distribution (with variance $\mathbf{D}$). The highlighted points are examples of parameters with high likelihood that lie close to the mean (blue), and low likelihood that are far from the mean (red). The examples with low likelihood all have high blood pressure, but are not consistent in how they differ from the population mean. One subject ($\mathbf{b}_i = \{10.0, 0.48\}$) starts close to the population mean but rises

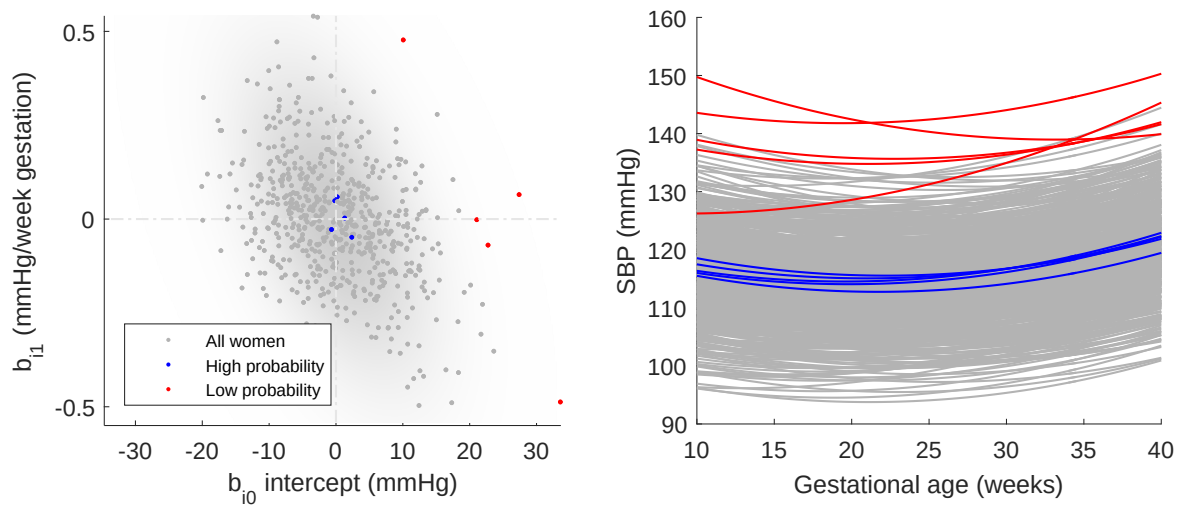


Figure 5.6: Distribution of the subject-specific parameters, $\mathbf{b}$, for antenatal SBP (left) which describe the trajectory over weeks of gestation (right). Antenatal SBP has two subject-specific parameters (intercept, b_0 , and slope, b_1) and therefore $\mathbf{b}$ is a 2-D distribution. The highlighted points show examples of subject-specific parameters with high likelihood which lie near the population mean (blue) and subject-specific parameters with low likelihood which are far from the population mean (red).

quickly; while another subject ($\mathbf{b}_i = \{33.50, -0.49\}$) starts far from the population mean but has a negative slope, so becomes closer to the population mean with time. Four of the five women highlighted with low likelihood were diagnosed with gestational hypertension or pre-eclampsia during their pregnancy.

5.4.1 Subgroup investigation

Subject-specific trends inferred from the multilevel model can enable insight into subgroups of women. We illustrate this with two groups of women: those that developed a hypertensive disorder of pregnancy (antenatal trends); and those that delivered preterm (postnatal trends). These are exploratory analyses as the number of women in the affected groups are small. Information available prospectively can be incorporated into a multilevel model and allows comparison between groups, as will be investigated for parity and BMI in Section 5.5.

Hypertension in pregnancy

Women who develop gestational hypertension, pre-eclampsia or severe pre-eclampsia are described as developing a hypertensive disorder of pregnancy (HDP)⁵ [31]. 46/620 (7.4%) of the women in the 4P study developed HDP. The antenatal SBP and DBP trajectories for these women are shown in Figure 5.7. Predictably, the blood pressure trajectories for both SBP and DBP were higher for women with HDP than for those who did not develop hypertension. The middle panel shows the marginal distribution over the two dimensions of the subject-specific parameter, $\mathbf{b}$. The marginal distribution of the intercept parameter, $p(b_{i0})$, shows a right shift for women with HDP. That is, by 10 weeks of gestation, the women who develop hypertension already have higher BP than pregnant women who do not go on to develop HDP.

The BP shift for HDP women is even more prominent when the interaction between the intercept and slope parameters is observed, for example along the dominant axis of the conditional distribution $p(\mathbf{b}_i|\mathbf{D})$ (in the top panel of Figure 5.7 the dominant axis is in the diagonal direction, not explicitly shown). For instance, SBP subject-specific parameters for women with HDP either have a high intercept and low slope (SBP starts high and remains high), or have an intercept below the mean and high slope parameter (SBP starts low and rises quickly during pregnancy).

Preterm deliveries

Women who deliver before 37⁺⁰ weeks of gestation are considered to have delivered *preterm*. Preterm birth is the single biggest cause of neonatal mortality and morbidity in the UK, affecting around 7% of live births [153]. Preterm delivery may be elected due to maternal or fetal risk (25% of cases), or outside the control of women or health professionals [154]. Multiple gestations and a maternal age greater than 35 years are risk factors for preterm birth.

⁵Hypertension as a pregnancy-related complication was introduced in Section 2.1.2.

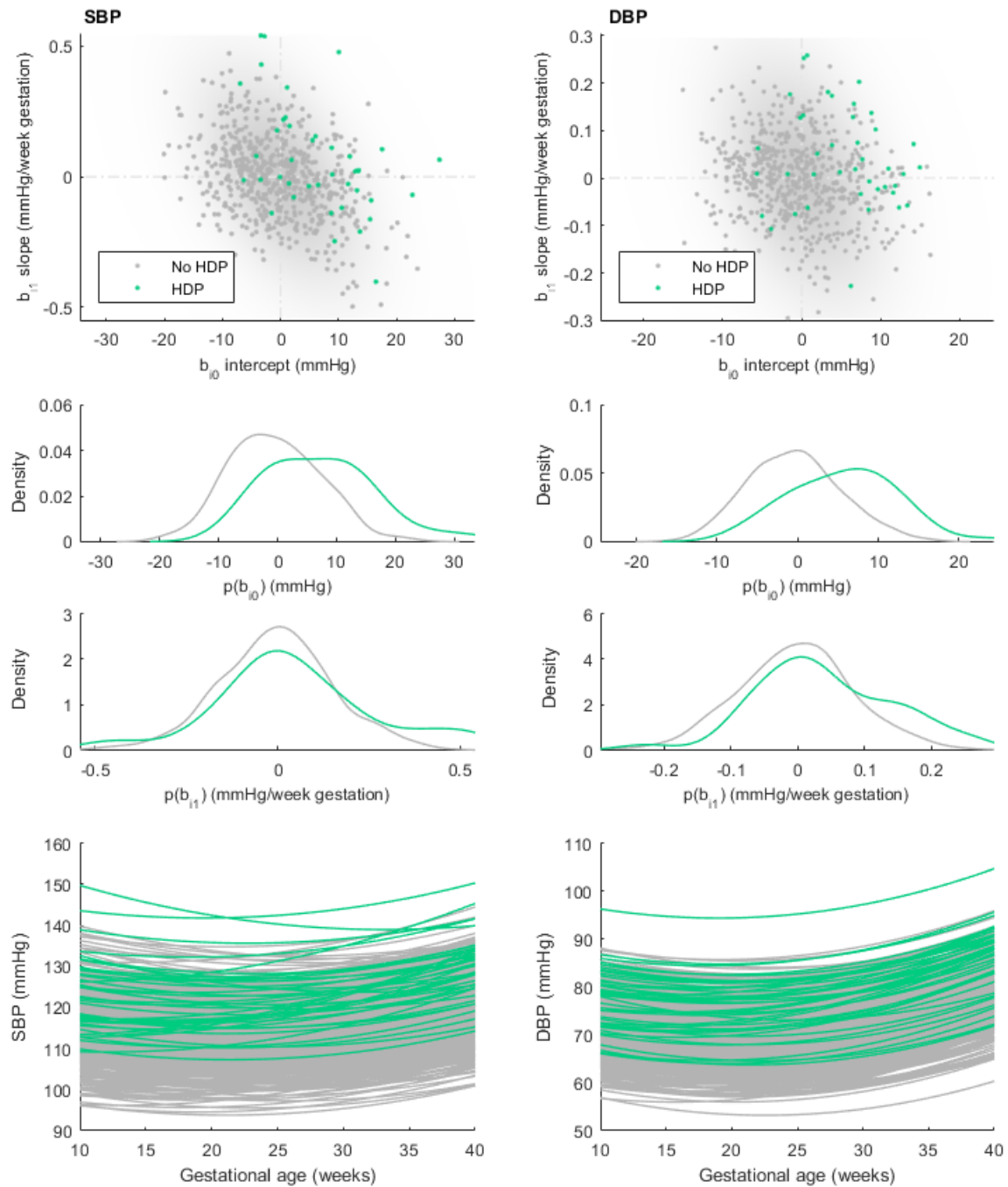


Figure 5.7: Antenatal SBP (left) and DBP (right) information for those women who developed gestational hypertension, pre-eclampsia, or severe pre-eclampsia during pregnancy. The top and middle panels show the distribution of subject-specific parameters, and the bottom graphs display the corresponding trajectories over weeks of gestation.

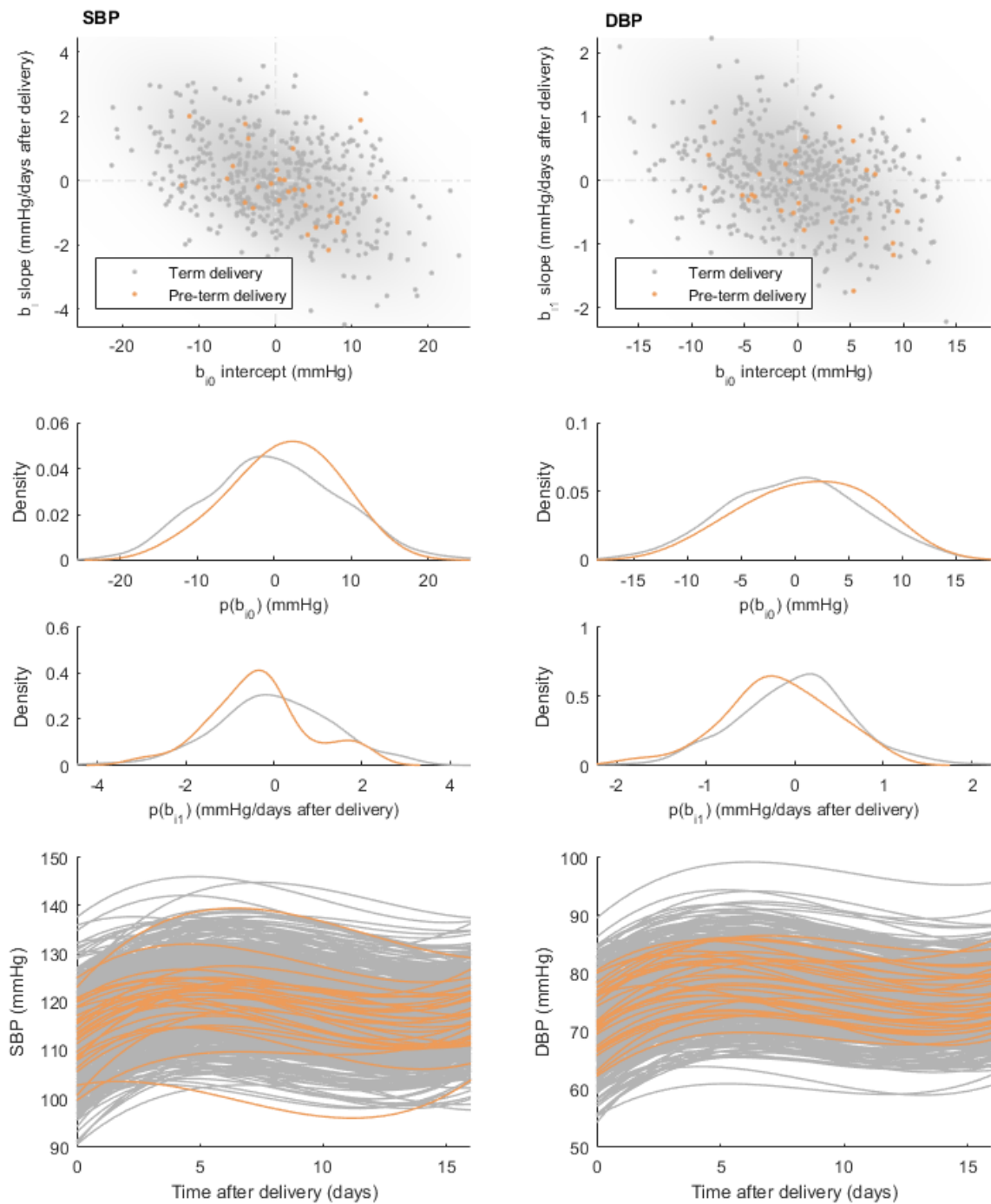


Figure 5.8: Postnatal SBP (left) and DBP (right) information for those women who delivered pre-term (before 37 weeks of gestation). The top and middle panels show the distribution of subject-specific parameters, and the bottom graphs display the corresponding trajectories over the days after delivery.

In the 4P study, 45/538 (7.6%) women delivered preterm. The postnatal SBP and DBP trajectories for these women are shown in Figure 5.8. Note that although postnatal SBP and DBP have three subject-specific terms (standard polynomial order $Q = 3$), only the first two terms (b_{i0} intercept, and b_{i1} slope) are displayed. There is a negative correlation between the intercept and slope parameters for both SBP and DBP. This means that women who have a high BP immediately after delivery are likely to experience a greater decrease in BP over time than those who have a relatively low BP. The subject-specific parameters of the women who delivered preterm are dispersed evenly among the parameters of those who delivered at term. The marginalised distribution and trajectories over time for the two groups are therefore similar.

5.5 Inclusion of additional covariates

Statistical models can incorporate information thought to influence the outcome variable. Previous models of BP during pregnancy adjusted for maternal age, parity, BMI, smoking habit, and exercise habits, among others (see Table 5.2). It is thought that these factors may influence vital-sign readings and so need to be accounted for when combining information from many women. The inclusion of such information may also allow reference ranges to be more descriptive of a particular group of women, or allow an early warning score tool to better identify women with abnormal values. The development of an obstetric early warning score will be discussed in Chapter 6.

In this section we investigate the difference between subgroups in the data and discuss whether this information should be included as a covariate in vital-sign models. We approach this in two ways: by fitting separate models to data for different categories; and by including the categories as a covariate in the model. The first approach will be illustrated using parity, and the second approach illustrated using BMI as a continuous variable.

5.5.1 Parity

Women in the 4P study ranged from a parity of 0 (256/620, 41.3%) to a parity of 5 (3/620, 0.5%). Maternal age increased with parity from 30.7 years (parity 0), to 33.6 years (parity 5). Statistical models reported in the literature usually treat parity as a dichotomous variable, such that nulliparous women (parity = 0) are separated from multiparous women (parity ≥ 1). This is because women in their first pregnancy have been shown to have different labours and higher rates of maternal complication than those women with parity 1 to 3 [155]. After parity 4 the risks of obstetric complication and other adverse outcomes again increase, although the number of women with a parity ≥ 4 is usually small (9 women in the 4P cohort, 3% of the multiparous group). It is therefore not appropriate to treat parity as a discrete counting variable. Of the six papers presented in Table 5.2 that modelled BP in the antenatal period, three included parity as a covariate [144, 61, 145], and two separated women by parity [71, 72]. All five of these studies considered parity to have two categories, separating nulliparous from multiparous women.

We fitted separate standard polynomial multilevel models to each vital sign for nulliparous and multiparous women, using the polynomial orders determined in Appendix C.7.3 (see Table C.1, model (a)). Figure 5.9 shows the 3rd, 50th, and 97th centile trajectories for antenatal and postnatal SBP, DBP, and HR.

Vital sign differences between nulliparous and multiparous women were greatest in SBP, DBP, and HR; although few differences were statistically significant. For SBP and DBP, the mean difference between nulliparous and multiparous women at 10 weeks gestation was not significant ($p=0.060$ and $p=0.224$ for SBP and DBP respectively). At 40 weeks gestation, the mean (95% CI) difference between nulliparous and multiparous women for SBP was 3.0 (0.9, 5.1) mmHg ($p=0.002$), and for DBP was 3.1 (1.6, 4.6) mmHg ($p<0.001$). The change in DBP from the nadir to its value at 40 weeks gestation

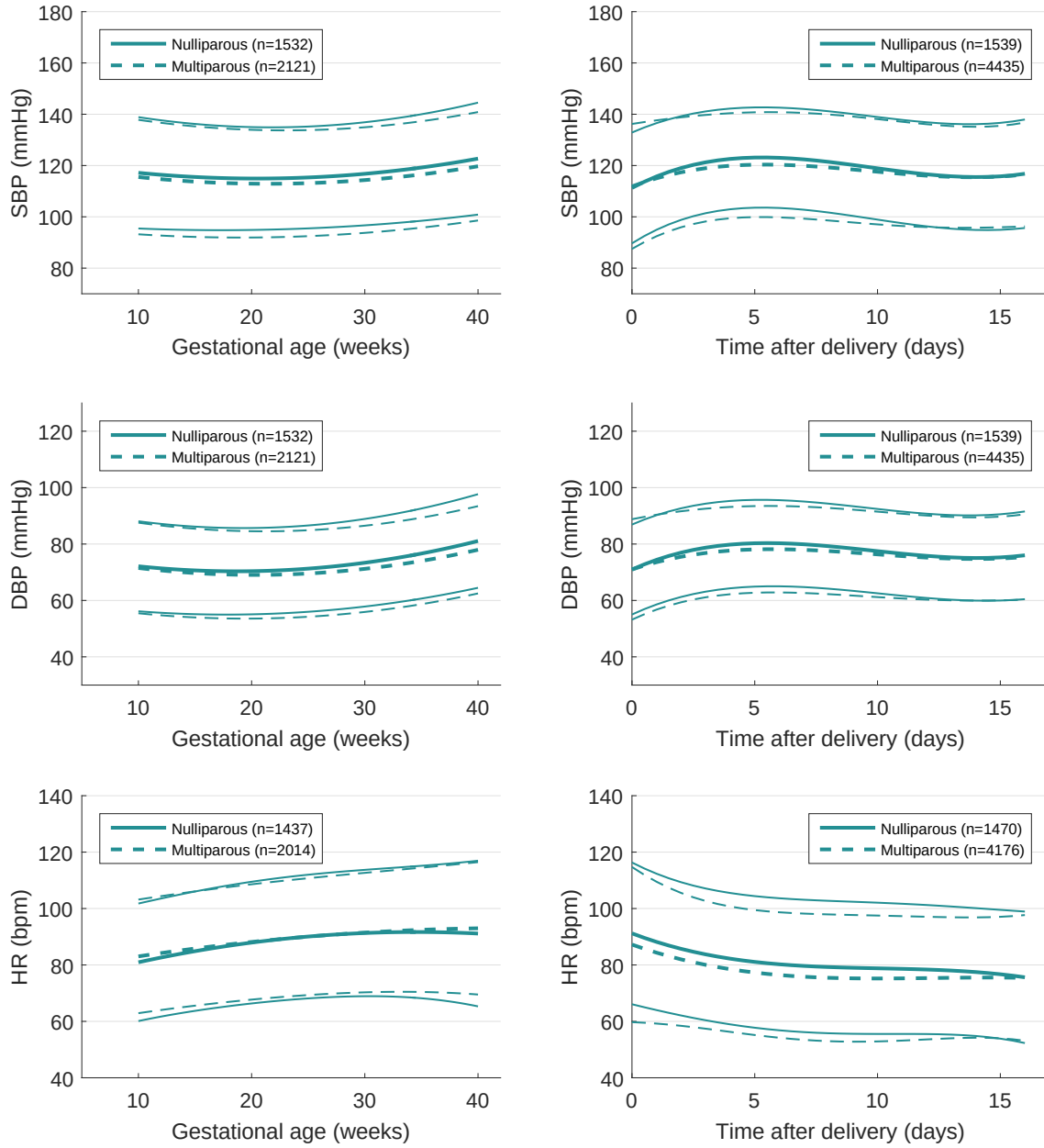


Figure 5.9: Standard polynomial multilevel models fitted separately to vital sign data from nulliparous (solid line) and multiparous (dashed line) women. The 3rd, 50th and 97th centiles are shown for antenatal (left column) and postnatal (right column) vital sign trends of SBP (top), DBP (middle) and HR (bottom).

was significantly greater for nulliparous women than for multiparous women, a mean (95% CI) difference of 1.8 (0.4, 3.2) mmHg ($p=0.007$). The shape of the postnatal SBP and DBP trajectories is different between nulliparous and multiparous women. The mean (95% CI) difference of the slope parameter, β_1 , between the nulliparous and multiparous women is 1.5 (-0.1, 3.1) mmHg ($p=0.033$) for SBP and 1.0 (-0.1, 2.1) mmHg ($p=0.038$) for DBP, where nulliparous women had a greater BP increase than multiparous women. Some of the difference in shape could be attributed to difference in the number of data points ($n = 1,532$ and $n = 2,121$ for nulliparous and multiparous women respectively).

For HR, the mean (95% CI) difference between nulliparous and multiparous women was 2.1 (0, 4.1) bpm ($p=0.023$) at 10 weeks gestation, and 1.9 (0.5, 4.2) bpm ($p=0.063$) at 40 weeks gestation. Immediately after delivery the mean (95% CI) difference was 4.0 (-0.7, 8.6) bpm ($p=0.046$). This difference in HR after delivery is maintained until approximately day 14.

The differences in SBP, DBP, and HR, between nulliparous and multiparous women reported here are small compared to the changes in vital signs associated with pregnancy (see Section 5.3). These differences are therefore unlikely to be clinically significant.

The other vital signs, Temp, SpO₂ and RR, showed little difference between trajectories for nulliparous and multiparous women.

Comparison to AVON study data

In 2015, Macdonald-Wallis *et al.* [72] investigated the effect of parity on SBP and DBP during the antenatal period, using longitudinal data collected from a large cohort of women ($n = 7,504$) enrolled in the AVON study (Bristol, UK). They fitted a cubic spline multilevel model separately to nulliparous and multiparous women in the AVON cohort. This model was described as model (c) in Section 5.2.3. Here we compare the results of their model with the same restricted cubic splines fitted to the nulliparous and multiparous women in the 4P cohort. Figure 5.10 shows the mean SBP and DBP

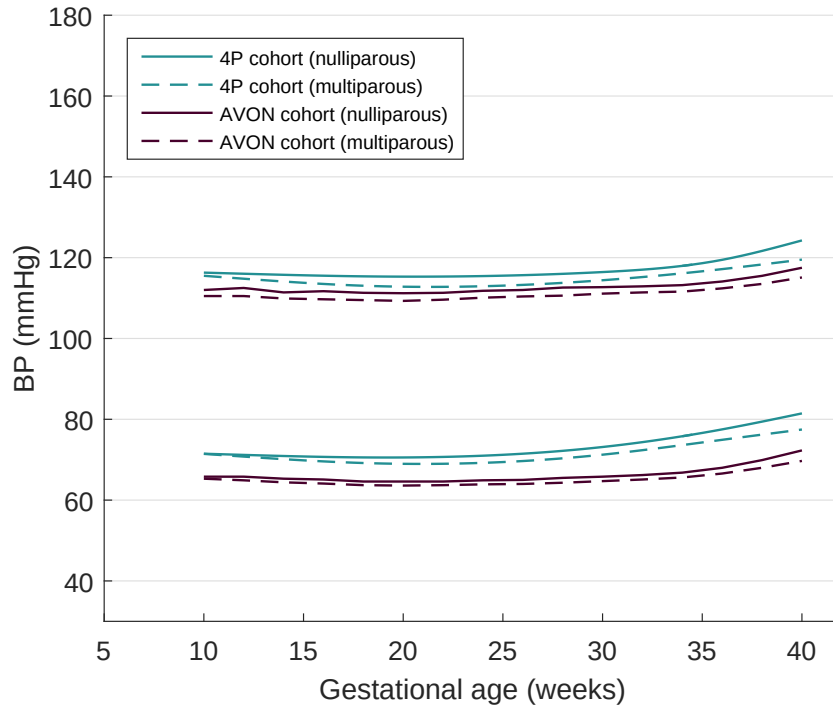


Figure 5.10: Comparison of cubic spline multilevel model fitted to 4P (green) and Macdonald-Wallis et al. (MW) (purple) measurements of SBP and DBP. Separate models are fitted to nulliparous women (solid line) and multiparous women (dashed line).

trajectories from model (c) for both cohorts of women.

Both the 4P and AVON cohort models showed the same relationship between nulliparous and multiparous women. Nulliparous women recorded higher BP values than multiparous women, and the difference increased during gestation. The difference in mean trajectory between the nulliparous and multiparous SBP models was 0.8 mmHg (4P) and 1.5 mmHg (AVON) at 10 weeks gestation, and 4.8 mmHg (4P) and 2.4 mmHg (AVON) at 40 weeks gestation. The difference in mean trajectory between the nulliparous and multiparous DBP models was 0 mmHg (4P) and 1.5 mmHg (AVON) at 10 weeks gestation, and 4.0 mmHg (4P) and 2.4 mmHg (AVON) at 40 weeks gestation. Confidence intervals for the mean trajectory of the AVON model is not available, and so we are unable to determine the statistical significance of these differences.

The 4P cohort model records higher mean SBP and DBP values than the AVON cohort model. The difference in mean for the SBP models was 4.3 mmHg (nulliparous)

and 5.0 mmHg (multiparous) at 10 weeks gestation, and 6.8 mmHg (nulliparous) and 4.4 mmHg (multiparous) at 40 weeks gestation. The difference in means for the DBP models was 5.7 mmHg (nulliparous) and 6.1 mmHg (multiparous) at 10 weeks gestation, and 9.2 mmHg (nulliparous) and 7.8 mmHg (multiparous) at 40 weeks gestation. Examination of the raw centiles⁶ reveal that this difference is present in the raw data, and may be explained by the difference in clinical care between 1991-1992 when the AVON study was performed, and 2012-2016 during the collection of the 4P study data. The shape of the trend between the studies is similar, with a rise in SBP and DBP towards the end of pregnancy, slightly greater in the 4P cohort.

5.5.2 BMI

BMI is a measure of obesity and a high BMI has been associated with adverse pregnancy outcomes [156, 157]. NICE defines four categories of BMI: $< 18.5 \text{ kgm}^{-2}$ as underweight, $18.5 - 24.9 \text{ kgm}^{-2}$ as normal weight, $25.0 - 29.9 \text{ kgm}^{-2}$ as overweight, and $> 30.0 \text{ kgm}^{-2}$ as obese [158, 159]. BMI is usually treated as a categorical variable with women grouped into these four categories [69], or as a dichotomous variable for $\text{BMI} < 25$ (underweight and normal weight) and $\text{BMI} \geq 25$ (overweight or obese) [61, 71].

In the 4P study, BMI of the women is collected at recruitment, usually between 11 and 14 weeks of gestation. Here we include recruitment BMI as a *continuous* variable as this includes all of the available information. We include an intercept term, a quadratic term, and the interaction between BMI and time (i.e. gestational age for an antenatal model) at level 2 (population level) of the standard polynomial multilevel model. Unfortunately BMI is only recorded once in the 4P study, so the change in BMI during pregnancy cannot be included. BMI is transformed using a linear transformation to remove the mean, thus let $m = \text{BMI} - 24.0$. For antenatal SBP, standard polynomial orders of $P = 2$ and $Q = 1$ for the population and subject-specific levels respectively were chosen (see

⁶Macdonald-Wallis *et al.* reported empirical centiles for every two weeks of gestation.

Appendix C.7.3). The standard polynomial multilevel model including BMI is described by:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2 + \beta_3 m_i + \beta_4 m_i^2 + \beta_5 m_i x_{ij} + b_{i0} + b_{i1} x_{ij} \quad (5.8)$$

To illustrate the effect of BMI on vital sign trajectories, in Figure 5.11 we present the predictive curve over time at a BMI of 18.5, 25.0, 30.0, and 35.0 kgm^{-2} for antenatal and postnatal SBP, DBP, and HR (50th centile only, outer centiles are not shown). BMI has the greatest impact on antenatal SBP. Mean (95% CI) difference between the 50th centiles of the SBP model for BMI = 18.5 and the SBP model for BMI = 35.0 was 17.4 (14.1, 20.7) mmHg at 10 weeks gestation ($p < 0.001$) and 9.6 (6.2, 12.9) mmHg at 40 weeks gestation ($p < 0.001$), where a low BMI is associated with a low SBP and a high BMI is associated with a high SBP. Similarly for the DBP models, mean (95% CI) difference was 12.7 (10.4, 15.1) mmHg at 10 weeks gestation ($p < 0.001$), and 8.0 (5.5, 10.6) mmHg at 40 weeks gestation ($p < 0.001$). The decrease in difference between the BMI = 18.5 model and BMI = 35.0 model for both SBP and DBP during pregnancy is the effect of the quadratic BMI term, $\beta_4 m_i^2$, which describes how the curvature of the trajectory changes as BMI changes. The opposite trend is observed for antenatal HR. The mean (95% CI) difference between the 50th centiles of the HR model for BMI = 18.5 and the HR model for BMI = 35.0 was 0.3 (-3.1, 3.7) bpm at 10 weeks gestation ($p = 0.444$), but increased to 8.3 (4.4, 12.2) bpm at 40 weeks gestation ($p < 0.001$). The 50th centile of the HR model for BMI = 35.0 had the greatest change, increasing from 83.2 (80.4, 86.0) bpm at 10 weeks gestation to 98.3 (95.3, 101.4) bpm at 40 weeks gestation; a change of 15.2 (12.2, 18.1) bpm ($p < 0.001$).

Postnatal vital-sign trends show relatively little difference between the different BMIs. Immediately after delivery, mean (95% CI) difference between the 50th centiles of the model for BMI = 18.5 and the model for BMI = 35.0 was 10.0 (6.3, 13.6) mmHg for SBP ($p < 0.001$) and 5.3 (2.4, 8.2) mmHg for DBP ($p < 0.001$). As BMI changes, the

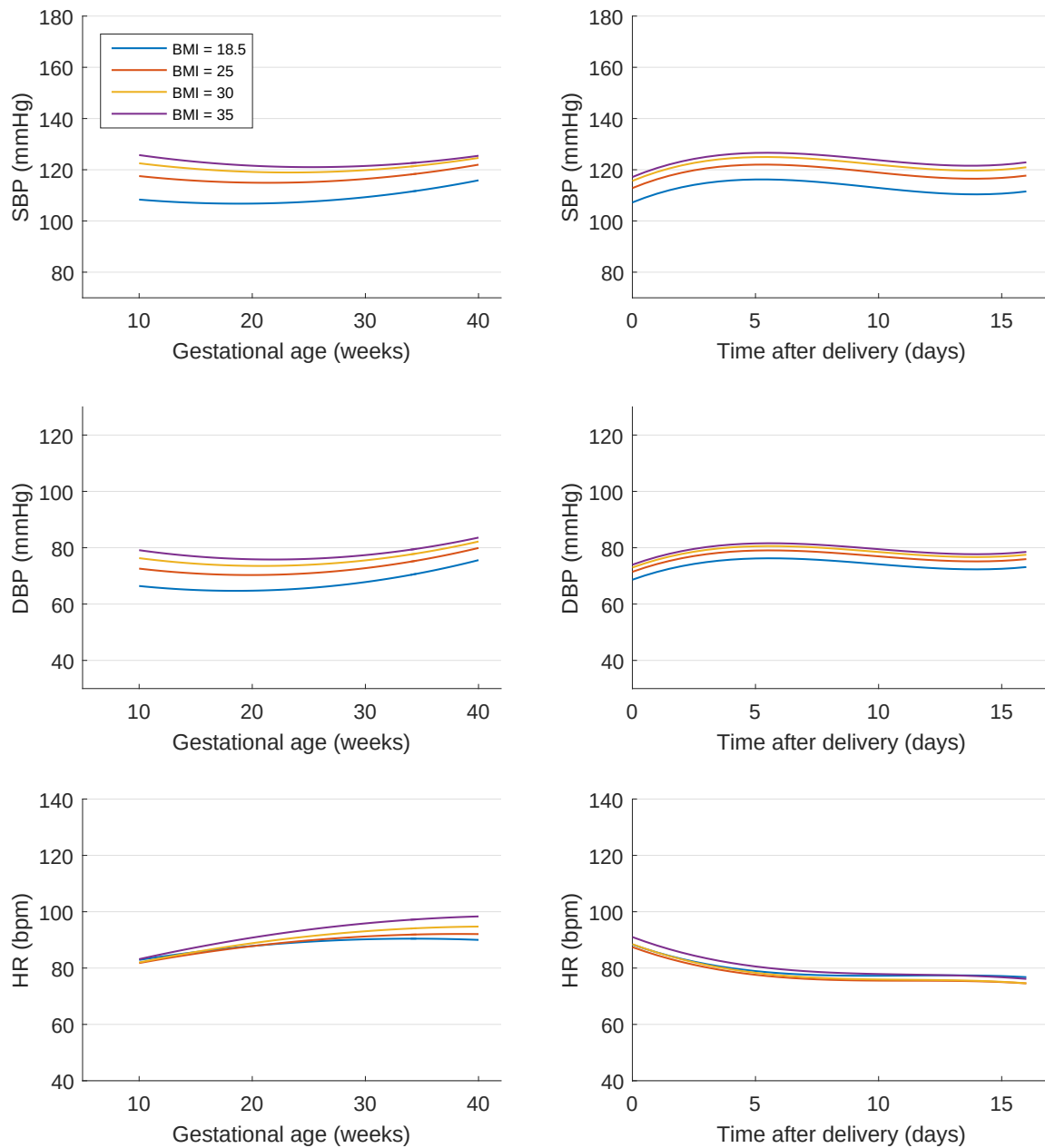


Figure 5.11: Mean trajectories of four different BMIs for antenatal (left) and postnatal (right) data for vital signs SBP (top), DBP (middle) and HR (bottom). BMI was included in a standard polynomial multilevel model as a continuous covariate.

shape of the vital-sign trajectory does not change significantly. BMI has little effect on Temp, SpO₂ and RR.

The treatment of BMI as a continuous variable in a multilevel model produces useful results. A quadratic BMI term allowed the model to describe changes in shape trajectory, particularly relevant for antenatal SBP, DBP, and HR. The antenatal BP association with BMI observed here, that a low BMI is associated with a low BP and *vice versa*, has been observed by other BP studies in pregnancy [61, 72, 69, 70]. Studies exploring the effect of BMI on the other vital signs in pregnancy have not been identified, nor have we found studies that treated BMI as a continuous variable and reported the results.

5.6 Discussion

In this chapter we have developed and presented reference ranges for antenatal and postnatal vital-sign trends using thousands of measurements collected during the 4P study. With the exception of antenatal SBP and DBP, obstetric vital-sign trends have not previously been modelled. The modelled trajectories supported previously suspected changes in antenatal SBP, DBP and HR, and revealed new information about Temp, SpO₂ and RR. Postnatal trajectories were particularly informative as there is little information available on physiology after delivery.

Previous attempts at modelling BP during pregnancy have mostly relied on multilevel models to derive reference ranges, and this was the method adopted in this chapter. The strength of a multilevel model lies primarily in its hierarchical nature. The model takes into account structures present in the data, such as repeated measures within a subject that are present in a longitudinally collected dataset. Variance is therefore appropriately modelled. The hierarchical structure allows prediction inference to be performed at both the population-level and subject-level. Trajectories from individual women can therefore be compared, so as to explore the difference between subgroups of women.

A primary weakness of the linear multilevel model is the assumption of normality. Between-subject and within-subject variances are assumed to be normally distributed, which is largely appropriate for vital-sign data except perhaps for oxygen saturation. SpO_2 is a one-sided distribution (bounded by 100% at the upper end), and is therefore likely to be skewed. Other statistical models, such as the L-M-S method, directly account for skewness and therefore may be more appropriate for such data, so should be explored further [142]. The subject-specific parameter distribution, $\mathbf{b}_i$, was also assumed to be Gaussian. This may mask underlying subgroups of trajectory shape. A more flexible distribution, such as a mixture of Gaussians, could be explored to identify possible subgroups.

Furthermore, the multilevel models considered here were parametric, that is the number of parameters were fixed, as opposed to growing with the dataset. While the standard polynomial appeared sufficient to model vital-sign trends, other models could be considered, such as kernel regression or Gaussian Process regression for which the number of parameters does not constrain the shape of the trend. Modelling in this way may be particularly advantageous for the 24 to 48 hours immediately after delivery where few data were available and the standard polynomial and fractional polynomial models diverged. The postnatal measure of time was in days after delivery for ease of interpretation, however measuring time in hours since delivery would be important to capture the transient changes in vital signs during this critical time. It would also be helpful if measurements were made more frequently during the first 48 hours, for example every 8 hours.

The addition of covariate information to the multilevel model was also explored. BMI was included as a continuous variable, which has not previously been done for obstetric applications. SBP and DBP in particular changed significantly with BMI. Differences in parity were investigated by fitting separate models to nulliparous and

multiparous women. The shape of the mean trajectories were similar between the two groups, suggesting that its inclusion as a parameter in the multilevel model would be appropriate. Other variables to consider include maternal age and smoking status; for postnatal trends, gestational age of delivery and type of delivery could also be included.

The method to decide which covariate information to include in a model remains controversial. The statistical literature suggests either a top-down or bottom-up approach, whereby parameters are removed or introduced gradually, performing strict statistical tests to determine the significance of each parameter [160]. This often leads to “over-fitting” when the model is only applicable to the available data. Alternatively, model parameters are chosen based on the quantity of data available, such that potentially important variables with little data may be omitted because chance variation cannot be ruled out as an alternative explanation for their predictive power. The Bayesian literature suggests all possibly relevant information should be included in a model. The possible loss of precision when including unimportant predictors is usually viewed as a relatively small price to pay for the general validity of predictions and inferences about estimands of interest [161]. However, both of these approaches must be mediated by the application of such a model. The obstetric vital-sign reference ranges developed in this chapter are ultimately for use in hospitals, where including too much information may inhibit their ease-of-use, yet excluding irrelevant information will make it a useless tool.

Associations between vital-sign trajectories can be explored through a multivariate model, which will be explored in Chapter 7. The next chapter discusses the translation of vital sign reference charts to a clinical tool using a track-and-trigger system.

Development of an Obstetric Early Warning Score

In Chapter 2 we introduced the concept of track-and-trigger systems for identifying and initiating a response to deteriorating physiology. Current track-and-trigger systems that have been adapted for maternity are based on limited knowledge of physiology in pregnancy. The antenatal and postnatal vital sign reference ranges presented in Chapter 5 were informed by thousands of vital sign measurements collected in the 4P study, and therefore provide an ideal model from which to create an evidence-based alerting system. In this chapter we review existing obstetric track-and-trigger systems, and propose a new obstetric alerting system that is specific to the stage of pregnancy.

6.1 Alerting systems in obstetrics

Since 2007, CEMD reports have emphasised the importance of routine monitoring of pregnant women and recommended the implementation of an EWS specific to obstetrics [6, 14, 36, 162, 163]. Standard adult alerting systems, such as the UK national NEWS, are not appropriate for the obstetric population since physiology adapts to the pregnant state [44, 151]. An example maternal EWS was included in the 2007 CEMD report, named the Modified Early Obstetric Warning Score (MEOWS) [36]. There has since been a national drive to implement MEOWS or a similar obstetric EWS in all hospitals

in the UK. By 2014, a survey of midwifery centres at NHS secondary care organisations reported that 99% (106/107) of respondents had implemented a maternal EWS [164]. One-third of these centres had directly implemented the recommended MEOWS score, while the remaining centres used a modified version or their own adapted EWS. This rate of maternal EWS implementation was an improvement from 2007, where a similar survey of 150 UK maternity units found that only 30 (19%) maternity units regularly used any form of EWS for obstetric patients, and just nine (6%) were using a modified EWS specific for obstetrics [151]. These surveys also highlighted the lack of consistency between obstetric EWS, such as disagreement about which physiological parameters to monitor, where vital sign thresholds should be set, and when an urgent clinical response should be triggered [164, 165].

The EWS is the most common form of track-and-trigger chart implemented in hospitals in the UK, and therefore the focus of the work in this chapter. EWS systems are usually presented as a paper chart, where each vital sign range is divided into a series of boxes available for selection by the recording clinician. A typical scoring system uses seven bands, e.g. {3, 2, 1, 0, 1, 2, and 3}, to describe the range of abnormally low values (3) to normal values (0) to abnormally high values (3) for each physiological parameter. This is the scoring system adopted by NEWS (see Section 2.2.2). In maternity, the scoring bands are often coloured white, yellow, orange, and red for the regions corresponding to a score of 0, 1, 2, and 3 respectively. The sum of scores is the EWS aggregate score, which is compared to the trigger threshold to determine whether a clinical response is initiated. Note we use the word *cut-off* to describe the values of each physiological parameter which separate the different scoring bands, and *trigger threshold* to describe the aggregate scoring criterion which determines whether a response should be triggered.

Some EWS systems in maternity determine whether a clinical response is triggered by counting the number of coloured boxes, rather than assigning numerical values. These

systems typically use five bands (instead of seven) for each physiological parameter, e.g. {red, yellow, white, yellow, and red}, to describe the range from abnormally low values (red) to normal values (white) to abnormally high values (red). A clinical response is usually triggered by two yellow parameters or one red parameter. Colour-coded systems were advocated by the 2007 CEMD report because they are simple to use and interpret [36].

Several names have been used to refer to an EWS adapted for obstetrics. Here we adopt the acronym ObsEWS (Obstetric Early Warning Score) to refer to an EWS adopted for use in maternity, as recently proposed by Quinn *et al.* [48].

6.1.1 Identification of obstetric EWS in the literature

To illustrate the inconsistencies in current ObsEWS, and thus emphasise the need for a standardised system, here we compare and contrast ObsEWS and other relevant systems presented in the literature. Four ObsEWS (labelled A-D) were identified in journal publications. Two other published track-and-trigger systems were also considered relevant: a single-parameter maternal warning system proposed for national use in the USA (E); and the first iteration of a response system which is initiated by vital sign triggers (F). The details of these publications are presented in Table 6.1. We have excluded ObsEWS that appear in textbooks (such as the PROMPT manual [64] and the MOET manual [166]) since these are examples of locally developed charts and no descriptions are given. However, for context, we include local ObsEWS currently in use at the three 4P study sites (labelled OXF, LON, and NCL for the Oxford, London, and Newcastle centres respectively). The cut-off values of these nine systems are presented graphically in Figure 6.1 using the yellow-orange-red colour scheme most common in ObsEWS systems.

The majority of the alerting systems presented here (5/9) use a colour-coded scoring system to trigger a clinical response (A, B, C, OXF, and NCL). They all follow the typical colour-coded format with two abnormal regions per vital sign - yellow (moder-

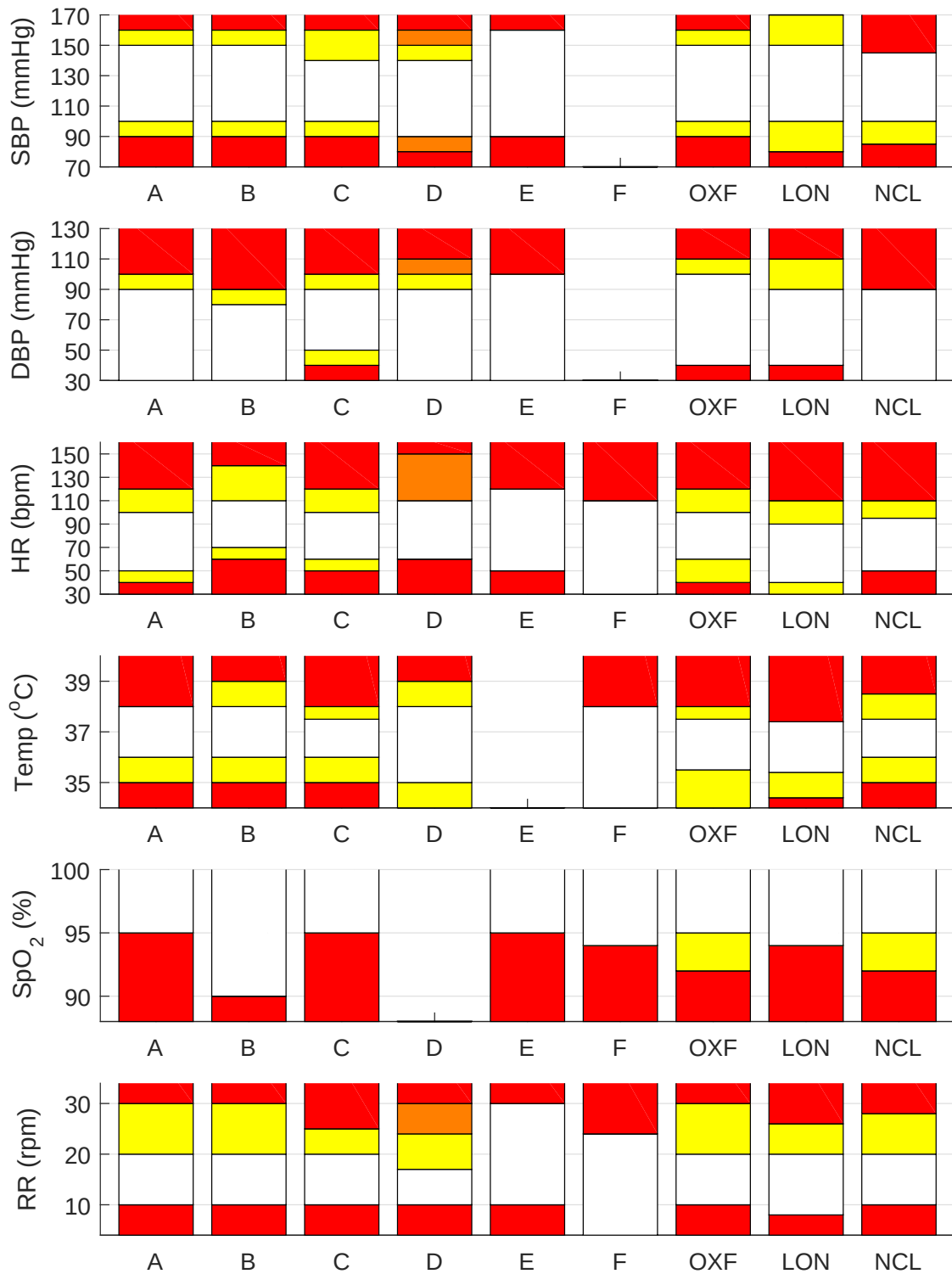


Figure 6.1: Vital sign cut-off values from ObsEWS and other relevant systems reported in the literature. The white region indicates a normal vital sign value, the yellow and orange regions indicate slightly abnormal values, and the red region represents most abnormal physiology. Letters A-F refer to the publications described in Table 6.1. OXF, LON, and NCL are the local ObsEWS used at the 4P study sites, Oxford, London, and Newcastle respectively.

	Year	Author	Country	ObsEWS name	Description	Design process	Scoring system
A	2007	CEMD [36]	UK	MEOWS	Locally developed	None given	CC
B	2009	Swanton <i>et al.</i> [151]	UK	Combined ObsEWS	Combination of nine local ObsEWS	Expert opinion	CC
C	2013	RCPI [152]	Ireland	I-MEWS	Irish national ObsEWS	Literature review and clinical consensus	CC
D	2013	Carle <i>et al.</i> [167]	UK	Statistical and clinical ObsEWS	Data-driven approach	Parameters informed from logisitic regression model of ICU data	AGG
E	2014	Mhrye <i>et al.</i> [168]	USA	MEWS	Proposed ObsEWS for United States	Literature review and clinical consensus	SP
F	2016	Hedriana <i>et al.</i> [19]	USA	MEWT	Early warning tool that includes treatment pathways	Literature review, clinical consensus and ICU data	SP*

CC: colour-coded system; AGG: aggregate scoring system; SP: single-parameter system

*F has a single cut-off value for each parameter like a single-parameter system, but two abnormal measurements are required to initiate a clinical response.

Table 6.1: *ObsEWS identified in the literature.*

ately abnormal) and red (very abnormal) - and alert when two yellow parameters or one red parameter are scored. Aggregate scoring systems, where a numerical “score” is calculated, are used by two systems (D, LON). LON again has two abnormal regions per parameter, assigning a score of either 0 (normal), 1 (moderately abnormal), or 2 (very abnormal) per vital sign. Graded clinical responses are triggered at an aggregate score of 2, 4, and 6. D is the only ObsEWS reported here to include a third abnormal region (coloured orange) for each parameter. Numeric scores for each vital sign therefore range from 0 (normal) to 3 (very abnormal), and clinical responses are initiated for aggregate scores of 4 and 6. This scoring system is most similar to NEWS. The remaining two systems (E, F) have a single cut-off (red) for each parameter, and trigger a response when any one (E) or two (F) parameters are in the red region. F additionally specifies the abnormal condition must be sustained for 30 minutes or more before an alert is triggered.

Six of the nine systems (A, B, C, OXF, LON, NCL) include all vital signs parameters in their alerting system: SBP, DBP, HR, Temp, SpO₂, and RR. One system (F) specifies a BP cut-off in the form of Mean Arterial Pressure (MAP) rather than SBP and DBP; another (D) measures the amount of oxygen required to maintain SpO₂ at 96% rather than the SpO₂ value. The final system (E) excludes temperature because it was not considered important by the system designers [168].

Although the vital signs included in these ObsEWS are largely agreed upon, additional parameters were often included in the alerting system. These may be subjective (such as neuroresponse or a pain score), or measurable (such as blood glucose or urine output). Neuroresponse was the most common, included in every system reviewed here.

6.1.2 Methods of design

An obvious reason for the inconsistencies between ObsEWS is the lack of data-driven approaches to development. Of the systems identified, the earliest three provide little information on the design process. The first nationally recommended MEOWS (A) reproduced the locally developed ObsEWS from Aberdeen Maternity Hospital [36]. We have not been able to identify an explanation of the design process. The ObsEWS presented by Swanton *et al.* (B) was a combination of the best elements of nine UK ObsEWS, decided by expert opinion [151]. No further details were given. The I-MEWS (C) has been implemented as the national ObsEWS in Ireland, yet the only justification provided for design decisions was a review of the international literature on ObsEWS [152]. Similarly, the local ObsEWS used in the Oxford, London, and Newcastle 4P sites (OXF, LON, and NCL) do not appear in journal publications, and thus no explanation of the design process has been found.

The first evidence-based ObsEWS was presented by Carle *et al.* in 2013 (D) [167]. Their aggregate-weighted ObsEWS was based on a statistical model, trained and validated using 13 years of ICU audit data from England, Wales, and Northern Ireland (the ICNARC dataset¹). Audit data contains the most abnormal physiological measurements collected within the first 24 hours of an ICU admission (for example, the lowest and highest heart rate recorded for a patient). This dataset therefore reflects only physiological extremes. All available physiological parameters (including HR, SBP, DBP, RR, and Temp) collected from direct obstetric admissions (4,440 women) were entered into a multiple logistic regression model where the outcome was patient mortality. The resulting coefficients (specific to one decimal place) were used to weight the corresponding physiological parameters in a statistical ObsEWS. However, the authors developed a

¹The Intensive Care National Audit and Research Centre (ICNARC) Case Mix Programme (CMP) database began in 1995 and is regularly updated with ICU data. The dataset analysed by Carle *et al.* was collected between 1995 and 2008.

second *clinical* ObsEWS using weightings and physiological parameters more familiar to clinicians, since they acknowledged the importance of an ObsEWS meeting the expectations of those who would be using it. The clinical ObsEWS therefore used whole number weightings between 0 (normal) and 3 (most abnormal) for the physiological parameters, and added five limits (high SBP, high DBP, high Temp, low HR, and low RR) which were not included in the logistic regression model.

Although the ObsEWS developed by Carle *et al.* are data-driven, the ICU nature of the dataset has been the source of much criticism. In a recent editorial, Smith and Prytherch suggest that an ObsEWS developed from such a dataset would only discriminate the risk of in-ICU death [169]. The score should, as a minimum, be tested statistically against data from the general obstetric inpatient population before implementation in the general obstetric ward is considered.

In 2014, Mhrye *et al.* [168] proposed a national maternal early warning system for the United States: the Maternal Early Warning Criteria (MEWC) (E). The MEWC is designed as a single-parameter risk assessment system whereby a response is triggered for any single abnormal parameter. A single-parameter system is thought to reduce false alarms that contribute to desensitisation, mistrust, and lack of caregiver response [20]. In addition, a single-parameter system does not require staff to document, calculate, and interpret a numerical score, which can be cumbersome when the score is usually normal. Against this, the “graded” approach of aggregate systems has been shown to provide a more robust measure of acute-illness severity for adult, non-obstetric patients [40].

Another relevant maternal warning system is the Maternal Early Warning “Triggers” (MEWT) (F) proposed by Hedriana *et al.* [19]. These are a set of single-threshold vital sign triggers where two or more abnormal vital signs are required to trigger a response. The authors also propose a response system of four clinical pathways which aim to identify the area of concern (e.g. sepsis, see Section 2.1.2) and make use of

existing policies and procedures in that area. This MEWT system is of note because the motivation of its authors in developing a maternal warning tool was to evaluate the use of *continuous* critical vital sign values for deciding on follow-up treatment. Continuous vital sign data from 50 obstetric ICU patients and 50 obstetric ward patients (the controls) informed their decision about the persistence requirement for vital sign triggers in their system. The physiological parameters to include in the early-warning system and cut-offs at which to trigger had been chosen previously by literature review and clinical consensus. They observed that, when the vital signs of control patients exceeded a threshold, the abnormal physiology did not persist. In contrast, the abnormal physiology of the ICU patients persisted for more than 30 minutes. A vital sign was therefore only considered abnormal when exceeding the threshold for more than 30 minutes, and two persistent abnormal vital signs were required to trigger a clinical response.

Of the nine maternal alerting systems presented here, only one used a statistical model to influence design (D), and two used data to inform or justify design choices (D, F). ObsEWS systems are largely based on expert opinion or opinions from literature, which are often not described or justified. There is an apparent lack of knowledge about vital sign trajectories expected in pregnancy, and therefore where to set cut-off values. None of the ObsEWS systems identified took into account the stage of pregnancy when determining vital sign thresholds. There is also a lack of agreement about which physiological parameters to include in an ObsEWS (decisions to include or exclude certain parameters were usually based on opinion) and on how to structure the trigger systems.

6.1.3 Methods of validation

The need for validation of ObsEWS systems is an ongoing challenge. Despite the rapid introduction of ObsEWS into the maternity services in the UK, few local charts have been validated and clinical audit has not taken place [165, 164, 170, 171]. In Section 2.2.1,

we introduced the commonly used metrics for validation of track-and-trigger systems. The ideal ObsEWS as a predictive tool would have a high sensitivity (all those with the defined outcome trigger a response) and high specificity (few unnecessary triggers). Five of the nine ObsEWS identified in Section 6.1.1 (Table 6.1) have been evaluated in a validation study (A, B, C, D, and F). Here we present and discuss the main findings from five validation studies identified in the literature, summarised in Table 6.2.

The earliest identified ObsEWS validation study was in 2012 by Singh *et al.* [20]. They validated MEOWS (A) using vital sign data from 676 obstetric inpatients at Barnet hospital, UK, collected prospectively over a period of two months. The outcome measure was maternal morbidity (pregnancy-related complications such as obstetric haemorrhage, pre-eclampsia, or suspected infection, defined before the study began), positive for 86 (13%) women. They found the MEOWS sensitivity was 89% and the specificity 79%, both higher than for a non-obstetric EWS. Women who triggered an ObsEWS response were more likely either to have or to develop morbidity than those who did not trigger. MEOWS had a poor positive predictive value (39%), probably due to the low prevalence of maternal morbidity in the studied population.

Two of the maternal warning systems identified in Section 6.1.1 included a validation study in the same publication that presented the design (D, F). These are the two systems whose design was data-driven. Carle *et al.* validated their statistical and clinical ObsEWS (D) against two other ObsEWS (A, B). They used a held-out section of their ICU audit dataset with which to perform the analysis, with mortality as the outcome. The statistical and clinical ObsEWS performed the best, both with an AUROC greater than 0.95. However, this validation study suffers from the previous criticism that only in-ICU death can be assessed by an ICU audit dataset. Data from obstetric inpatients on a general ward is required for a full validation.

Year	Author	Country	ObsEWS validated	Type of data	Type of study	Outcome measured	Main findings
2012	Singh <i>et al.</i> [20]	UK	A	Obstetric inpatient data, $n = 676$	Prospective	Morbidity (several complications)	A performed better than non-obstetric EWS
2013	Carle <i>et al.</i> [167]	UK	A, B, D	Obstetric ICU adult data, $n = 2200$	Retrospective	Mortality	D performed better than A and B, measured by AUROC
2015	Maguire <i>et al.</i> [170]	Ireland	C	Obstetric inpatients with bacteremia, $n = 81$	Prospective and retrospective	Vital sign recording rate	Dataset too small to suggest a change in outcome
2015	Edwards <i>et al.</i> [171]	USA	A, B, D and three others	Obstetric patients with chorioamnionitis, $n = 364$	Prospective	Morbidity (severe sepsis)	Large differences in performance between ObsEWS. Over-detection of severe sepsis
2016	Hedriana <i>et al.</i> [19]	USA	F	Obstetric ward and ICU data, $n = 50$	Retrospective	ICU admissions	Requirement of sustained abnormal physiology increases performance

Table 6.2: *Studies identified in the literature that validate ObsEWS systems.*

Hedriana *et al.* [19] also included validation results for their MEWT system (F) in the same publication that described its design. The same dataset was used for design and validation, with the outcome as ICU admission (50% rate). The MEWT had a sensitivity of 72%, a specificity of 96%, and a high positive predictive value of 95% [19]. The authors showed that the requirement of persistent abnormal physiology (>30 minutes) improved the specificity and positive predictive value when compared to the same thresholds triggered by an isolated abnormal value.

The most comprehensive ObsEWS validation study was performed by Edwards *et al.* [171]. They retrospectively assessed the predictive power of six ObsEWS (A, B, D, an example included in the PROMPT medical textbook [64], and two example NHS ObsEWS systems included in the Maternal Critical Care Working Group publication, labelled MC1 and MC2 [172]) to identify severe sepsis in a population with chorioamnionitis². Their dataset consisted of physiological measurements from 364 women with chorioamnionitis, five of which developed severe sepsis. Results showed large disparities between the ObsEWS. Three ObsEWS (A, B, PROMPT) had a high sensitivity (100%) yet low specificity (between 3.6 and 29%); the other three ObsEWS (D, MC1, MC2) had moderate sensitivity (40 to 60%) yet good specificity (84.4 to 96.9%). All ObsEWS had very poor positive-predictive values (from <2% to 15%), probably due to the very low number of positive cases in the dataset (5/364, 1.4%). AUROC ranged from 0.52 (A, PROMPT) to 0.68 (MC2). In general the ObsEWS tested here over detected severe sepsis; thus, the authors suggest that a more specific secondary test is required to identify true cases.

The final validation study concerned the Irish I-MEWS (C) score, which in April 2013 was implemented in all 19 maternity units in Ireland [170]. The effect of the I-MEWS system (C) was assessed by identifying whether recording of vital signs had improved

²A bacterial infection of the fetal amnion and chorion membranes that increases the risk of sepsis.

for women who had developed maternal bacteremia³ at a large maternity hospital in Dublin, Ireland (The Coombe Women and Infants University Hospital). Vital signs were extracted retrospectively from medical charts completed between 2009 and April 2013 (61 patients), and prospectively from the I-MEWS charts for the 12 months after its introduction (20 patients). The introduction of I-MEWS was associated with an improvement in the recording of vital signs, particularly RR. Despite covering a 5-year period, the number of people identified with maternal bacteremia was too small to suggest whether I-MEWS had caused a change in clinical outcome, so no typical validation metrics were presented.

In this section, we have examined five publications which present validation studies for one or more ObsEWS. All five employed a different outcome measure. The two that used a form of morbidity as the outcome (all maternal complications [20]; or severe-sepsis [171]) used well-defined criteria specified before the beginning of the study. It is thought that morbidity, rather than mortality, is a more appropriate outcome measure for ObsEWS since the number of maternal deaths is fortunately low in developed countries. All studies were affected by a low prevalence of the adverse event (positive outcomes) in their dataset. As a result, none of the five studies are able to come to any firm or robust conclusion about the performance of ObsEWS.

6.1.4 Future direction for an obstetric early warning score

Our findings demonstrate inconsistencies in existing ObsEWS systems. These need to be resolved with a standardised obstetric alerting system, similar to the standard UK national NEWS developed for the adult non-pregnant population [151]. A major barrier to development of such a system has been the lack of available evidence from which to inform cut-offs and trigger thresholds [48]. As discussed in Section 2.3.2, a systematic

³Bacterial growth in blood culture specimens from pregnant or postpartum women. If maternal bacteremia is associated with sepsis, severe sepsis, septic shock, or multiple organ dysfunction syndrome it may lead to significant morbidity and mortality.

review of the literature revealed few large studies collecting vital sign information from women experiencing a normal pregnancy. Furthermore, synthesis of BP and HR information from these studies showed large heterogeneity in the antenatal period. Limited information was available for the intrapartum or immediate postpartum periods, or for SpO₂, Temp, and RR.

Another barrier a standardised ObsEWS will need to overcome is their infrequent use in the postnatal period [6]. One centre, surveyed by Mackintosh *et al.* [42], reported just 22% postnatal usage even though use of a postnatal ObsEWS was a required guideline at that centre. Several studies have also identified the high absence of RR measurements in the postnatal period compared to other vital signs [42, 173]. The lack of adherence to an ObsEWS in the postnatal period is thought to be because nurses perceive women to be recovering from the natural event of childbirth, and therefore conclude an ObsEWS is to be neither necessary or relevant. While childbirth is a natural event, the female body is recovering from an altered physiological state, and normality cannot be presumed without measurement [6]. Many complications also occur in the few days after delivery, such as obstetric haemorrhage and sepsis, as previously discussed. The relevance of an ObsEWS could be addressed by developing different EWS systems for each stage of pregnancy.

A recent paper by Quinn *et al.* [48] calls for a standardised ObsEWS and presents clear guidelines for construction. The ObsEWS should be applied to all women presenting to acute care wards who are known to be pregnant or who are within 42 days (6 weeks) of delivery. The only six physiological parameters that should be included are: SBP, DBP, HR, Temp, SpO₂, and RR. This differs from the adult NEWS, where DBP is not recorded and level of consciousness is included [44]. The paper explains that a reduced level of consciousness is not an early warning sign but an indication of established critical illness, and DBP is an important indicator of gestational hypertension

and pre-eclampsia. These choices of vital signs are supported by other expert opinion [164, 165]. The scoring system of a national UK ObsEWS should align with that of the UK NEWS so that staff who primarily work with the non-pregnant population can easily adapt without confusion.

In the next section, we use vital sign information collected in the 4P study to develop an evidence-based ObsEWS specific to the antepartum and postpartum periods, incorporating the suggestions by Quinn *et al.* [48]. An antenatal ObsEWS (C-MEOWS) was developed by another member of our research group using preliminary analysis of 4P data [124]. Their work used data from fewer women than the dataset described in this thesis (439 compared to 620 women in this thesis), and only focused on the antenatal period. Our work builds on their preliminary analysis.

6.2 Centile-based early warning scores

Previous data-driven approaches to developing EWS systems, such as Carle *et al.* [167] and the ViEWS system which informed NEWS [45], have typically used a binary outcome (e.g. death/survival) to determine cut-off values for each physiological parameter. However, as observed by Tarassenko *et al.* [174], this approach may discriminate against patients with redeemable deterioration. They proposed an alternative approach where cut-off values are based on the extreme centiles of the statistical distribution of vital signs in the healthy population. This would alert clinicians to patients who exhibit abnormal physiology, irrespective of outcome, in line with the original objectives of EWS systems [175].

Tarassenko *et al.* [174] developed the centile-based early warning score (CEWS), informed by 64,000 hours of continuous vital sign data collected from the bedside monitors of 863 acutely ill hospitalised inpatients (non-pregnant adults). Their CEWS used seven scoring bands for each vital sign, numbered $\{3, 2, 1, 0, 1, 2, 3\}$ from most ab-

normally low values to most abnormally high values. The cut-off values for the first scoring band were set at the 10th and 90th population centiles; the cut-off values for the second scoring band were set at the 5th and 95th centiles; and the cut-off values for the highest scoring band were set at the 1st and 99th centiles. Thus, the most extreme 2% of measurements scored a 3; the next 3-10% scored a 2; and the next 11-20% scored a 1. The 80% of measurements with the most “normal” values (closest to the mean) did not contribute to the aggregate score. In this section we apply the same centile approach to the vital sign measurements in the 4P data to design a new ObsEWS for the maternity population (named C-ObsEWS).

6.2.1 Scoring bands of a centile-based ObsEWS

We use the outer centiles of the univariate standard polynomial multilevel models developed in Chapter 5 to set the cut-off values for each ObsEWS scoring band. Multilevel models use a hierarchical structure to model the dependence present in longitudinal data, and thus accurately describe the variance in the population. The multilevel model assumes a Gaussian distribution for each vital sign at each week of gestation and at each day after delivery. Centiles are straightforward to calculate from the Gaussian distribution. The standard polynomial form of time provided the best fit to both the antenatal and postnatal 4P vital sign data (Appendix C.7.3), using approximately 3,500 and 5,500 sets of measurements.

In line with the principles of the centile-based approach proposed by Tarassenko *et al.* [174] and the recommendations by Quinn *et al.* [48], we use seven scoring bands with a range of scores from 0 (normal) to 3 (most abnormal); the same scoring range as the UK national NEWS. We set the cut-off values for each vital sign at the 1st, 5th, 10th, 90th, 95th, and 99th centiles of their distribution for each week of gestation and each day after delivery. Figure 6.2 illustrates the selection of centiles for SBP at 25 weeks gestation. This process is performed for all vital signs except for SpO₂ which is

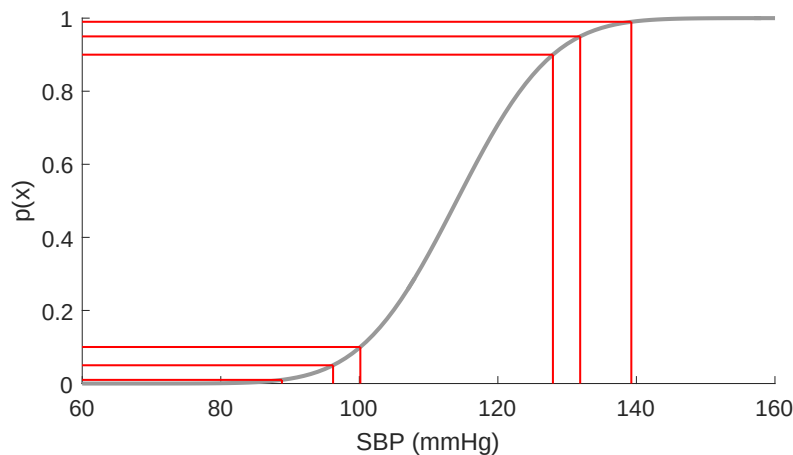


Figure 6.2: Identification of the 1st, 5th, 10th, 90th, 95th and 99th centiles (red lines) on the cumulative distribution of SBP at 25 weeks gestation. The cut-off values for the C-ObsEWS system are informed by these centiles.

a one-sided Gaussian distribution (only low SpO_2 values are considered abnormal; high SpO_2 values such as 99% and 100% are recorded for healthy women). We therefore select only the 10th, 5th, and 1st centiles to inform the cut-off values for SpO_2 .

As suggested by Quinn *et al.* [48], our C-ObsEWS uses six vital signs (SBP, DBP, HR, Temp, SpO_2 , and RR) and no other parameters. We propose a scoring system that varies with the stage of pregnancy, and therefore present different systems for the antenatal and postnatal periods. This addresses the concern that the current ObsEWS is not relevant for the postnatal population [42].

Our C-ObsEWS systems are presented in Figure 6.3 (SBP, DBP, and HR) and Figure 6.4 (Temp, SpO_2 , and RR). The system is coloured yellow-orange-red, corresponding to scores of 1, 2, and 3 respectively. A vital sign in the white region contributes a score of 0 to the overall ObsEWS score. The most apparent difference between the C-ObsEWS alerting cut-offs and other ObsEWS in the literature (see Figure 6.1), is the narrow distance between the lower and upper red cut-offs (1st centile and 99th centile).

To assess whether the C-ObsEWS is an appropriate early warning tool in obstetrics, we should ideally validate the system using an unseen dataset with labelled outcomes (such as ICU admission or development of a pregnancy-related complication). We cur-

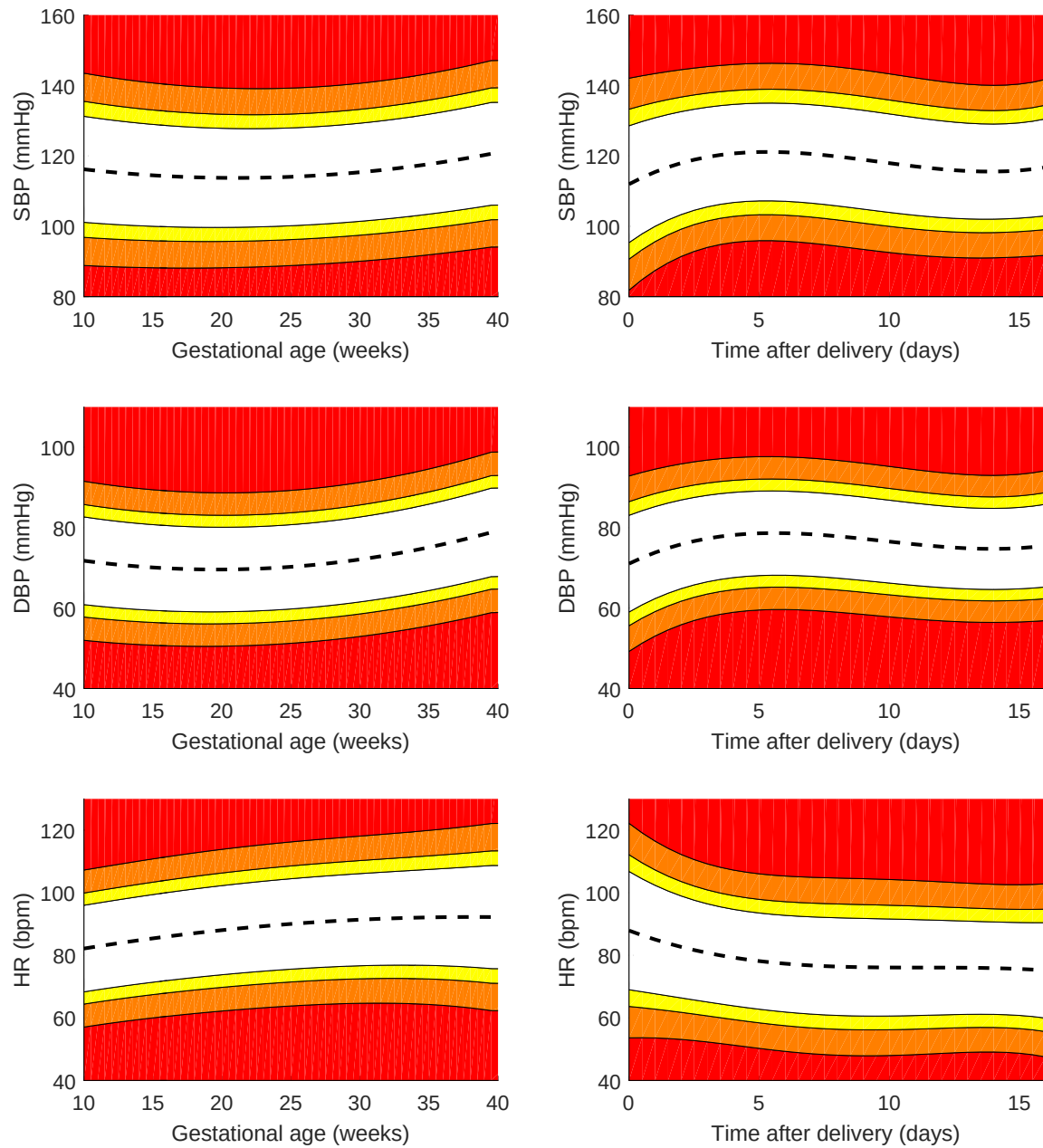


Figure 6.3: Proposed C-ObsEWS for SBP (top), DBP (middle) and HR (bottom) for the antenatal period (left column) and postnatal period (right column).

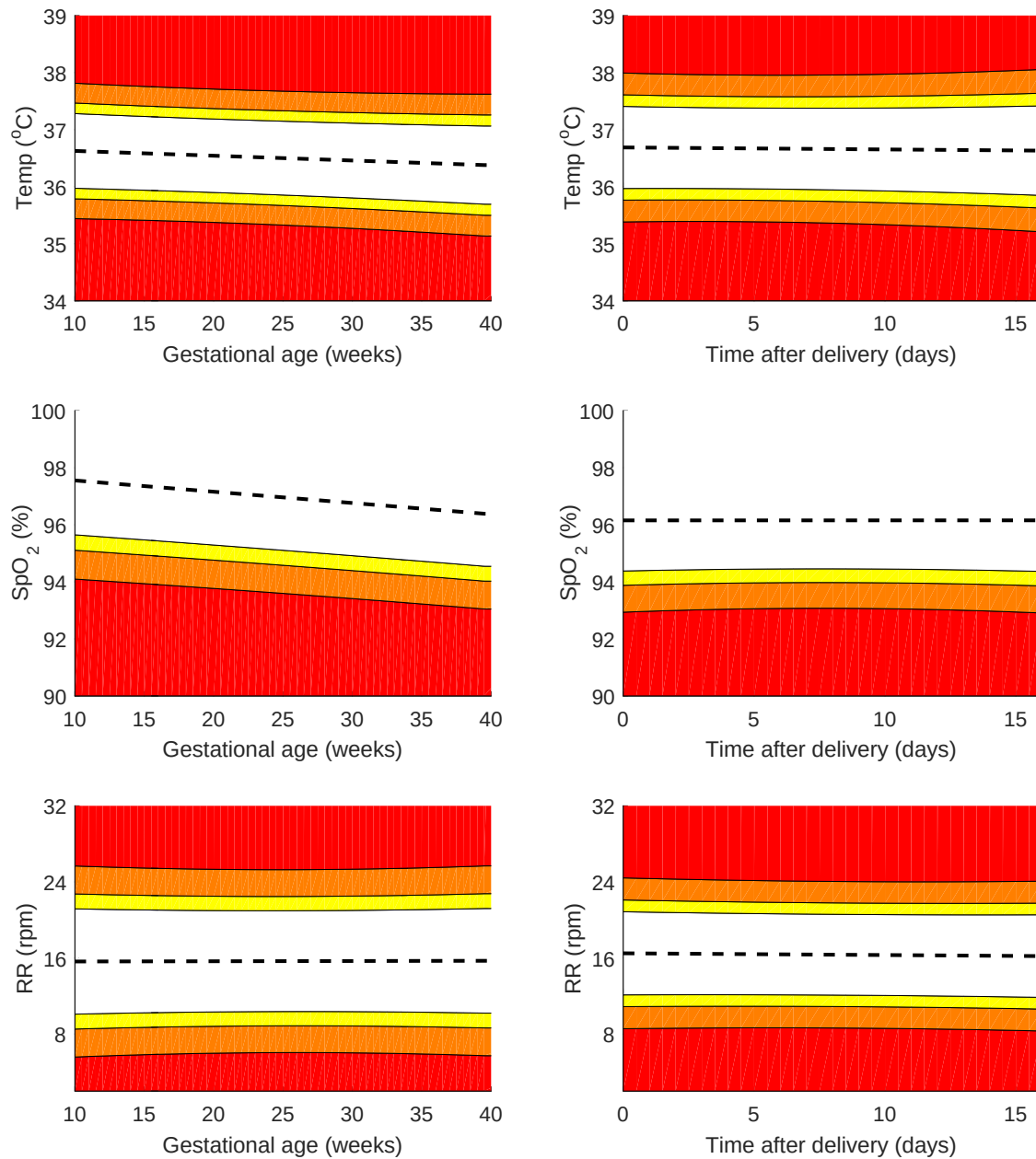


Figure 6.4: Proposed C-ObsEWS for Temp (top), SpO₂ (middle) and RR (bottom) for the antenatal period (left column) and postnatal period (right column).

rently do not have access to such a dataset. Instead, we discuss the potential impact of the C-ObsEWS system here by re-visiting the 4P dataset, and in Section 6.3 we examine two case studies of women who developed sepsis in the postnatal period.

6.2.2 Re-visiting the 4P data

The C-ObsEWS design was informed by 10,000 sets of vital sign measurements collected from 1 August 2012 to 15 July 2016 in the 4P study (see Table 5.3). Here, we investigate what proportion of those measurements would trigger an alert on the C-ObsEWS. We compare the results with the Oxford MEOWS chart (OXF in Figure 6.1), since this is the local ObsEWS system for the majority of women in the 4P study.

Figure 6.5 shows the proportion of 4P measurements which would be in the yellow, orange (C-ObsEWS only), or red regions of the two ObsEWS systems. For SBP, DBP, HR, and Temp, the C-ObsEWS is consistent in the proportion of alerts per vital sign, where approximately 10% of measurements cause a yellow alert, 8% cause an orange alert, and 2% cause a red alert. These numbers are what we expect since we chose the corresponding centiles to inform the placement of C-ObsEWS cut-off values. The proportion of alerts is lower for SpO₂ since we have used a one-sided distribution. RR has a peak of orange readings, probably due to the quantised nature of RR recordings during Phase 1 of the 4P study⁴. The cut-off value for the RR yellow region is around 21 rpm and the next cut-off is at 22 rpm; thus, only RR measurements performed for 60 seconds (and therefore not restricted to a multiple of 2 or 4) could be in the yellow region.

The Oxford MEOWS alerts the most frequently for HR and SpO₂, and least frequently for DBP. The number of yellow HR and SpO₂ triggers are similar to the number of C-ObsEWS yellow triggers. However, red alerts for Oxford MEOWS are rare. Less

⁴Phase 1 respiratory recordings were either 15 or 30 seconds in duration and therefore the respiratory rates are multiples of 4 and 2 respectively. See Table 3.2 in Section 3.2.

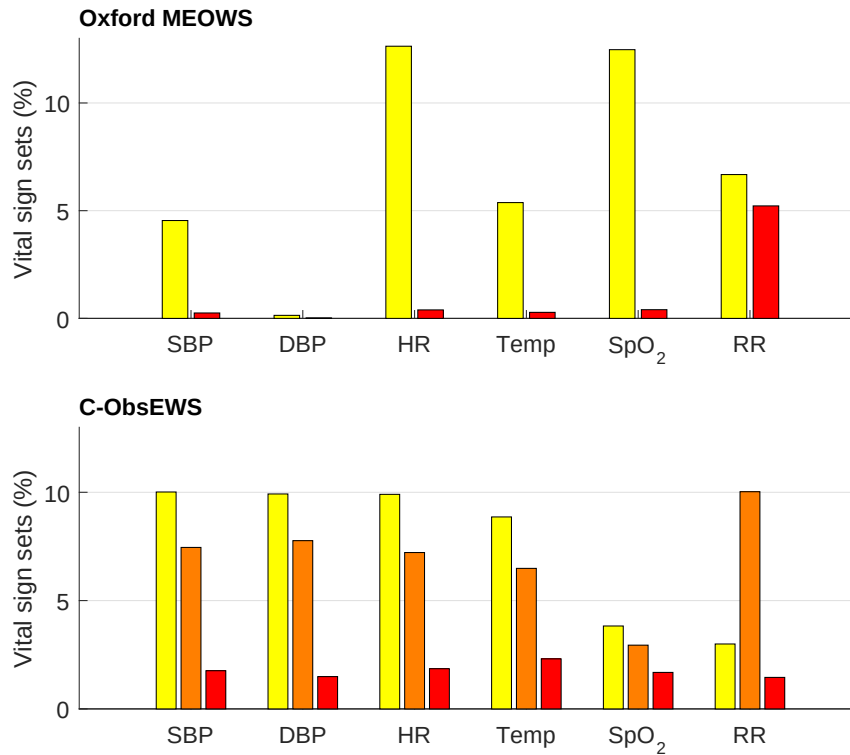


Figure 6.5: Percentage of 4P vital sign measurements that fall in the yellow, orange, or red regions for the Oxford MEOVS chart (top) and proposed C-ObsEWS system (bottom).

than 0.4% of SBP, DBP, HR, Temp, and SpO₂ measurements are in the red region. DBP is even more surprising with just 14 (0.1%) yellow alerts, and 2 (< 0.1%) red alerts. A high SBP and DBP is exhibited by 7.4% of women in the 4P study developed gestational hypertension or pre-eclampsia (see Section 3.3.2). As hypertension is indicated by high BP⁵, we would expect at least this number of yellow or red triggers for SBP and DBP combined. The Oxford MEOVS chart is thus not very “early” in its warning. The lack of DBP triggers also emphasises just how little is known about DBP in pregnancy [19].

Action from an EWS is initiated when the aggregate score reaches the trigger threshold. The Oxford MEOVS chart is a colour-coded system where two yellow parameters or one red parameter would initiate a clinical response. This translates to a numerical score of 1 and 2 to the yellow and red regions respectively, where a response would be

⁵Gestational hypertension and pre-eclampsia are defined in Section 2.1.2

initiated at an aggregate score of 2. As previously mentioned, C-ObsEWS scores for each vital sign range from 0 (white region) to 3 (red region). An aggregate threshold for a response has not been proposed, so we adopt the first (lowest) NEWS trigger threshold of 5, since NEWS also use a range of 0 to 3 for each parameter.

Figure 6.6 presents the aggregate scores for each set of vital sign measurements (top graph) and corresponding number of women (bottom graph) for C-ObsEWS and Oxford MEOWS. Aggregate scores range from 0 to 4 for Oxford MEOWS, and from 0 to 12 for C-ObsEWS. A greater number of vital sign sets cause a trigger for Oxford MEOWS (2 or above, 8.0%) than for C-ObsEWS (5 or above, 5.8%). This corresponds to 369 (59.5%) and 273 (44.0%) women for Oxford MEOWS and C-ObsEWS respectively. However, the greater range in aggregate scores of the C-ObsEWS scoring system would allow more stratification of physiological deterioration.

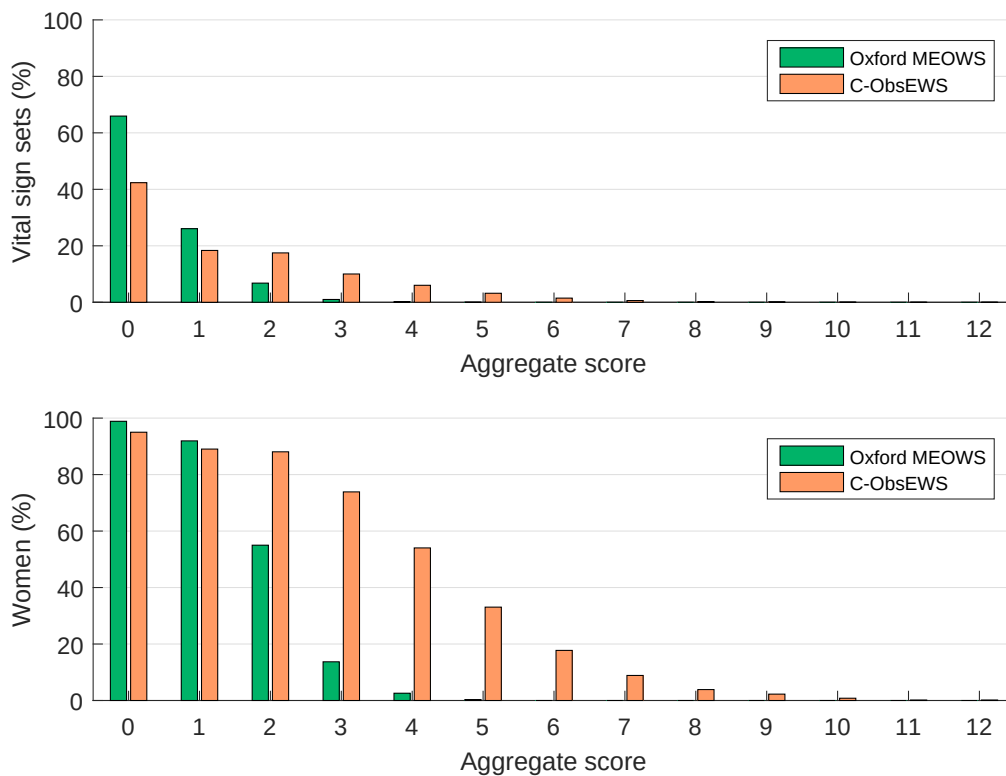


Figure 6.6: Total EWSscore for 4P vital sign measurements using the Oxford MEOWS chart (green) and proposed C-ObsEWS system (orange).

In this section we have compared the number of vital sign alerts from the C-ObsEWS with the Oxford MEOWS chart, thought to be representative of existing ObsEWS in use today. We observed that the C-ObsEWS system generated more alerts for individual vital signs than Oxford MEOWS, and that the aggregate C-ObsEWS scores covered a greater range than that of Oxford MEOWS. The next step in validation would be to match alerts with a clinical outcome. The 4P dataset contains few women with a poor clinical outcome, such as development of a pregnancy-related complication, and thus further validation can not be performed. As we have designed C-ObsEWS to alert for physiological deterioration, we would expect to see an increase in sensitivity but a reduction in positive predictive value when compared to existing ObsEWS.

6.3 Women who developed sepsis

Sepsis is one of the leading causes of maternal death in the UK, and thus has received much recent attention by the medical community. The 2011 and 2014 CEMD reports have emphasised a “Think Sepsis” slogan which encourages clinicians to think of sepsis at an early stage when presented with an unwell pregnant or recently pregnant woman [6, 162]. One of the five validation studies presented in Section 6.1.3 was concerned with the predictive power of the ObsEWS for severe sepsis [171]. Furthermore, a new component of the Confidential Enquiry program, which investigates a selected maternal morbidity (the Confidential Enquiries into Maternal Morbidity, CEMM), began in 2014 with the focus on maternal sepsis.

Sepsis is defined as an infection with the presence of two or more SIRS criteria⁶. The SIRS triggers are a high or low Temp ($> 38^{\circ}\text{C}$ or $< 36^{\circ}\text{C}$), high HR (> 90 bpm), high RR (> 20 rpm), and a high white-cell count. However, the SIRS criteria are defined for the non-pregnant population, and are often met within the normal healthy pregnant state. A systematic review by Bauer *et al.* [176] compared expected pregnancy physiology

⁶Sepsis was introduced in Section 2.1.2.

reported in the literature with the SIRS criteria. They illustrated how, during some stages of pregnancy, between 20% and 30% of the pregnant population would trigger one or more SIRS criteria. The C-ObsEWS scoring system developed in this chapter (see Section 6.2.1) confirms the overlap of SIRS criteria with expected pregnancy physiology. For example, a HR of 90 bpm is considered normal (white region, score of 0) for the duration of the antenatal period and the first five days after delivery (see Figure 6.3).

While specific sepsis screening tools have been introduced to structure bedside clinical assessment in the non-pregnant population, tools for sepsis in maternity have been neglected. Furthermore, the literature lacks evidence-based criteria for SIRS that specifically adjusts for the physiological changes which occur during pregnancy [177].

The CEMM report on maternal sepsis highlighted delays in diagnosis and incomplete vital sign assessment as two reasons for the development of severe sepsis and consequent mortality [6]. In particular, the reviews found that the absence of vital sign readings was most obvious in the postnatal period, and RR was the most frequent physiological measure omitted. Incomplete sets of vital sign observations are considered a lost opportunity for earlier diagnosis. Here we investigate whether, in addition to complete vital sign recordings, the design of ObsEWS itself could improve earlier recognition.

Case studies

A random sample of twelve women that developed postnatal sepsis were identified from those who had delivered at the John Radcliffe Hospital, Oxford, UK, between October 2014 and April 2015. None of these women were enrolled in the 4P study. We transcribed their vital-sign observations that had been recorded on paper charts into an electronic format. Time since delivery was measured in calendar days. Missing time stamps were interpolated between known time stamps.

Of these twelve women, two remained in hospital for a few days after delivery (in both cases due to a forceps delivery and obstetric haemorrhage) with regular vital sign

monitoring. They therefore had an adequate number of vital sign measurements from which to interpret an EWS score trajectory. These women's hospital stays are described below.

Woman A: Woman A experienced a normal pregnancy. She had a forceps delivery six days before her EDD, with an epidural. She lost 800mL of blood. She remained in hospital after delivery due to neck and back pain, and was given pain relief medication. Five days after delivery she was started on antibiotics to cover possible endometritis⁷. Two days later she was discharged from hospital. On day 11 after delivery (four days later) she was re-admitted to hospital with suspected endometritis. She was immediately started on intravenous⁸ fluids and antibiotics. She was discharged the next day.

Woman B: Woman B experienced a normal pregnancy, except for two hospital admissions due to reduced fetal movements. She had a forceps delivery seven days after her EDD and lost 700mL of blood. After delivery she remained in hospital to recover from the forceps delivery. She was discharged four days later. Ten days after delivery, she was re-admitted to hospital with suspected endometritis. She was given oral antibiotics. No other details were recorded. On day 13 after delivery she was discharged from hospital in the morning. Later that day (10pm) she was re-admitted with severe abdominal pain and nausea. She was discharged on day 18 after delivery, re-admitted on day 22 with persistent endometritis, then finally recovered and was discharged from hospital 24 days after delivery.

Both of these women had suspected endometritis before being discharged from hospital, and both were readmitted with confirmed sepsis (SIRS criteria with infection confirmed by blood culture, in these cases from endometritis). Woman B was re-admitted three times before finally recovering more than three weeks after delivery. Here we investigate whether an evidence-based ObsEWS, such as C-ObsEWS proposed in Section

⁷Inflammation of the lining of the uterus, usually due to infection. Endometritis is a common cause of maternal sepsis.

⁸Fluids or drugs that are delivered intravenously (IV), are fed directly into a vein.

6.2.1, could have better identified physiological deterioration and thus prevented the return to hospital or delayed discharge.

Postnatal vital sign observations for Woman A and Woman B are presented in Figures 6.7 and 6.8 respectively. Observations are plotted on both C-ObsEWS and the local Oxford MEOWS chart for comparison. The aggregate EWS score for each set of observations for each system is also presented. In the absence of data beyond 16 days after delivery from which to construct C-ObsEWS (see Section 3.3.3), the yellow, orange, and red C-ObsEWS cut-offs beyond day 16 were kept constant.

For Woman A, the EWS score for the first five days in hospital are vastly different between Oxford MEOWS and C-ObsEWS (Figure 6.7). The former records some fluctuation in the first two days after delivery, and then mostly remains at a score of 0 until discharge. In contrast, the C-ObsEWS score is oscillating around 5, which would trigger a clinical response. She appears to be getting better on days 4 and 5 after delivery, but on day 6 her BP drops (100/70 mmHg), her HR increases (105 bpm) and her RR increases (28 rpm). The woman is discharged at day 7 with a C-ObsEWS score of 5. The main cause for the difference between the two EWS scores is the DBP measurement. This woman has a low BP, and thus is continuously in the red region for DBP, and SBP is often also below the red cut-off. During pregnancy she had a relatively low SBP and DBP, primarily in the yellow region of the antenatal C-ObsEWS system⁹. She had an obstetric haemorrhage which would account for hypotension immediately after delivery, but persistent postnatal hypotension in the red regions is unusually low for her. As hypotension is an indicator of severe sepsis and septic shock, this is an important warning sign. When she is re-admitted at day 11 after delivery, her BP is initially normal (112/80mmHg, white region) but then drops as her HR increases and her temperature drops. She is discharged the next day, with a C-ObsEWS score of 4, and fortunately

⁹BP measurements for Woman A during the antenatal period were (as SBP/DBP in mmHg): 85/50 (10 weeks gestation), 89/64 (13 weeks), 100/58 (16 weeks), 110/70 (19 weeks), 97/67 (26 weeks), 100/60 (29 weeks), 99/60 (35 weeks), 100/60 (37 weeks), 106/60 (39 weeks).

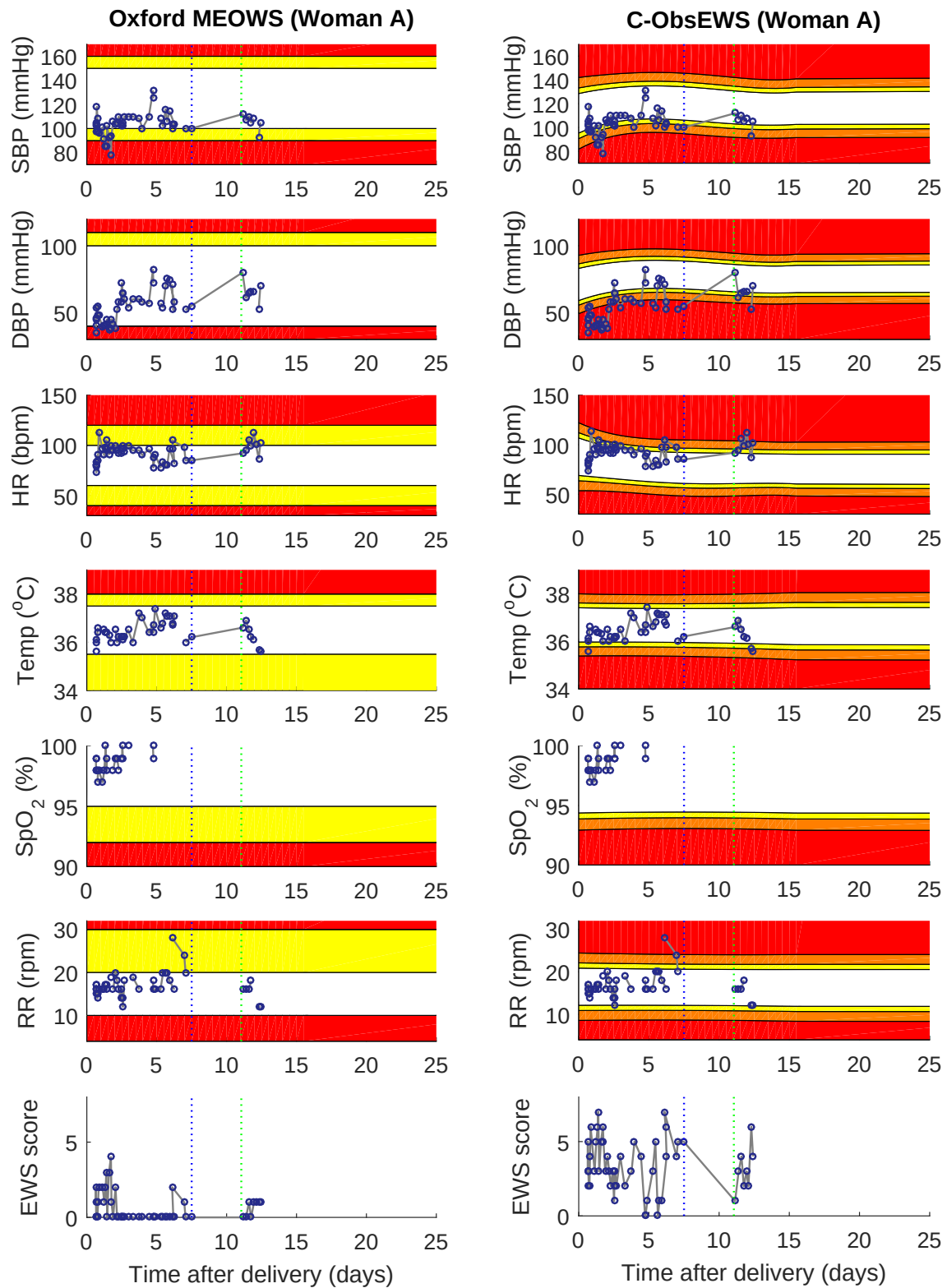


Figure 6.7: Example Woman A. Vital signs recorded for a woman who developed sepsis after delivery. Vertical dashed lines represent discharge (blue) and readmission (green) into hospital.

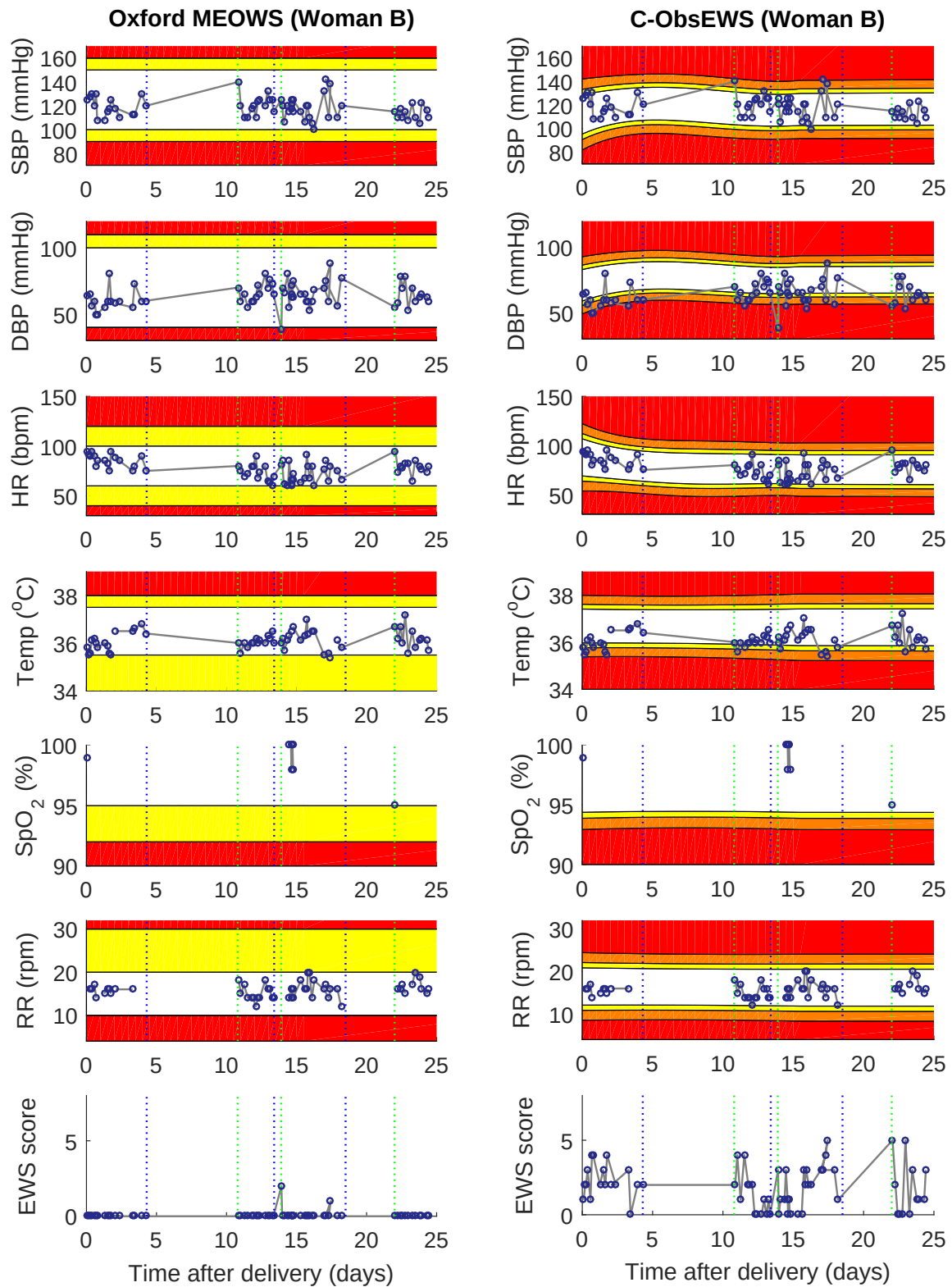


Figure 6.8: Example Woman B. Vital signs recorded for a woman who developed sepsis after delivery. Vertical dashed lines represent discharge (blue) and readmission (green) into hospital.

recovers at home.

The same pattern in EWS scores is observed for Woman B. The aggregate score from Oxford MEOWS remains at 0 with two deviations at day 13 and 17. In contrast, the aggregate C-ObsEWS score is rarely 0. Immediately after delivery the C-ObsEWS scores between 2 and 3, which would not be cause for alarm. In her subsequent re-admissions she would have triggered a clinical response multiple times, which may have made a difference to her care. The difference between EWS scores is again caused by low DBP (around 50mmHg, SBP is in the normal range), and low temperature (lowest of 35.4°C). These are both warning signs of sepsis.

The case studies described here illustrate the fact that C-ObsEWS would create a different clinical response to existing obstetric EWS systems such as the Oxford MEOWS. The cut-offs for DBP caused the greatest difference in the aggregate score. This confirms the suspicion of Singh *et al.* [20] that the cut-off values for low BP in existing ObsEWS charts were incorrectly set. In these examples we only had access to vital sign measurements recorded while in hospital. If home measurements had been available, it may have been possible to evaluate the postnatal changes in physiology more accurately.

6.4 Discussion

In this chapter we have presented the need for a standardised evidence-based ObsEWS, and proposed the C-ObsEWS system to meet that need.

The C-ObsEWS has three distinct advantages over existing ObsEWS systems. Firstly, the cut-offs for each vital sign are informed by thousands of vital sign measurements acquired during a normal pregnancy. The distance between the most abnormal cut-offs (red, score of 3) are narrower than other ObsEWS, which have initially been informed by clinical consensus, literature review, or ICU data. Secondly, the C-ObsEWS is specific to the stage of gestation. We propose separate systems for the antenatal and postnatal

periods, and within each system the cut-off values for each vital sign vary with time. This addresses concerns that ObsEWS systems are not relevant to the stage of pregnancy [42]. Such a system may be difficult to implement in paper charts, but would be appropriate in an electronic system for which cut-off values corresponding to each stage of pregnancy could be stored or computed. Thirdly, we demonstrated that the C-ObsEWS produces a wider range of aggregate scores than existing charts. This allows greater stratification of patients with physiological deterioration.

An outstanding requirement is an obstetric dataset with clear outcome labels with which to validate the C-ObsEWS. We have qualitatively demonstrated the sensitivity of the C-ObsEWS by investigating two case studies of women who developed sepsis in the postnatal period. In both cases, the sensitivity of the C-ObsEWS score could have prompted improved clinical care.

The scoring range for each parameter in C-ObsEWS was chosen to be the same as the UK national NEWS score. This was suggested by Quinn *et al.* [48] so that medical staff could transition between obstetric wards and other hospital wards without confusion. The linear weighting of vital signs is familiar for clinicians and thus easy for them to interpret. However, more complex scoring combinations may better describe physiological deterioration. The main obstacle to implementation of such systems has been the paper-based nature of typical EWS systems, which often require hospital staff to calculate aggregate scores in stressful situations. There are now several examples of electronic vital sign recording systems successfully deployed within the NHS [178]. The SEND system in Oxford allows patients' vital signs to be entered and reviewed via a bedside tablet, and is now used for routine observations on adult wards in all acute NHS hospitals in Oxfordshire [179].

A complex scoring system could also take into account interaction between vital

signs to better describe physiological deterioration. For example, hypovolemic¹⁰ shock is indicated by the combination of an increased HR and a lowered BP. During postpartum haemorrhage, blood loss is compensated for by an increase in HR until no further increase is possible, after which SBP begins to fall [180]. Thus a young, healthy woman could lose up to 30% of her blood volume before either HR or SBP are outside of their normal clinical range. The use of the expected values for the relevant stage of gestation may make HR and SBP more sensitive indicators, as discussed in Section 5.3.1. However, the Shock Index (SI), which is defined as HR divided by SBP, has been shown to be better than HR and SBP at predicting blood loss and subsequent maternal morbidity [181, 182]. A normal SI for the obstetric population is between 0.7 and 0.9 [180]. An SI using values which take into account gestational age may be even more sensitive. In Chapter 7 we investigate the interaction between vital signs beyond bivariate correlations and incorporate it into a complex scoring system for the C-ObsEWS.

Other future improvements to C-ObsEWS include the use of continuous monitoring and individual patient modelling. Continuous monitoring was used by Hedriana *et al.* [19] in their MEWT tool (F) to determine when vital signs were persistently abnormal, rather than an isolated incident. The main advantage of continuous monitoring should be identification of patient deterioration in real-time. Continuous vital sign data has been used in high-risk patients on adult wards, where the wealth of data enables the use of more complex mathematical methods such as Gaussian Processes [183]. These models can be specific to an individual, determining their personal vital sign ranges, rather than using population-determined thresholds. Patient-specific antenatal data could be used to optimise postnatal vital sign cutoffs.

¹⁰Hypovolemia is a decrease in blood volume

Multivariate systems

In the previous chapters we have focused on *individual* vital sign trajectories during pregnancy and after delivery. Although aggregate track-and-trigger systems produce a score that combines the abnormality of multiple parameters, these scores are typically linearly weighted and consider parameters independently. In other words, an abnormal value for one vital sign is treated as equally indicative of the patient state as an abnormal value for another vital sign, and the understanding of an expected vital sign range is not influenced by the values of other vital signs. In reality, the human body is a complex system and changes in physiology do not occur in isolation. In this chapter we explore the interaction between vital signs measured during pregnancy by considering a multivariate multilevel model. This is an extension of the multilevel model developed in Chapter 5. We also compare the multivariate results with the C-ObsEWS score developed in the previous chapter.

7.1 The contribution of multivariate models

A multivariate model is a joint model of all observed variables (e.g. all vital signs) that learns information about how these variables may be correlated. A joint modelling approach to the 4P dataset shall help to establish how the association between vital signs changes during pregnancy, and which combinations of vital signs are most abnormal.

There are a number of approaches to modelling multivariate longitudinal data, re-

viewed by Verbeke *et al.* [184]. For our analysis we select the family of multilevel models, since they have already been shown to be appropriate for modelling individual vital sign trends in the 4P data (Chapter 5). Multilevel models can be extended to multiple dimensions by incorporating additional population and subject-specific parameters [185]. The association between outputs is described in the population level covariance matrix. Multilevel models are versatile since they can include both continuous and discrete data, and allow for unbalanced data. Furthermore, the parameters of multilevel models are interpreted in terms of the observed values rather than the transformed parameter space characteristic of other multivariate modelling systems [184].

There have been no previous attempts at applying multivariate multilevel models to longitudinal vital sign data collected in pregnancy or in the postnatal period. In this chapter we focus on the antenatal period, as vital sign values vary more during pregnancy than after delivery.

7.2 Multivariate multilevel models

7.2.1 Dataset

We exclude measurements of SpO₂ from the multivariate analysis. SpO₂ is a one-sided distribution, bounded by 100% on the upper end. In the univariate analysis of SpO₂ in Chapter 5, the mean value for SpO₂ decreased from 98% to 96% during the course of pregnancy. SpO₂ values above the mean value are healthy, thus SpO₂ is not appropriate to include as one of the dimensions in a multivariate Gaussian distribution. For the analysis in this chapter, we consider the following set of vital signs: SBP, DBP, HR, Temp, and RR.

We use vital sign data collected from the 620 *healthy* women in the 4P study. We only use antenatal data, and require vital signs sets to have at least three of the five vital sign measurements recorded. If only one or two measurements are missing, they are replaced

by the population mean value of that vital sign over the whole of the pregnancy. 3,673 vital sign sets were recorded from 4P women in the antenatal period. 684 (18.6%) sets had one vital sign value missing (most commonly Temp), 108 (2.9%) sets had two vital sign values missing, and 15 (0.4%) sets had more than three vital sign values missing, so these 15 sets were excluded from analysis. Hence 3,658 vital sign sets from 620 women were used for multivariate analysis, of which 900 vital sign values were replaced by the corresponding mean (for 792 vital sign sets, 21.7%).

7.2.2 Description of the multivariate model

The multivariate multilevel model has the same hierarchical structure as the univariate multilevel model introduced in Chapter 5 (Section 5.2.2). For longitudinal data, as in the 4P study, repeated measurement occasions (level 1, subscript j) are grouped by subject (level 2, subscript i). For the multivariate model, each observation is now a *set* of vital sign measurements, i.e. $\mathbf{y}_{ij}$, the j^{th} observation on the i^{th} subject measured at time x_{ij} , is a vector of dimension 5 (SBP, DBP, HR, Temp, and RR).

The population parameters, $\boldsymbol{\beta}$, include one or more parameters per output variable to describe the shape of the the multivariate population trajectories. Similarly, the subject-specific parameters, $\mathbf{b}_i$, include subsets of parameters for each output variable to capture variation in the individual subjects that is not explained by the population model. The X_{ij} and Z_{ij} matrices contain the population and subject-specific covariates respectively, usually a function of time (e.g. a polynomial). For both the X_{ij} and Z_{ij} matrices, we use the superscript (v) to refer to the subset of *columns* that are associated with vital sign v . The subscript i selects the *rows* that refer to the measurement occasions for subject i . The multivariate model for vital sign v is therefore described by:

$$y_{ij}^{(v)} = X_{ij}^{(v)} \boldsymbol{\beta}^{(v)} + Z_{ij}^{(v)} \mathbf{b}_i^{(v)} + \epsilon_{ij}^{(v)} \quad (7.1)$$

Variance is described by the within-subject (level 1) covariance matrix $\mathbf{R}$, and the between-subject (level 2) covariance matrix, $\mathbf{D}$. As for the univariate model, we assume the residual vectors, $\boldsymbol{\epsilon}_{ij}$, are uncorrelated and normally distributed with a constant within-subject variance, σ^2 . The within-subject covariance matrix for output v is therefore $\mathbf{R}^{(v)} = \sigma^{2(v)}\mathbf{I}$. We assign an unstructured matrix to the between-subject covariance matrix, $\mathbf{D}$, which describes the variance of the subject-specific parameters, $\mathbf{b}_i$.

The covariance matrix of the population model is a combination of the within-subject and between-subject variances. For each pair of output variables, e.g. A and B , the covariance is calculated using Equation 7.2, which describes how each vital sign output varies with respect to the other vital signs.

$$\text{Cov}(A, B) = \mathbf{Z}^{(A)}\mathbf{D}^{(A,B)}\mathbf{Z}^{\mathbf{T}(B)} + \mathbf{R}^{(A,B)} \quad (7.2)$$

7.2.3 Functional form of time

In Chapter 5 we considered three functional forms of time: (a) standard polynomials, (b) fractional polynomials, and (c) cubic splines (see Section 5.2.3). We chose standard polynomials to model the antenatal and postnatal univariate vital sign trends (see Appendix C.7.3). The standard polynomial model is described by Equation 7.3 where P is the polynomial order that describes the population trend, and Q is the polynomial order that describes the subject-specific trajectories.

$$y_{ij} = \sum_{p=0}^P \beta_p x_{ij}^p + \sum_{q=0}^Q b_{iq} x_{ij}^q \quad (7.3)$$

Since the standard polynomial functional form has already been shown to be most appropriate to model vital sign trajectories in the 4P data, we select the same standard polynomial form for the multivariate multilevel model. The polynomial orders selected for the antenatal univariate vital sign models are shown again in Table 7.1 (excluding

	P	Q
<i>Antenatal</i>		
SBP	2	1
DBP	2	1
HR	2	2
Temp	1	1
RR	1	1
Total parameters	13	11

Table 7.1: Standard polynomial model orders selected for the antenatal univariate multilevel models. P is the polynomial order of the population model, Q is the polynomial order for the subject-specific model. The total number of parameters includes cases where $P = 0$ and $Q = 0$.

SpO₂ since it is not included in the multivariate model).

Including powers of zero, there are 13 population parameters ($\boldsymbol{\beta} = \beta_0, \beta_1, \dots, \beta_{12}$) which describe the multivariate population trajectory, and 11 subject-specific parameters per subject ($\mathbf{b}_i = b_{i0}, b_{i1}, \dots, b_{i10}$) which describe the multivariate subject-specific trajectories. The between-subject covariance matrix, $\mathbf{D}$, is therefore an 11×11 matrix (since $\mathbf{b}_i \sim \mathcal{N}(0, \mathbf{D})$). The equations for each vital sign are given below:

$$y_{ij}^{(SBP)} = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2 + b_{i0} + b_{i1} x_{ij} \quad (7.4)$$

$$y_{ij}^{(DBP)} = \beta_3 + \beta_4 x_{ij} + \beta_5 x_{ij}^2 + b_{i2} + b_{i3} x_{ij} \quad (7.5)$$

$$y_{ij}^{(HR)} = \beta_6 + \beta_7 x_{ij} + \beta_8 x_{ij}^2 + b_{i4} + b_{i5} x_{ij} + b_{i6} x_{ij}^2 \quad (7.6)$$

$$y_{ij}^{(Temp)} = \beta_9 + \beta_{10} x_{ij} + b_{i7} + b_{i8} x_{ij} \quad (7.7)$$

$$y_{ij}^{(RR)} = \beta_{11} + \beta_{12} x_{ij} + b_{i9} + b_{i10} x_{ij} \quad (7.8)$$

The multivariate multilevel model was solved using the same Bayesian inference methodology as for the univariate multilevel models (see Section 5.2.4). Prediction was performed from 10 to 40 weeks of gestation which, for a 5-dimensional (5-D) multivariate model, yields a 5-D Gaussian distribution at every time point.

7.3 Vital sign correlations

The mean population trajectory for each dimension of the multivariate model is similar to the mean population trajectory from the univariate model of the same vital sign. The contribution of the multivariate model is in the correlations between vital signs. The correlation coefficient, ρ , is a scaling of the covariance at the population level (see Equation 7.2) between -1 (strong negative correlation) and 1 (strong positive correlation). A correlation of 0 indicates no relationship between the variables. For variables A and B , Pearson's correlation coefficient is calculated by:

$$\rho_{AB} = \text{Corr}(A, B) = \frac{\text{Cov}(A, B)}{\sqrt{\text{Cov}(A, A)\text{Cov}(B, B)}} \quad (7.9)$$

We visualise the correlation between vital signs in Figure 7.1 by pair-wise (2-D, bivariate) vital sign distributions at the 25th week of gestation. For each pair of vital signs we present the 98th centile¹, which is an ellipse that encompasses the same proportion of the population as the red cut-offs calculated for C-ObsEWS from the univariate models (the 1st and 99th centiles, see Section 6.2.1). For comparison between multivariate and univariate distributions, in Figure 7.1 we also include the corresponding 1st and 99th univariate centiles.

Correlated vital sign pairs are those where the axes of the bivariate ellipse are rotated beyond 0° and 90°. The most highly correlated pair of vital signs are unsurprisingly SBP and DBP ($\rho = 0.50$ at 25 weeks gestation, as in Figure 7.1). DBP and HR are also positively correlated ($\rho = 0.21$ at 25 weeks of gestation). There is little correlation between the other vital sign pairs. Temp and RR are the only pair with negative correlation (i.e. a high RR is correlated with a low Temp).

¹For a bivariate Gaussian, the proportion of the variance that corresponds to a centile follows the χ^2 -distribution with 2 degrees of freedom. The 98th centile has a χ^2 value of 7.82.

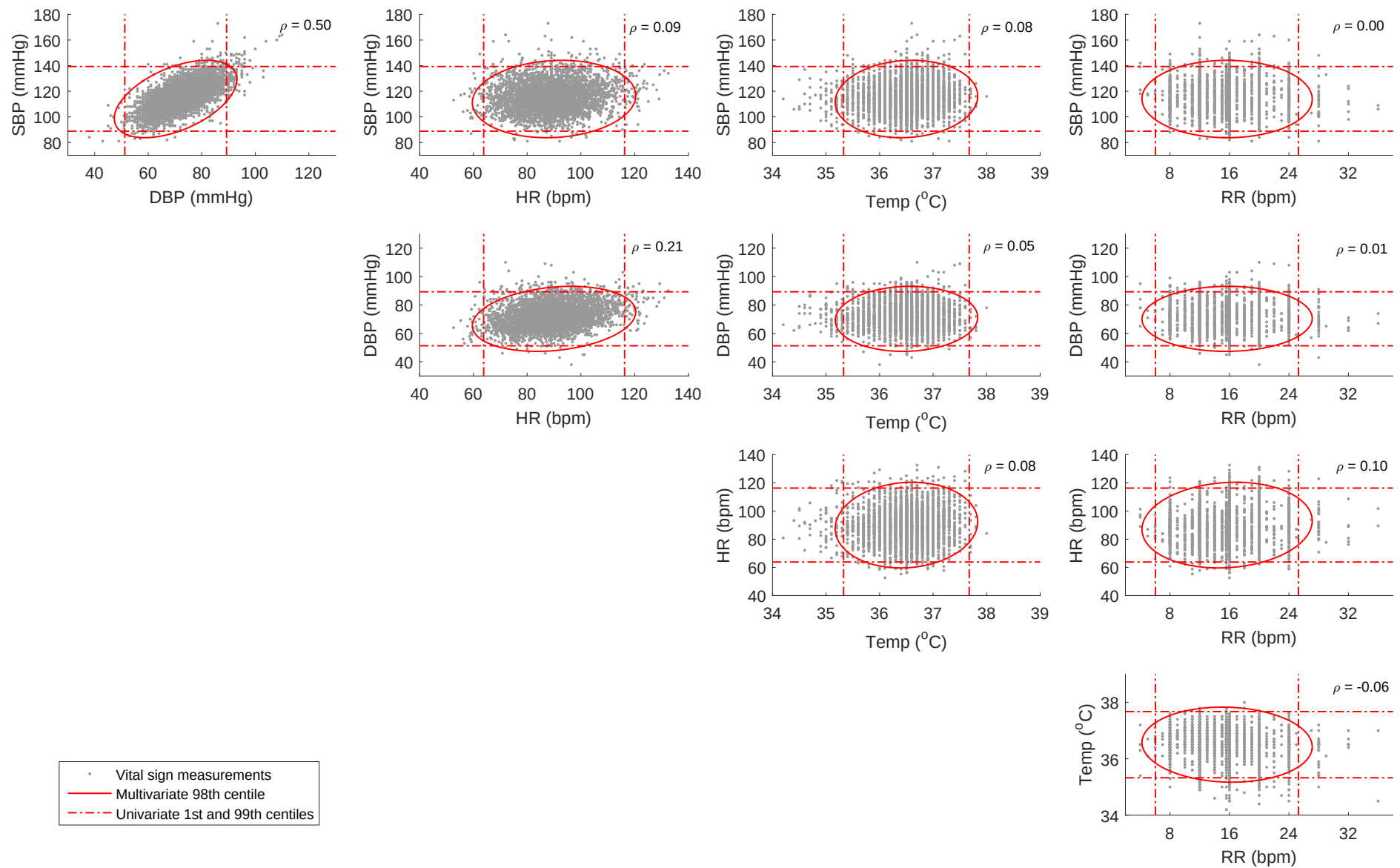


Figure 7.1: Relationships between each pair of vital signs, where ρ is the correlation coefficient. Shown are all vital sign sets (grey dots), the 98th centile (red) from the multivariate model, and the C-ObsEWS red cut-offs (dashed lines) derived from the univariate vital sign models developed in Chapter 5. Centiles are from the 25th week of gestation.

The correlations reported in Figure 7.1 are a snapshot of one time point in the antenatal period. The multivariate multilevel model also gives us information about how the correlation between two variables changes over time. For most vital sign pairs, correlation varies by less than 0.05 during the antenatal period. The correlation between SBP and DBP varies the most from a maximum coefficient of 0.57 at 10 weeks of gestation to a minimum of 0.50 at 26 weeks of gestation.

7.4 A continuous measure of abnormality

The EWS systems discussed in this thesis have all used a discrete measure to describe abnormal physiology. Whole numbers are assigned to regions of abnormality and summed to produce a discrete aggregate score. In this section we quantify the departure of a set of vital sign measurements from the mean using a continuous *novelty score*, z , to illustrate the differences between a discrete and continuous score.

The likelihood of a set of measurements, $\mathbf{y}_{ij}$, is calculated from the Gaussian distribution of the population model at time x_{ij} , described by:

$$p(\mathbf{y}_{ij}|\theta) \sim \mathcal{N}(X_{ij}\boldsymbol{\beta}, \mathbf{ZDZ}^T + \boldsymbol{\sigma}^2) \quad (7.10)$$

We define the z -score as the natural logarithm of the normalised likelihood, so that the minimum z -score is 0 at any time point ($\ln(1) = 0$). Let p_{max} be the maximum possible value of p at gestational age x_{ij} . The z -score is therefore described by:

$$z_{ij} = \ln \left(\frac{p_{max}}{p(\mathbf{y}_{ij}|\theta)} \right) \quad (7.11)$$

A low z -score indicates a “normal” vital sign set (high likelihood), and a high z -score indicates an abnormal vital sign set (low likelihood). For example, at 25 weeks gestation, the set of vital signs { SBP/DBP = 114/70 mmHg, HR = 90 bpm, Temp = 36.5 °C, and

RR = 15.7 rpm} gives a z -score of 0 because this set is the mean of the 5-D Gaussian distribution at this time point. If SBP is increased by 15 mmHg to 129 mmHg and the rest of the vital signs kept the same, the z -score is 1.3. If SBP is increased a further 15 mmHg, the z -score rises to 5.2. In addition to the SBP changes, if HR is decreased to 60 bpm (30 bpm decrease), the z -score will be 9.1. The z -score is always a positive number, which does not distinguish between an abnormally high and an abnormally low vital sign value.

The z -scores of the 3,658 antenatal vital sign sets from the 4P study are shown in Figure 7.2. The median z -score is 1.9 with the majority of the scores below 5.0 (92.9%). A z -score above 15 is unusual, occurring for just 7 vital sign sets (0.2%). Two high scores were due to a high RR (40 rpm), another two were due to abnormal temperature values, and the remaining three had abnormally high SBP (>160 mmHg).

The z -score does not weight all vital signs equally. The multivariate multilevel model learned parameters that described expected vital sign combinations. Table 7.2 compares the scores assigned to each vital sign when the value is at the equivalent univariate 90th,

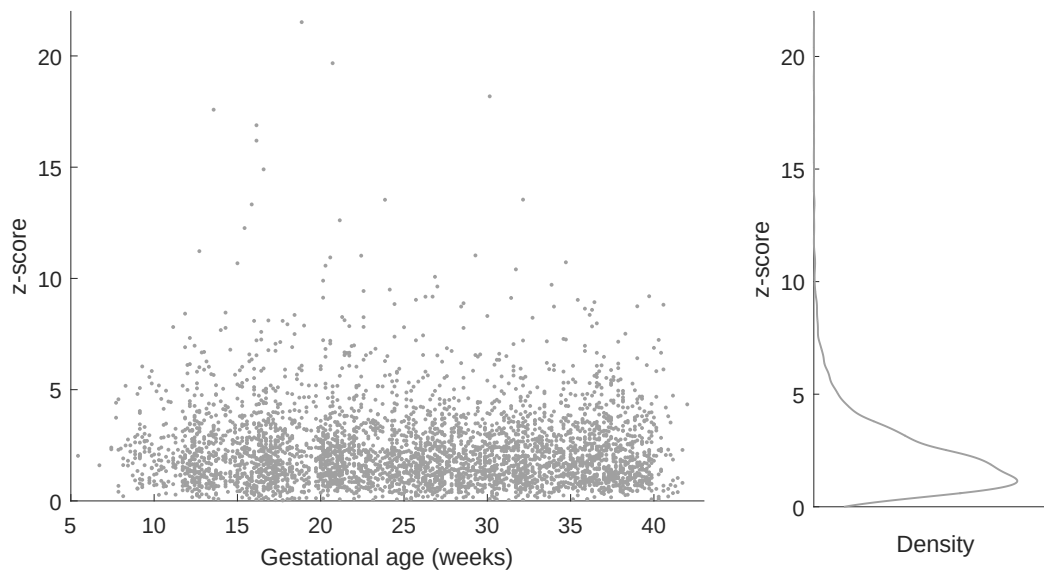


Figure 7.2: Z -score of each antenatal vital sign set from the 4P study (left) and the density of scores assigned (right).

95th, and 99th centile (the yellow, orange, and red cut-off values for the C-ObsEWS system). Deviations in SBP and DBP carry the most weight for the z -score, and RR the least weight.

Centile	90 th	95 th	99 th
C-ObsEWS score	1	2	3
<i>z-scores</i>			
SBP	1.1	1.9	3.7
DBP	1.2	1.9	3.8
HR	0.9	1.6	3.1
Temp	0.9	1.5	3.1
RR	0.8	1.4	2.8

Table 7.2: z -score from the multivariate model when each vital sign in turn is deviated from the mean by the corresponding 90th, 95th and 99th univariate centiles (all other vital signs are at the mean value).

The z -score is similar to the aggregate C-ObsEWS score developed in Chapter 6 in that both scores are a measure of the degree of physiological abnormality. For both scores a higher number indicates a higher degree of physiological abnormality. We compare the z -score and C-ObsEWS scoring system in Figure 7.3 by computing the respective scores for the 4P antenatal vital sign sets ($n = 3,658$). As the assigned C-ObsEWS score increases, the median z -score also increases, as expected. However, the values assigned (irrespective of scale) do not always agree. The z -score from the multivariate model incorporates information about vital sign correlation, however since there is little correlation between vital sign pairs except for SBP and DBP (as shown in the previous section), we hypothesise that the difference observed between the z -score and C-ObsEWS score is primarily due to the quantised nature of the C-ObsEWS system, particularly evident for measurements in the extremes of physiology.

As previously mentioned, the C-ObsEWS score assigns *whole* numbers to regions of abnormality for each vital sign, from 0 (normal, white region) to 3 (most abnormal, red region). Two vital signs in the red region will therefore produce an aggregate C-ObsEWS

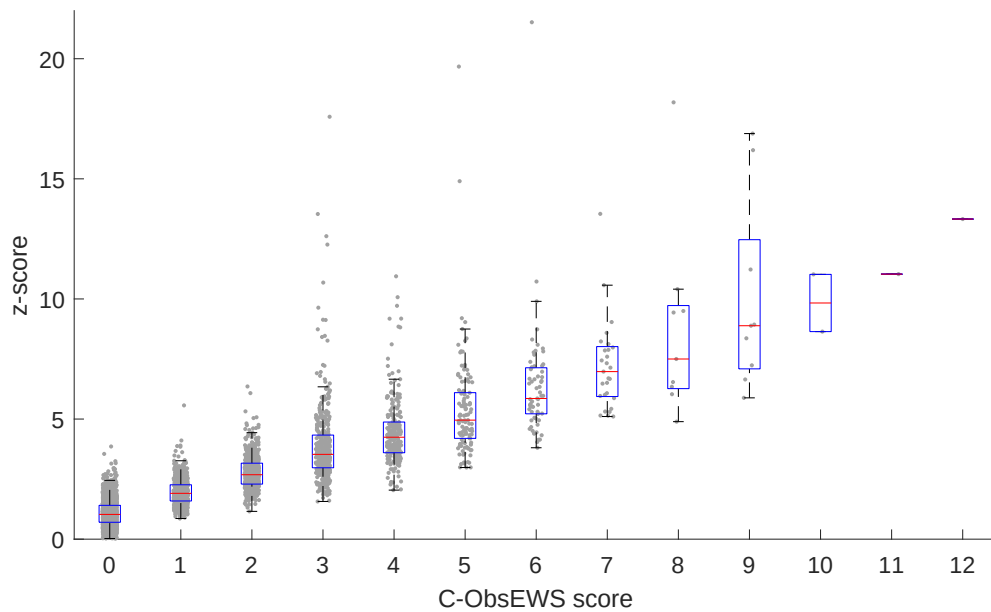


Figure 7.3: *Z-score of each antenatal vital sign set from the 4P study compared with the aggregate C-ObsEWS score (developed in Chapter 6) for the same vital sign set. The box-plot overlaid shows the median (red) and IQR (blue) of z-scores for each C-ObsEWS score.*

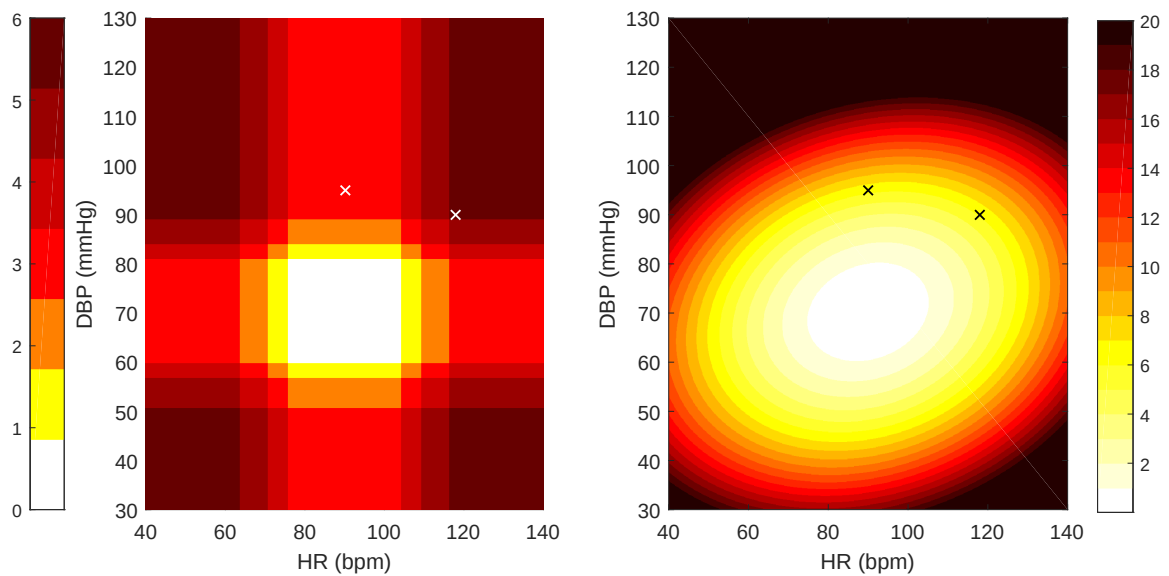


Figure 7.4: *Comparison between the discrete C-ObsEWS scoring system (left plot) and the continuous z-score from the multivariate model (right plot).*

score of 6. In contrast, the z -score is a *continuous* number based on the likelihood of a set of vital sign readings in a multivariate probability distribution. Consider the bivariate distribution of DBP and HR shown in Figure 7.4 (at 25 weeks gestation). The left plot shows the C-ObsEWS aggregate scores assigned to combinations of DBP and HR values, and the right plot shows the corresponding z -scores. These correlate to the z -score, but do not reflect the absolute z -score values.

In Figure 7.4, the two white crosses (left plot) and black crosses (right plot) show two abnormal pairs of {HR, DBP} measurements. The first is at {118 bpm, 90 mmHg} and falls in the red region (score 3) for both the DBP and HR univariate models, giving an aggregate score of 6. The second measurement is at {90 bpm, 95 mmHg} which is in the normal region for HR (score 0), but again the red region for DBP (score 3), giving an aggregate score of 3. The same two sets of measurement on the multivariate model both fall in the outer 2% of the population. Assuming all other vital sign measurements are normal, the z -scores for the two sets of measurements would be 6.2 and 6.4 respectively. Thus in this example, two sets of abnormal vital sign measurements give different C-ObsEWS scores yet similar z -scores.

7.5 Discussion

In this chapter we fitted a multivariate multilevel model to antenatal vital sign measurements to explore previously unknown relationships between vital signs during pregnancy. The model learned the correlation between vital signs and described the 5-D trajectory of SBP, DBP, HR, Temp and RR during gestation. The most highly correlated pair of vital signs are SBP and DBP, and some correlation is present between DBP and HR. The other vital sign pairs have a correlation coefficient less than 0.1. Over the course of pregnancy there are few changes in vital sign correlation.

We also used the multivariate model to illustrate the difference between a discrete

and continuous measure of physiological abnormality. The C-ObsEWS system developed in Chapter 6 assigns each vital sign measurement a value between 0 and 3, then sums the values from all vital signs to give a discrete aggregate score. In contrast, we quantified deviation from the mean in the multivariate model in a continuous novelty score, z , in terms of the likelihood of the underlying probability distribution. The quantisation of the C-ObsEWS system is most apparent for very abnormal physiological measurements. We illustrated how sets of vital signs could give different C-ObsEWS scores yet a similar z -score using the distribution of DBP and HR at 25 weeks gestation. A validation study would be required to determine whether a continuous description of abnormality would enable better discrimination between patients who require clinical intervention and those who do not, than the standard ObsEWS scoring systems currently implemented. One barrier to the use of a continuous number is the unknown clinical interpretation. Hospital staff are used to calculating and interpreting vital sign measurements using whole numbers, so a new system would have to be introduced gradually.

The main limitation of the multivariate model described in this chapter, is the exclusion of SpO_2 . SpO_2 is one of the six vital signs important for routine monitoring, and it therefore would be necessary to include it in a model that is used in clinical care [7]. Other techniques for multivariate data analysis could be considered, such as Parzen windows, which uses kernel density estimation to determine the probability density function. This is a non-parametric technique that does not assume a distribution function, and so would allow the inclusion of SpO_2 .

In summary, the work presented in this chapter explores two new directions for a hospital alerting system: the consideration of correlation between vital signs, and the quantification of physiological abnormality using a continuous variable. These features would be too complex to implement in paper charts, so would require an electronic system in which scores were calculated automatically.

Conclusions and future work

Timely recognition of physiological deterioration can reduce the prevalence and severity of pregnancy-related complications, however the normal physiological changes in pregnancy are known to complicate early recognition and yet these changes are not well documented. The overall aims of the investigations presented in this thesis were two-fold. Firstly we aimed to develop algorithms to ensure good-quality vital sign data. Respiratory rate was a particular focus since it is the only vital sign that is not currently routinely measured on hospital wards using an electronic device. Secondly, we aimed to establish the pregnancy-associated changes in vital signs (SBP, DBP, HR, Temp, SpO₂, and RR), and use this knowledge to propose an evidence-based hospital alerting system specific to the obstetric population. In this chapter we summarise the main results presented in this thesis and discuss general directions for future work.

8.1 Summary of thesis

The first half of the thesis focused on the collection of good-quality vital sign data during all stages of pregnancy. In Chapter 2 we provided the motivation for this work. A systematic review of studies that reported vital sign measurements during pregnancy revealed that only BP and HR changes during the antenatal period were well understood. Few studies reported measurements of Temp, SpO₂, or RR during pregnancy, or described changes in the immediate postpartum period for any vital sign. Data from

those studies that reported BP and HR measurements were synthesised using a multilevel model for aggregate data. Both SBP and DBP dipped in mid-pregnancy, and HR increased steadily throughout pregnancy. Our investigations were limited due to the poor-quality of many studies (assessed using the QUADAS-2 criteria [56]) which contributed to heterogeneity in the dataset. Particular areas of concern were a lack of accuracy in measurement and reporting of gestational age, and little evidence that equipment had been calibrated or validated for use in pregnancy.

The limited understanding of pregnancy-associated vital sign changes led to the design of the 4P (Pregnancy Physiology Pattern Prediction) study, introduced in Chapter 3. The 4P study started in Oxford in August 2012, designed to collect vital sign data and other demographic information from pregnant women during pregnancy, labour, and after delivery. This was the first prospective, multi-centre study to collect vital sign measurements longitudinally through all stages of pregnancy. The 4P digital health system was designed to aid the collection of good-quality data. Hospital-grade medical devices were combined with a custom-designed tablet application and website to enable real-time transmission and review of vital sign data. The 4P system was successfully used by midwives, and women during the self-monitoring phase. For the investigations in this thesis, we used 10,000 sets of vital sign measurements collected from the 620 healthy women in the 4P study who had delivered by the 15th July 2016. By the 9th January 2017, when this thesis was submitted, a total of 14,792 vital sign measurement sets had been collected from 1,054 women across three hospital sites, of whom 877 had delivered. Recruitment to the 4P study has now closed, and follow-up measurements and data cleaning is expected to be completed by July 2017.

To further ensure good-quality vital sign data we developed an SQI for the PPG waveform (obtained from the pulse oximeter). This used dynamic time warping to identify poor-quality beats in the PPG. Corresponding poor-quality measurements of

HR and SpO₂ were removed. If less than 30 seconds of good-quality signal remained after exclusion of poor-quality sections, the whole recording was discarded. Our SQI strategy excluded the majority of erroneous SpO₂ values, most of which were below 90%.

In Chapter 4 we focused on the quality of RR measurements. We used frequency-domain and time-domain methods to estimate RR from the PPG waveform (pulse oximeter) and accelerometer and gyroscope waveforms (smartphone). Analysis was performed using 697 sets of recordings collected from system 3 of the 4P study (for which midwives counted respirations for 60 seconds). RR estimates from the y -axis of the gyroscope were closest to the reference RR, with a mean absolute error of 1.13 bpm. This error is low because we only used recordings for which we could validate the reference RR. While counting respirations, midwives indicated the time of each breath by pressing a keypad button on the tablet application. If the RR calculated from the mean length between keypad taps did not agree to within 2 bpm of the RR reported by the midwife, the recording was not used to evaluate the RR algorithms.

We recommend using the pulse oximeter and/or smartphone for RR measurements in remote monitoring. To ensure a good-quality RR estimate at the time of measurement, information about signal quality should be included. We assessed the quality of the derived respiratory waveforms (after filtering and other pre-processing steps) using autocorrelation. When poor-quality signals were removed, the mean absolute error of RR estimates decreased. If a recording is of poor quality at the time of remote measurement, a new recording could be requested. During the self-monitoring phase of the 4P study, women monitored their own vital signs at home, but could not measure their own RR. The methodology presented in this thesis to derive RR from the pulse oximeter would allow knowledge of respiratory information for 5,264 PPG waveforms collected during self-monitoring in the 20 days after delivery.

The second half of the thesis focused on the creation of an obstetric alerting tool. In Chapter 5 we applied multilevel models to the 4P vital sign data to understand the trajectories of each vital sign during pregnancy and in the two weeks after delivery. Multilevel models are appropriate for longitudinal studies because they use a series of levels to account for hierarchical structures in the data. We chose the standard polynomial to describe the shape of vital sign trajectories over time, because of a better fit to the vital sign data than the fractional polynomial and restricted cubic spline formulations. We presented reference ranges for SBP, DBP, HR, Temp, SpO₂, and RR, for both the antenatal and postnatal periods. The shape of the trajectories confirmed the known trends of antenatal BP and HR established in Chapter 2, and gave new information about other vital sign trends. Of note was the observed increase in SpO₂ at the beginning of pregnancy (97.5% at 10 weeks gestation) and the subsequent decrease during gestation (96.5% at 40 weeks gestation). Changes in SpO₂ during pregnancy have not previously been reported.

In addition to vital sign changes at the population level, we investigated the vital sign changes associated with parity and BMI by fitting separate multilevel models to each group (parity) and including additional covariates in the model (BMI). We showed that nulliparous women had a slightly higher SBP and DBP than multiparous women during pregnancy, and a higher HR in the two weeks after delivery. BMI was included as a continuous variable, and affected the expected BP and HR trajectories during pregnancy. The difference in expected values of SBP and DBP with BMI was large at the beginning of pregnancy and then decreased throughout pregnancy. BMI did not have an effect on HR values in the initial weeks of pregnancy, but a difference in expected HR values was observed towards the end of gestation.

In Chapter 6 we translated the univariate vital sign trajectories into an obstetric early warning score (ObsEWS). A review of existing ObsEWS systems highlighted the

inconsistencies in cut-off values and trigger thresholds, and the lack of evidence-based approaches to design. We developed the C-ObsEWS system based on recommendations by Quinn *et al.* [48]. All six vital signs were included (SBP, DBP, HR, Temp, SpO₂, and RR). Vital sign cut-offs were based on the outer centiles on the population as proposed by Tarassenko *et al.* [174] so that the system would alert for patients exhibiting the extremes of expected physiology [37]. Furthermore, the C-ObsEWS proposed here was specific to the time of gestation and number of days after delivery. This addresses a key concern that early warning systems are not relevant for the postnatal period [42].

We recommend the use of an ObsEWS to identify those inpatients with deteriorating physiology. The C-ObsEWS system proposed in this thesis was developed from 10,000 vital sign measurements from 620 healthy women. We demonstrated the impact of such an early warning score by examining the case studies of two women who developed sepsis in the postnatal period. The proposed C-ObsEWS system was more sensitive to vital sign changes than the Oxford MEOWS chart, particularly for DBP values, which may have led to different clinical care.

8.2 Future work

At the end of each chapter, we have suggested improvements to the analysis methods presented. Here, we discuss general ideas for future direction of an ObsEWS.

8.2.1 Validation of an obstetric early warning score

The next step in the design of the C-ObsEWS system, is a validation study. In this thesis we illustrated the difference between the C-ObsEWS and the Oxford MEOWS with two example patients who developed sepsis in the postnatal period. However, large-scale analysis is required to set the trigger threshold for a clinical response, ensure the C-ObsEWS correctly identifies patients with deteriorating physiology, and to compare the C-ObsEWS with existing ObsEWS systems.

In order to perform validation, an important question is: what outcomes should an ObsEWS alert for? We have suggested admission to ICU or the development of a specific maternal complication. The recent definition of maternal morbidity presented by the WHO working group could be considered a good place to start [17]. A validation dataset is then required. This must contain vital sign measurements from patients on a general obstetric ward (non-ICU), with a reasonable number of cases of the chosen outcome. The trigger threshold can be set using an AUROC curve to achieve a balance between sensitivity and specificity.

The second step in the assessment of the C-ObsEWS system should be a prospective clinical intervention trial. However, these studies are not practical. They would require cluster-randomisation, because two different alerting systems cannot be deployed within the same institution, and therefore a considerable amount of time and money to complete (for example, the MERIT study included 23 hospitals in Australia yet no difference in outcome was recorded between the clusters [186]). In addition, a hard outcome (i.e. death) is required because different hospitals use different decision criteria, for example the criteria for admission to ICU. An alternative approach could be a stepped-wedge implementation, as recently demonstrated in the deployment of an electronic EWS recording system in Oxfordshire [187]. This approach proposes that evaluating roll-outs by using the effects of the steps (no randomisation) helps to manage the effects of other changes over time, which is a major draw back of the *before-and-after* study approach. As obstetric complications are fortunately rare, a prospective study of any design would have to span multiple centres over a considerable time-period.

The investigations in this thesis have demonstrated that vital sign changes are present in both the antenatal and postnatal stages of pregnancy. Currently, pregnant women and recently pregnant women are monitored using the same ObsEWS system which does not recognise pregnancy-associated vital sign changes. We designed the C-ObsEWS system

to change with time. In clinical care this means two early warning scoring systems are required: one for the antenatal stage of pregnancy (gest-ObsEWS), and a second system for the postnatal stage of pregnancy (post-ObsEWS).

With respect to the implementation of an ObsEWS in clinical practice, electronic systems are able to store and calculate algorithms in real-time. An electronic system for collection of vital sign data would therefore allow the implementation of C-ObsEWS where cut-off values vary with gestational age or time after delivery. Subject-specific covariates such as maternal age, BMI or parity could also be included. Furthermore, electronic systems overcome current obstacles for effective EWS implementation such as human error in the recording and computing of aggregate scores [188].

8.2.2 Further development of an early warning score

In Chapter 7 of this thesis we introduced a multivariate multilevel model to describe vital sign changes during pregnancy in multiple dimensions. Further investigation is required as to how a multivariate model can be used to improve clinical practice. We demonstrated how a multivariate model captures correlation between vital signs, and how the abnormality of a set of vital sign measurements could be quantified using a continuous novelty score. The probability distribution described by this multilevel model assumed a multivariate Gaussian distribution. An alternative approach to defining a probability distribution function is to use a non-parametric model, such as Parzen windows. This model uses a kernel function to describe the density observed in a dataset. It would allow the inclusion of SpO_2 , however the aspect of time (i.e. gestational age) would be lost. Retrospective analysis should be performed to compare univariate multilevel models, the multivariate multilevel model, and a non-parametric approach such as Parzen windows.

An electronic system for collection of vital sign data would also be advantageous for the implementation of a multivariate alerting system. This would readily allow the

incorporation of non-integer variable weightings, or the calculation of a fully continuous novelty score.

8.2.3 Monitoring during labour

The investigations in this thesis have focused on the antenatal and postnatal periods of pregnancy. The next step is the investigation of normal vital sign values during labour (the intrapartum period). The 4P study has collected vital sign information recorded on partogram charts, which could be used to examine the common patterns in vital sign progression during labour. An open question is whether antenatal reference ranges should inform the expected vital sign values during labour, depending on the woman's gestational age at the time of labour. A woman who delivers pre-term would not have experienced the changes in vital signs expected in the final stages of pregnancy.

8.3 Final remarks

There is a long way to go before a nationally agreed ObsEWS system is validated and ready for implementation in clinical practice; the investigations presented in this thesis are only an initial step. Yet it is intuitive that in-depth knowledge of pregnancy-associated vital sign changes will enable earlier detection of physiological deterioration, lead to an improvement in maternal care, and ultimately reduce the prevalence and severity of serious maternal morbidity and mortality.

Definition of healthy

Under the definition of *healthy*, *normal* or *low-risk*, women with the following conditions were excluded from the data set used for analysis in this thesis (see Section 3.3.1):

- 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 were *not* reasons for exclusion:

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

Supplementary material: Pulse oximeter quality

B.1 Peak detection algorithm

The following algorithm was used to determine the extrema in a signal. Extrema are either peaks (negative turning points) or troughs (positive turning points). As this algorithm was most commonly applied to the photoplethysmograph (PPG) signal from the pulse oximeter, particular attention was made to the exclusion of the dicrotic notch.

1. All extrema are identified using the zero-crossing points of the differentiated signal from a 5pt differentiating filter,

$$y[n] = 0.3x[n - 2] + 0.7x[n - 1] - 0.7x[n + 1] - 0.3x[n + 2]$$

2. Peaks below a predetermined threshold (peak-threshold), and troughs above the mean of the signal are discarded.
3. A peak must be followed by a trough and vice versa. If two peaks or two troughs are consecutive, the extremum furthest from the mean of the signal is kept and the other discarded.

The peak-threshold was set as $mean(PPG) + 0.2IQR(PPG)$, determined empirically by examination of 26 PPG traces with a prominent dicrotic notch. Figure B.1 illustrates

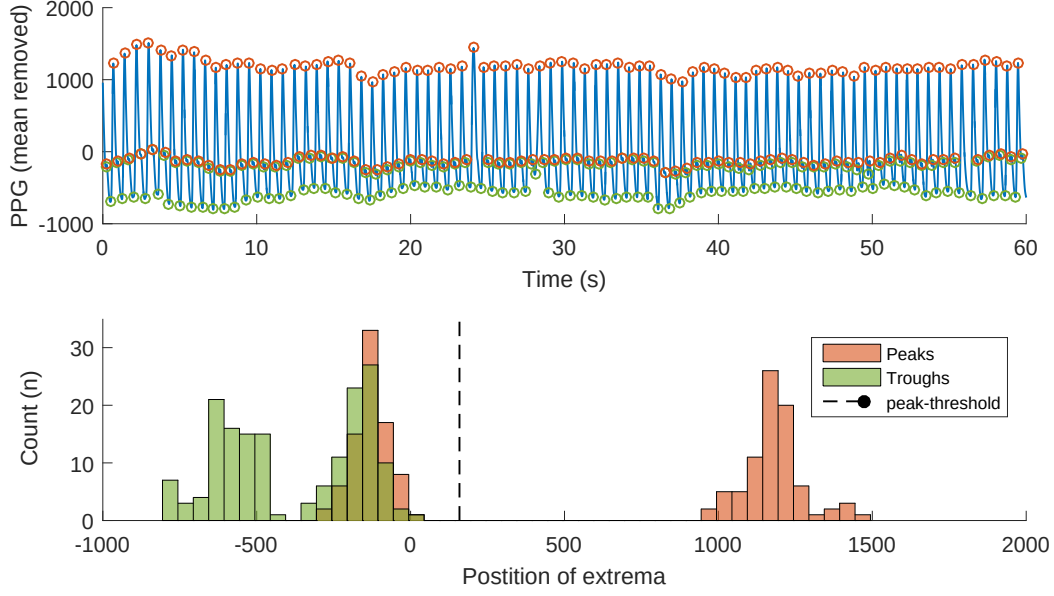


Figure B.1: *Top:* Peaks (orange) and troughs (green) identified in a PPG signal. *Bottom:* There is a clear separation between the peaks corresponding to the dicrotic notch, and the peaks of the overall signal. The peak-threshold eliminates all false peaks.

the relative position of dicrotic notch extrema with respect to the desired peaks of the overall signal. In this example there is a clear distinction between the two sets of peaks, so a threshold can cleanly remove the undesired peaks. However, not all signals are clean and regular. In particular, three types of PPG traces provide challenges for peak detection. Firstly, signals with a large baseline wander (relative to the height of the signal) means that some dicrotic notch peaks will be similar to the true peaks of the signal. We therefore include a threshold for troughs in addition to the peak-threshold. Combined with the final condition (peaks and troughs must be alternate), this is successful at dealing with large baseline wander. Secondly, a change in the amplitude of a signal leads to a shift in the position of the true PPG peaks. Amplitude changes can occur either suddenly or gradually due to the automatic gain control feedback system within the pulse oximeter device. We therefore set the threshold as close as possible to the dicrotic notch. Thirdly, signals with large motion artefacts distort the range and distribution of the signal. Additionally, no two PPG signals have the same amplitude.

We therefore set the peak-threshold relative to the amplitude of the signal in question, and in terms of the interquartile range (IQR) which, unlike standard deviation, is a measure of variance minimally affected by changes in range.

We have chosen not to include a refractory period (a minimum time between consecutive peaks) in this peak detection algorithm, despite it being a popular choice to eliminate unwanted extrema. A refractory period is dependent on the first extremum being identified correctly, and does not take into account missed beats or unpredictable motion artefacts. The peak detection algorithm described here deals with the entire signal at once, and is appropriate and effective for the signals described in this thesis.

B.2 SQI(Flags)

The SQI(Flags) identifies poor quality sections of the PPG signal using the flags generated by the Nonin pulse oximeter device. The Nonin flags that indicate poor signal quality are:

- Artefact (flagged once per beat)
- Sustained artefact (absence of consecutive good pulse signals)
- Sensor error (unusable data; set when no finger is detected)
- Pulse quality:
 - Pulse quality = 1: poor quality pulse detected
 - Pulse quality = 2: moderate quality pulse detected

Nonin flags are either true or false, except for pulse quality which returns a numeric range from 0 to 3 where 0 is no pulse and 3 is a good quality pulse. The most obvious indicator of poor quality is the *sensor error* flag which indicates that there is no finger detected. This usually correlates with an invalid reading (where vital sign values returned by the device are 511 rpm for HR and 127% for SpO₂) which were used to identify poor quality sections in SQI(Invalid).

The following rules were used to create SQI(Flags):

1. All occasions where a Nonin flag was true were marked as poor quality.
2. As artefacts were typically only flagged once per beat, a buffer (T_{pulse}) was included immediately before and after every artefact flag to label the whole beat as artefactual. This was set to $T_{pulse} = 1.0$ seconds, corresponding to the top 5% pulse lengths (HR of 60 rpm).
3. 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 $T_{cons} = 3$ seconds, approximately 3 beats.

B.3 SQI(DTW)

The SQI(DTW) is a signal quality index for the PPG that identifies individual beats that are of poor quality. We presume that all beats in a PPG signal are similar, allowing for small modulations in height and width due to respiration and HRV. It should therefore be possible to map all beats onto an ideal template beat using dynamic time warping (DTW). If an individual beat correlates poorly with the template after DTW, we mark the beat as poor quality. DTW has been used by Qi and Clifford (2012) [85] to mark and then predict poor quality beats in a PPG signal. Our SQI(DTW) is based on their method to mark poor quality beats. We do not incorporate a prediction stage since we are not interested in implementing an SQI in real-time monitoring. Our algorithm includes the following four steps:

1: Split PPG signal into beats

The PPG signal was split into individual beats by first identifying the extrema using the peak detection algorithm described in Appendix B.1. A beat was defined as lasting from a trough to the next consecutive trough. The beats were normalised such that the start of the beat (trough) was at 0 and the peak at 1. If no beats were identified, no further analysis was performed and the entire PPG signal was considered poor-quality.

2: Create beat template

A beat template, T1, was generated by averaging all beats within the PPG signal. Beats that had already been identified as poor quality by SQI(Invalid) and SQI(Flags) were excluded. If there were less than 5 beats available to construct the template, no further analysis was carried out and the entire PPG was considered poor-quality since DTW could not be performed.

Template T1 was set to have length L, the median length of all beats. The median length, rather than the mean, was chosen because beats affected by motion artefact can often be two or three times the length of a regular beat. In such a case the template length may exceed the true beat length and thus contain some of the upward stroke of the next beat. During step 3, when individual beats are fitted to the template using DTW, low correlations would not allow the beat quality to properly be assessed.

After the construction of T1, all beats were directly correlated with T1 (no DTW performed). Beats that did not highly correlate (where correlation coefficient < 0.8 , as used by Qi and Clifford) were excluded, and a new template, T2, constructed with the remaining beats. If more than half the beats were excluded then the threshold was reset to the median correlation coefficient so that an improved template could still be constructed. If the median correlation coefficient was less than 0.5, no further analysis was performed and the whole PPG signal was considered poor-quality since the beats were not consistent.

An example of beat template construction is shown in Figure B.2.

3: Determine Beat Quality (BQ)

To assess beat quality, each beat was fitted to the new template, T2, using DTW (Appendix B.4). However, before performing DTW, we first split the PPG beats and template T2 into straight line segments using the piecewise linear approximation algorithm described by Koski and Juhola (1996) [189]. Line segmentation aims to find the best

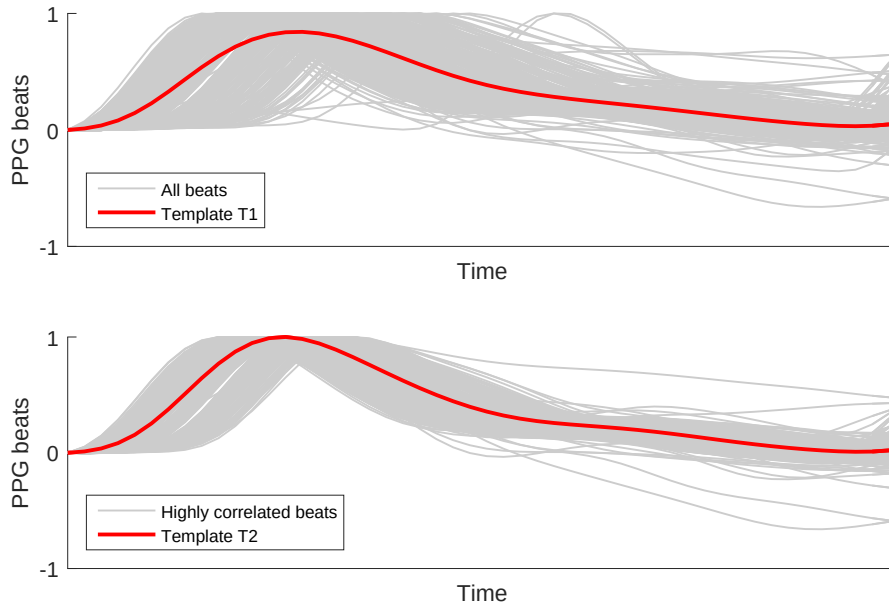


Figure B.2: Example of individual beats extracted from a PPG signal. **Top:** All beats are averaged to form an initial template, $T1$. **Bottom:** Beats with poor correlation to $T1$ are excluded and the template reconstructed to form $T2$.

approximation of the original signal with a continuous broken line. It is one way to extract important information from raw signals, and considerably speeds up computation time when using intensive algorithms such as DTW. We chose straight line segmentation because it was simple, easy to process, and appropriate for signals for which several consecutive points lie almost within a straight line, such as PPG signals.

To select each line segment, two steps are undertaken. Firstly, the number of points included in the line segment are increased until the maximum residual distance is above a predetermined bound. The residual distances are the distances between the individual points and the straight line between the first and last point in the segment. Secondly, the segment was cut at the point with maximum residual distance (furthest away from the straight line) and the second half discarded. The straight line and residual distances were recalculated for the new segment. The segment is iteratively cut in this way until the maximum residual distance falls under the predetermined bound once again. We empirically set the bound at 0.015 units (the amplitude of the beat was previously

normalised to 1). Figure B.3 shows a single PPG beat before and after line segmentation. This example is from a 4P Oxford participant at 32 weeks gestation.

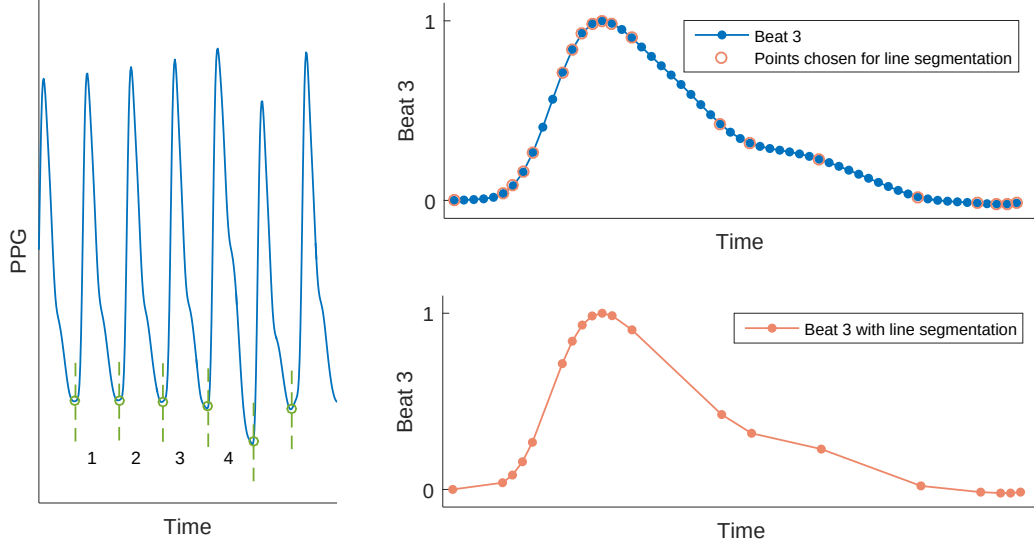


Figure B.3: Illustration of a PPG beat before (top right) and after (bottom right) line segmentation. Straight lines are chosen to reduce the computation time when performing dynamic time warping (DTW). This example is beat 3 from the PPG signal on the left.

After performing line segmentation of the template T2 and all individual beats, each beat was fitted to the template T2 using DTW (see Figure B.4). DTW is further explained in Appendix B.4. Beat Quality (BQ) was defined as the correlation coefficient of each dynamic-time-warped beat with T2. Negative BQ values were set to zero.

4: SQI(DTW)

For the majority of beats observed, BQ was between 0.97 and 1.00. As we were only concerned with excluding very poor quality beats, we set the threshold below this range. Thus, beats with a $BQ < 0.96$ were considered poor-quality and their SQI(DTW) was set to 1. If there were less than three beats between consecutive poor quality beats, the beats between these beats were also marked for exclusion. This is similar to T_{cons} in SQI(Flags) (Appendix B.2).

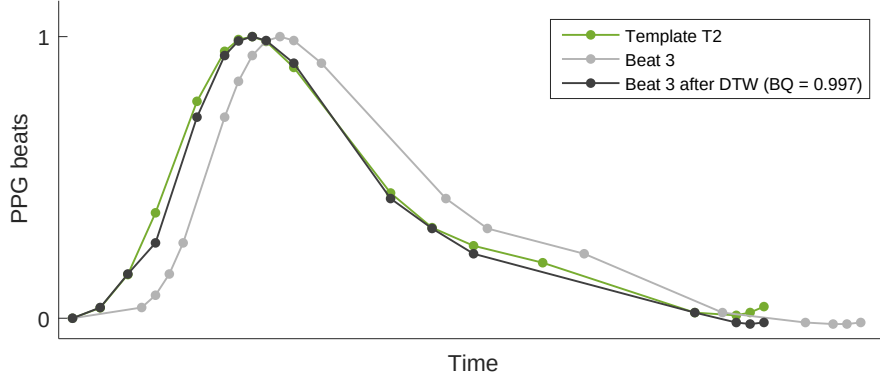


Figure B.4: The current beat (grey) is fitted to the template (green) using DTW to give a new beat (black). DTW warps the time scale so that the difference between slope segments is minimised. No adjustment is made to the signal amplitude. The BQ of this example is 0.997 which is considered good quality.

B.4 Dynamic Time Warping (DTW)

Dynamic time warping (DTW) finds an optimal alignment between two time-dependent sequences by warping (reducing or enlarging) time intervals. DTW therefore allows a measure of similarity between sequences that may vary in time or speed. In the signal quality index for PPG signals, SQI(DTW), individual PPG beats (trough to consecutive trough) are compared to a template using DTW. If the dynamic-time-warped beat highly correlates with the template the beat is considered good quality, else it is marked as poor quality. Here we explain DTW, adapted from Li and Clifford [85].

Consider two time series T and B , of lengths n and m respectively, such that

$$T = t_1, t_2, \dots, t_i, \dots, t_n \quad (\text{B.1})$$

$$B = b_1, b_2, \dots, b_j, \dots, b_m \quad (\text{B.2})$$

To align the two sequences using DTW, an $n \times m$ distance matrix (D) is constructed where the (i^{th}, j^{th}) element of the matrix contains the distance $d(t_i, b_j)$ between the two points t_i and b_j .

For our SQI(DTW) application, let T be the template and B an individual PPG beat.

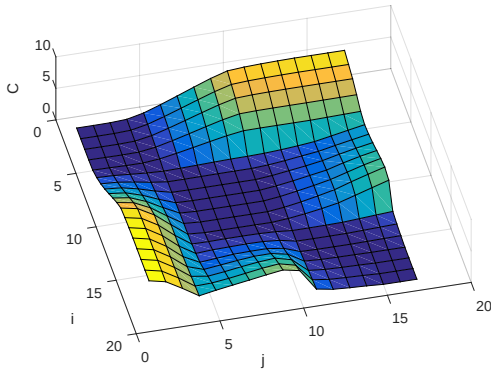
Given we split the time series T and B into line segments as described in section B.3, we define the distance between each short line pair $(d(t_i, b_j))$ as the absolute difference between the slopes of each short line.

The aim of DTW is to find an optimal path from $(1, 1)$ to (n, m) by minimising the cumulative distance of the path. We construct a cumulative matrix (C) which contains the sum of all previous distances. For line i and j , the cumulative distance $c_{i,j}$ is defined by:

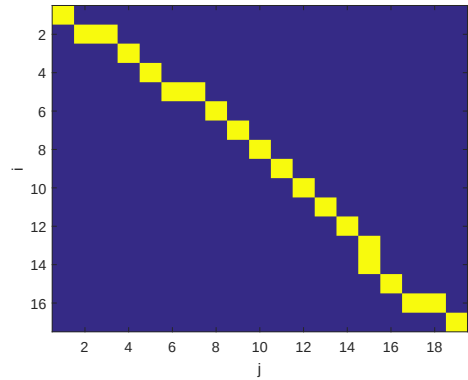
$$c_{i,j} = \begin{cases} c_{i-1,j} + d(t_i, b_j)l(t_i) \\ c_{i-1,j-1} + d(t_i, b_j)(l(t_i) + l(b_j)) \\ c_{i,j-1} + d(t_i, b_j)l(t_i) \end{cases} \quad (\text{B.3})$$

where $l(t_i)$ and $l(b_j)$ are the durations of lines t_i and b_j in the time series, respectively.

The optimum *warping* path is the path through the matrix, from $c(1, 1)$ to $c(n, m)$ that has the minimum cumulative distance. An example cumulative matrix and optimum warping path are shown in Figure B.5.



(a) Example cumulative matrix for DTW. The cost is the absolute difference in slope between each line segment.



(b) The optimum warping path that minimises the total cost (yellow blocks).

Figure B.5: DTW process.

Supplementary material: Multilevel models

This supplement provides additional information about multilevel models in the context of Bayes rule and numerical sampling, and specific information about how inference was performed for the multilevel models implemented in Chapter 5 and Chapter 7.

Multilevel models arise from the need to accurately model nested and clustered data structures. Cases where multiple observations are contributed from a single participant, such as in a longitudinal study, are examples of a nested data structure. In these cases, the standard assumption of independence is likely to be violated because of dependence among observations from the same participant. Multilevel models extend conventional linear regression to handle such dependence.

The multilevel models we consider in Chapter 5 are linear in their parameters, that is the relationship between the parameters, β , and covariate matrix, $\mathbf{X}_i$, is linear. Linear multilevel models are also referred to as linear mixed-effects models, random coefficients models, random parameter models, and nested models. We choose to use the term *multilevel model* because it emphasises the hierarchical structure and avoids the ambiguous terms *fixed* effects, *random* effects, and *mixed* effects.

C.1 Bayes rule and prior distributions

Bayesian models incorporate prior information about parameters using Bayes rule, as stated in Equation C.1. Here, y is the data and θ is a general term for unknown parameters that will be learned from the data.

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \quad (\text{C.1})$$

Bayes rule is often said as “the *posterior* is proportional to the *likelihood*, multiplied by the *prior*”. This statement eliminates the normalising factor, $p(y)$, which is not necessary to compute during inference. The *likelihood*, $p(y|\theta)$, describes the distribution of the data; the *prior*, $p(\theta)$, incorporates prior knowledge about the distribution of the parameters.

A hierarchical model, such as the multilevel model described in Chapter 5, includes another layer of parameters that are also estimated from the data. Hyperparameters, ϕ , describe the distribution of parameters, θ . The inclusion of hyperparameters adds a second level to Bayes rule that describes the posterior distribution of the hyperparameters, after integration over the parameters. The two levels are described by Equations C.2 and C.3. A Bayesian model can have any number of levels.

$$p(\theta|y, \phi) = \frac{p(y|\theta, \phi)p(\theta|\phi)}{p(y|\phi)} \quad (\text{C.2})$$

$$p(\phi|y) = \frac{p(y|\phi)p(\phi)}{p(y)} \quad (\text{C.3})$$

Prior distributions were assigned to all second-level parameters. The population parameters, β , were assigned a prior Gaussian distribution centred on zero; the within-subject variance, σ , was assigned a prior Gamma distribution. The between-subject covariance matrix $\mathbf{D}$ was decomposed into a vector of coefficient scales, σ_b , and a corre-

lation matrix, $\mathbf{\Omega}$, such that:

$$\mathbf{D} = \sigma_b \mathbf{I} \mathbf{\Omega} \sigma_b \mathbf{I} \quad (\text{C.4})$$

A uniform distribution was assigned as a non-informative prior for the coefficient scales, σ_b . An “LKJ” correlation distribution was chosen as a prior for the $\mathbf{\Omega}$ correlation matrix, as recommended by the designers of the Stan programming language (Carpenter *et al.* [190], see Appendix C.3). The LKJ distribution is described by Equation C.5, where the shape parameter, η , describes the height of the diagonal elements of $\mathbf{\Omega}$ compared to the off-diagonal elements.

$$\text{LKJ}(\mathbf{\Omega}|\eta) \propto |\mathbf{\Omega}|^{\eta-1} \quad (\text{C.5})$$

C.2 Inference using sampling methods

Numerical integration by simulation (“sampling”) is one way to summarise parameters in a probabilistic model when no exact analytical solution exists. The goal of sampling is to obtain a set of independent draws from the posterior distribution, $p(\theta|y)$, with enough draws, S , so that the quantities of interest, parameters θ , can be estimated with reasonable accuracy. Here we briefly discuss the principle of Markov chain simulation (also called Markov Chain Monte Carlo, or MCMC), a general method based on drawing values of θ from approximate distributions and then correcting those draws to better approximate the target posterior distribution $p(\theta|y)$. For more information on numerical integration by simulation, we recommend Chapters 10-13 of Gelman *et al.* [161], and Chapter 11 of Bishop [191].

MCMC sampling is performed sequentially by constructing a series of Markov chains¹. Each chain sequence is produced by starting at some point θ^0 and for each iteration s , drawing θ^s from a transition distribution $T_s(\theta^s|\theta^{s-1})$ that depends on the previous draw

¹As defined in probability theory, a first-order Markov chain is a sequence of random variables for which for any t , the distribution of θ^t given all previous θ 's depends only on the most recent value, θ^{t-1} .

θ^{s-1} . The new draw, θ^s , is rejected or accepted depending on its effect on the posterior density. For models in this thesis, sampling is performed using the Stan software described in Appendix C.3.

Once simulation is complete, parameters can be summarised from the Markov chain draws of the posterior distribution. Let θ^s be a draw of parameter θ from the posterior distribution $p(\theta|y)$ where $s = 1, \dots, S$ is the iteration number. The estimated mean parameter value, $\hat{\theta}$, is calculated by Equation C.6. This is sometimes called the Empirical Bayes solution.

$$\hat{\theta} = \mathbb{E}(\theta|y) = \frac{1}{S} \sum_{s=1}^S \theta^s \quad (\text{C.6})$$

Predictive inference is performed using all the draws from the posterior. For inputs $\mathbf{X}_i$ and $\mathbf{Z}_i$ the predicted outcome value, $\hat{y}_{ij}$, is described by Equation C.7.

$$\hat{y}_i = \mathbb{E}(\hat{y}_i|\boldsymbol{\beta}, \mathbf{b}_i, \mathbf{D}) = \frac{1}{S} \sum_{s=1}^S \mathbf{X}_i \boldsymbol{\beta}^s + \mathbf{Z}_i \mathbf{b}_i^s \quad (\text{C.7})$$

Given a new data set, $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Z}}$, the posterior predictive distribution is found by integrating over the parameters, or summing over the sampled parameters. For example, at the population level of the multilevel model, the posterior predictive solution for $\tilde{\mathbf{y}}$ is described by Equation C.8. Here, θ is a general term for all the model parameters.

$$\begin{aligned} p(\tilde{\mathbf{y}}|\mathbf{y}, \theta) &= \int p(\tilde{\mathbf{y}}|\mathbf{y}, \theta) p(\theta|\mathbf{y}, \boldsymbol{\beta}) d\theta \\ &= \frac{1}{S} \sum_{s=1}^S p(\tilde{\mathbf{y}}|\mathbf{y}, \theta^s) \end{aligned} \quad (\text{C.8})$$

The predicted population mean and variance are therefore described by Equations C.9

and C.10 respectively.

$$\tilde{\mathbf{y}} = \mathbb{E}(\tilde{\mathbf{y}}|\mathbf{y}, \boldsymbol{\beta}, \mathbf{D}, \sigma) = \frac{1}{S} \sum_{s=1}^S \tilde{\mathbf{X}} \boldsymbol{\beta}^s \quad (\text{C.9})$$

$$\text{Var}(\tilde{\mathbf{y}}) = \text{Var}(\tilde{\mathbf{y}}|\mathbf{y}, \boldsymbol{\beta}, \mathbf{D}, \sigma) = \frac{1}{S} \sum_{s=1}^S \tilde{\mathbf{Z}} \mathbf{D}^s \tilde{\mathbf{Z}}^T + (\sigma^s)^2 \quad (\text{C.10})$$

Similarly, the modelled trajectory for the i^{th} participant, $\tilde{\mathbf{y}}_i$, is estimated from the posterior predictive distribution at the subject-level of the multilevel model. Equation C.11 describes the mean trajectory for subject i . The variance for all subject-specific trajectories is described by the within-subject variance, σ^2 .

$$\tilde{\mathbf{y}}_i = \mathbb{E}(\tilde{\mathbf{y}}_i|\mathbf{y}, \boldsymbol{\beta}, \mathbf{b}_i) = \frac{1}{S} \sum_{s=1}^S (\tilde{\mathbf{X}}_i \boldsymbol{\beta}^s + \tilde{\mathbf{Z}}_i \mathbf{b}_i^s) \quad (\text{C.11})$$

The likelihood of each subject-specific parameter, $\mathbf{b}_i$, can be found using the posterior probability density distribution, $p(\mathbf{b}_i|\mathbf{D})$, described in Equation C.12.

$$p(\mathbf{b}_i|\mathbf{D}) = \frac{1}{S} \sum_{s=1}^S \mathcal{N}(\mathbf{b}_i^s|0, \mathbf{D}^s) \quad (\text{C.12})$$

C.3 Stan software

STAN² is a relatively new modelling language (2014) that can be called from a number of interfaces such as the Command line (CmdStan), R (RStan), Python (PyStan) or MATLAB (MatlabStan). Stan was originally developed to apply full Bayesian inference to multilevel generalised linear models, when the insufficiency of other Bayesian modelling languages, such as BUGS (Bayesian inference Using Gibbs Sampling) and JAGS (Just Another Gibbs Sampler), became apparent. Stan automatically compiles the model in C++ code and solves by Hamiltonian Monte Carlo sampling, using the

²Sampling Through Adaptive Neighbourhoods (STAN), also named in honour of Stanislaw Ulam, pioneer of the Monte Carlo sampling method.

no-U-turn method [190, 192].

Common statistics software such as STATA and SAS, and R packages for multilevel modelling, use point estimates for parameters using Maximum Likelihood or Restricted Maximum Likelihood (REML). In addition, these packages allow only limited customisation of covariance structures. In Stan, priors can be defined for all parameters, and covariance structures are easily customised. Sorensen *et al.* [193, 194] provide comprehensive tutorials on writing multilevel models in Stan.

C.4 Assessing convergence

When numerical simulation methods are used to summarise parameters in a model, it is important to check the sampling chains have converged to a single distribution. If convergence has not been achieved, either the model is not appropriate for the data, or sampling needs to be performed again with more iterations. Parameters in the multilevel models described in Chapter 5 were estimated using four Markov chains of 2000 samples, where the first 500 samples were discarded as warm-up iterations (see Section 5.2.5). We use two measures to check model convergence: Markov chain stationarity and effective sample size. For more information on monitoring convergence, we recommend Chapter 11 of Gelman *et al.* [161].

For a time-series to be *stationary*, its distribution must not change over time. Thus for a series of sequential Markov chains to be stationary, all the samples within each chain, and between chains, must be drawn from the same distribution. A traceplot is an easy way to observe stationarity. Here all the Markov chain samples are plotted consecutively and they should overlap, as in the example in Figure C.1. However, there are usually a large number of parameters in a model, each with their own traceplot. We therefore also use the quantitative $\hat{R}$ measure to assess stationarity. This scale calculates the between- and within-chain variance, and declines to 1 as the number of samples in

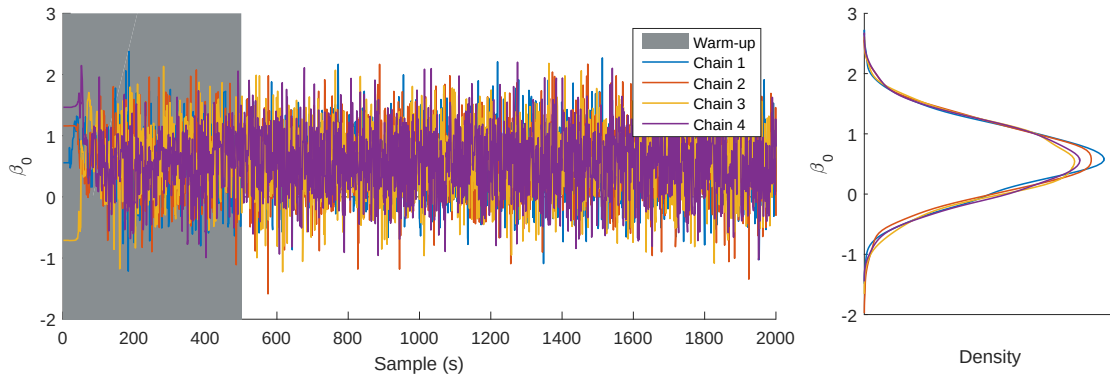


Figure C.1: An example of how stationarity can be visualised to assess convergence of a parameter. The traceplot (left) shows the value of each sample from the four Markov chains. The distributions of these values all overlap indicating convergence (right). This example is the β_0 parameter of antenatal SBP using a standard polynomial, with convergence parameters $\hat{R} = 1.0$ and effective sample size = 6000 (4 chains of 1500 samples once warm-up samples have been discarded).

each chain tends to infinity. We required $\hat{R} < 1.1$ for all parameters to consider a model converged.

Although a random process, Markov chains are sequential and therefore some correlation is present. *Effective sample size* is a measure of the equivalent number of independent samples in a Markov chain, and indicates the precision around the parameter estimate. This is calculated by estimating the correlation parameter from the between- and within-chain variance. The maximum number of independent samples possible is the total number of samples across all chains, in our case $4 \times 1500 = 6000$. We required the effective sample size to be less than 80 (the equivalent of twenty independent draws per chain) for all parameters, to consider a model converged.

C.5 WAIC Information Criterion

When choosing the best model, we want to estimate the predictive model accuracy, that describes how well a model fits to new data. Cross-validation is a natural solution to estimating predictive model accuracy, where a part of the training set is held out for validation, then the sets are rotated and the model re-fit several times. How-

ever, cross-validation is not suitable for the multilevel models described in Chapter 5, because it requires repeated model fits which are computationally intensive. Instead we use WAIC (Watanabe-Akaike or Widely Applicable Information Criterion) to assess predictive model accuracy [195].

WAIC estimates the expected log predictive density by correcting the observed log predictive density with a bias term. The bias term penalises for model complexity (number of parameters), and is estimated by the variance of the log likelihood. The formula for WAIC is presented in Equation C.13, where y_i is the i^{th} datapoint, and θ^s is the s^{th} draw of parameter θ from a sampling chain. All measures here are “pointwise” because integrals are replaced with samples drawn from the posterior distribution. WAIC is measured on the “deviance” scale, the same as other common information criteria such as AIC and BIC³. A lower WAIC value indicates a better predictive fit.

$$WAIC = -2 \left(\sum_{i=1}^n \log \left(\frac{1}{S} \sum_{s=1}^S p(y_i | \theta^s) \right) - \sum_{i=1}^n \text{Var}_s (\log p(y_i | \theta^s)) \right) \quad (\text{C.13})$$

WAIC closely approximates Bayesian cross-validation and is advantageous when comparing different models since it puts models with different parametrisations on a common scale.

C.6 Multilevel models for aggregate data

The multilevel model described so far in this supplement is for analysis of raw measurements, where the parameters are estimated entirely from the data. For application of the multilevel model on aggregate data (as in the synthesised dataset from the systematic review described in Section 2.3.2), we fix the within-group variance parameters to the variance of the mean that is reported in the study for that group. This weights the

³Akaike’s Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are useful for the comparison of singular models (where the parameters are non-nested, unlike the multilevel model for longitudinal data). These information criteria are of a similar form to WAIC, but differ in their approximation of the bias term.

contribution of each data point according to its variance, σ_{ij}^2 , and sample size, w_{ij} . The variance of the mean, SE^2 (standard error squared), is described by:

$$SE_{ij}^2 = \frac{\sigma_{ij}^2}{w_{ij}} \quad (\text{C.14})$$

The adapted within-study covariance matrix for group i , $\mathbf{R}_i$, is therefore:

$$\mathbf{R}_i = \begin{bmatrix} SE_{i1}^2 & & & \\ 0 & SE_{i2}^2 & & \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \cdots & SE_{in_i}^2 \end{bmatrix} \quad (\text{C.15})$$

This formulation of the multilevel model is similar to the model proposed by Ishak *et al.* [57]. However, in order to account for differences in sample size we use SE^2 instead of variance σ^2 . Each data point is therefore weighted by the reciprocal of the associated variance of the mean so that groups with a larger sample size are more influential.

C.7 Example of multilevel model development

In this section we illustrate the multilevel model development process using Systolic Blood Pressure (SBP) as an example. We first examine the model building process for a multilevel model with standard polynomials (model (a) described in Section 5.2.3). Then we compare the outputs of model (a), model (b), and model (c) for antenatal and postnatal SBP. Finally, we select the best model for each vital sign using WAIC information criterion.

C.7.1 Model building

The standard polynomial multilevel models (model (a)) were built using a bottom-up approach. For each model considered, convergence was checked using the WAIC criteria

described in Appendix C.4. The population polynomial order (P) and subject-specific polynomial order (Q) were decided by assessing model fit by visual inspection of centiles and the WAIC information criterion, described in Appendix C.5. We started with an intercept parameter for both the population and subject-specific levels ($P = Q = 0$), such that:

$$y_{ij} = \beta_0 + b_{0i} \quad (\text{C.16})$$

Population-level parameters

Population level parameters β were then added to the model. We varied the population polynomial order P from 0 (straight line) to 3 (cubic polynomial). At this stage the model is described by:

$$y_{ij} = \sum_{p=0}^P \beta_p x_{ij}^p + b_{0i}$$

Figure C.2 shows the 3rd, 50th and 97th centiles of these four models fitted to antenatal SBP. As additional population parameters were added to the model, the curvature of the mean trajectory increased to align with the mean empirical centiles. The curvature of the final two models ($P = 2$ and $P = 3$) were very similar. The WAIC values for these models were: 26036, 25908, 25789 and 25959 for $P = 0, 1, 2$ and 3 respectively. Model $P = 2$ was therefore selected as the best fit (lowest WAIC value).

Subject-specific parameters

After the population polynomial order was chosen ($P = 2$), subject-specific parameters were added to the model. We varied the subject-specific polynomial order Q from 0 (intercept only) to 2 (quadratic). The model is therefore described by:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2 + \sum_{q=0}^Q b_{qi} x_{ij}^q$$

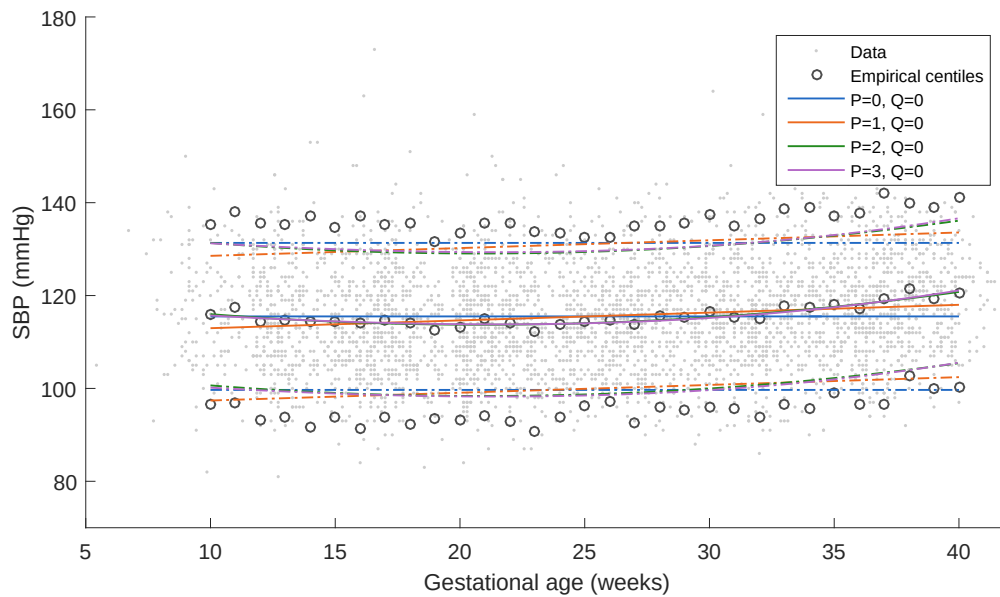


Figure C.2: Addition of population parameters to a standard polynomial multilevel model of antenatal SBP. The 50th centile is shown as a solid line, and outer centiles (3rd and 97th) as dashed lines.

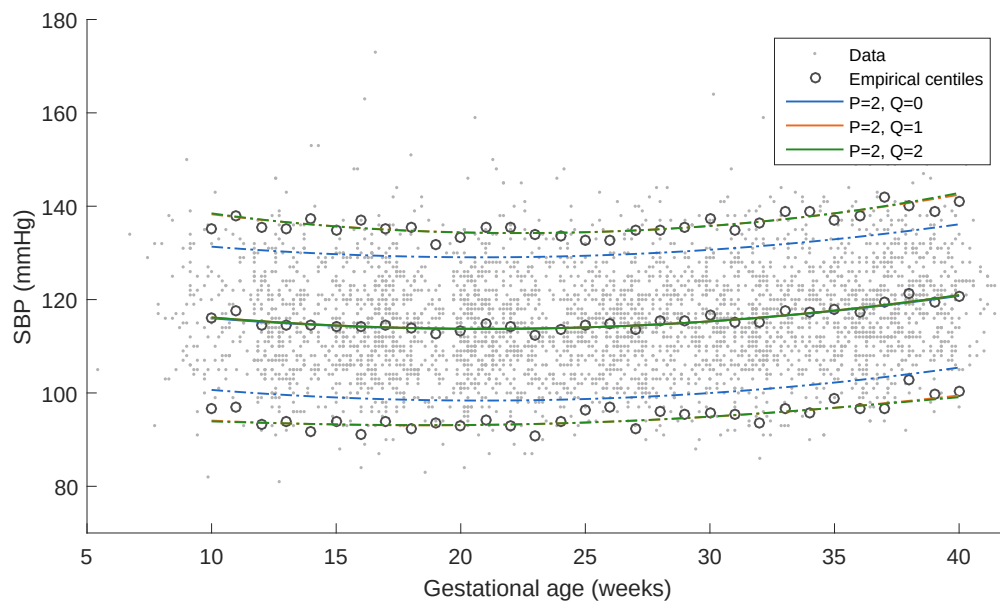


Figure C.3: Addition of subject-specific parameters to a standard polynomial multilevel model of antenatal SBP. The 50th centile is shown as a solid line, and outer centiles (3rd and 97th) as dashed lines. Note that the orange and green trajectories ($Q = 1$ and $Q = 2$ respectively) overlap.

Figure C.3 shows the 3rd, 50th and 97th centiles of these models fitted to antenatal SBP. Subject-specific parameters contribute to the variance of the population model, so the difference between these models is observed in the outer centiles. Note how the change from $Q = 0$ to $Q = 1$ increased the difference between the 3rd and 97th centiles to bring them in line with the empirical centiles.

Little difference was observed between $Q = 1$ and $Q = 2$. Investigation of a few individual trajectories also suggested that there was not much difference between these two models (for example, see Figure C.4). The WAIC values were 25789, 25639 and 25624 for $Q = 0, 1$ and 2 respectively. The difference between the latter two models is small (15 units) which confirmed that these models fit similarly to the data, as observed from the visual analysis. We choose the lower order $Q = 1$ as the final subject-specific polynomial order.

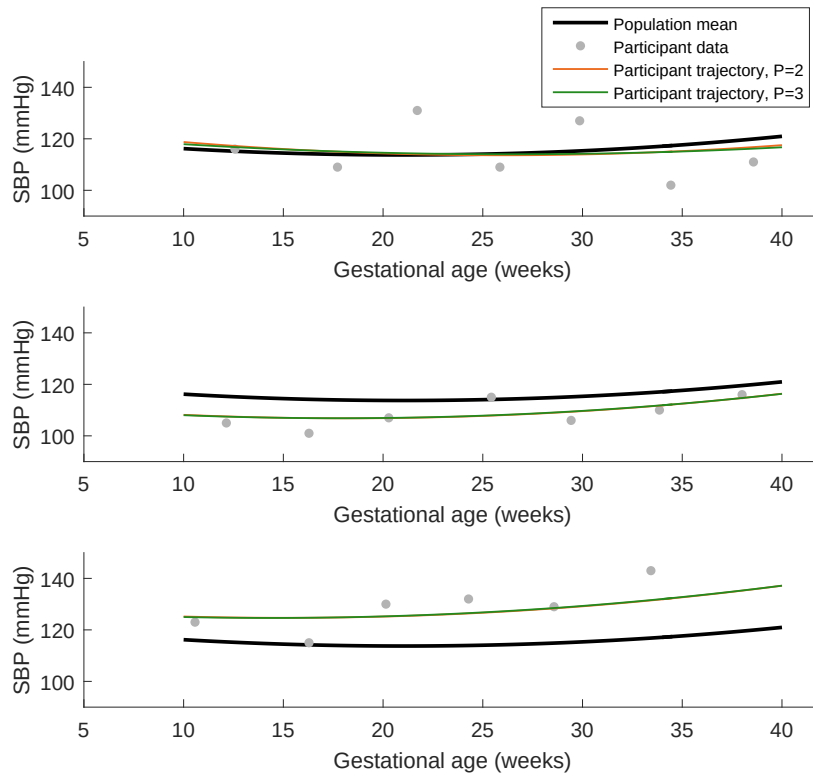


Figure C.4: Individual trajectories inferred from a standard polynomial multilevel model. Mean trajectories from subject-specific linear and quadratic models are presented ($Q = 1$ and $Q = 2$ respectively). Note these trajectories (orange and green) largely overlap.

The final multilevel model to describe the antenatal SBP trend therefore has $P = 2$ and $Q = 1$, described by:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2 + b_{0i} + b_{1i} x_{ij} \quad (\text{C.17})$$

C.7.2 Model comparison

To compare models (a), (b), and (c), we continue to use the example of SBP.

Antenatal SBP

Figure C.5 shows the fitted antenatal models. The standard polynomial order chosen for model (a) was quadratic in the population parameters ($P = 2$) and linear in the subject-

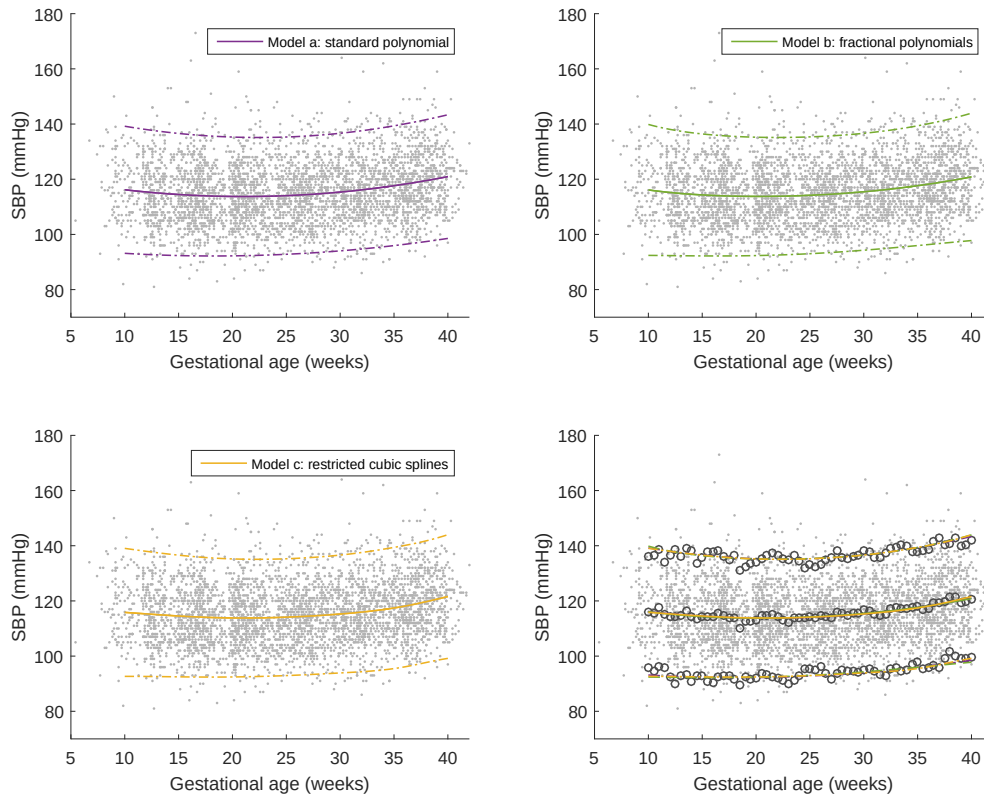


Figure C.5: Models (a), (b), and (c) applied to antenatal SBP measurements from the 4P study. The fourth panel (bottom right) shows all models with the 3rd, 50th and 97th empirical centiles for each week of gestation (there is little difference between the three models, so only the yellow line from model (c) is visible).

specific parameters ($Q = 1$). The fractional polynomial powers selected for model (b) were $p_1 = -0.5$ and $p_2 = 3$. All three models produce a similar mean trajectory and fit well to the empirical centiles. The maximum difference in the mean trajectory is 0.73 mmHg between model (b) and (c), at 40 weeks of gestation. This is a small difference given the resolution of an automatic blood pressure cuff is said to be 5 mmHg [150]. The WAIC information criterion, which assess model fit, indicates model (c) has the best fit, although the results are very similar for all three models (WAIC values were 25639, 25636, and 25597 respectively, where a lower number indicates a better fit). Overall, any of model (a), (b), or (c) would be appropriate to represent the changes in SBP observed in the antenatal period.

Postnatal SBP

The models fitted to the postnatal data are shown in Figure C.6. Only models (a) and (b) were considered for postnatal trends. The standard polynomial order chosen for model (a) was cubic in the population parameters ($P = 3$) and quadratic in the subject-specific parameters ($Q = 2$). The fractional polynomial powers selected for model (b) were repeated powers, $p_1 = p_2 = 0.5$, therefore the second polynomial term was multiplied by $\ln(\mathbf{x}_i)$. The final equation to describe the fractional polynomial model for postnatal SBP is:

$$y_{ij} = \beta_0 + \beta_1 x_{ij}^{0.5} + \beta_2 x_{ij}^{0.5} \log(x_{ij}) + b_{i0} + b_{i1} x_{ij}^{0.5} + b_{i2} x_{ij}^{0.5} \ln(x_{ij}) \quad (\text{C.18})$$

The trajectories for postnatal SBP model (a) and (b) most obviously differs in two places: between days 0 and 1, and between days 10 and 16 after delivery. In the first day after delivery, the mean trajectory of the standard polynomial model (a) gives a gradual increase from 112 to 115.4 mmHg. In contrast, the fractional polynomial (b) rises sharply from 107.5 to 116.3 mmHg and immediately after delivery the outer centiles

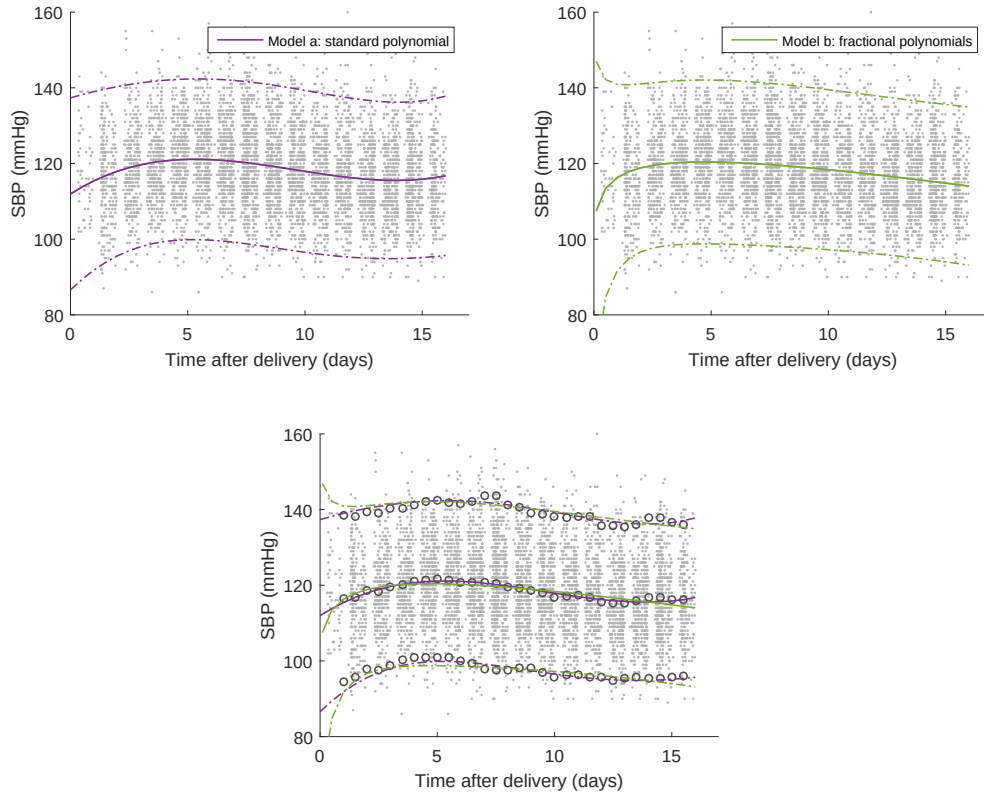


Figure C.6: Models (a) and (b) applied to postnatal SBP measurements from the 4P study. The bottom panel shows both model (a) and (b) with the 3rd, 50th and 97th empirical centiles for each day after delivery.

are widely spread. This is most likely an artefact which would indicate that the fractional polynomial model has not fitted properly to the data in this range, possibly due to the small number of SBP measurements recorded on day 0 ($n = 57$) compared to days 1-15 (minimum of $n = 241$ measurements).

Secondly, after day 10, the mean trajectory from model (a) flattens and then curves upwards from a local minimum of 115.5 mmHg at 14 days after delivery; whereas the mean trajectory of model (b) continues in a downward trend, dropping approximately 0.7 mmHg per day. The empirical centiles suggest that SBP flattens out, similar to the trajectory of model (a), although it is difficult to determine which is more appropriate without knowing what happens after day 15. The WAIC information criterion indicates model (a) fits to the data marginally better than model (b) (WAIC values were 41135

and 41210 respectively).

Sensitivity analysis for postnatal SBP

To further investigate why there is a difference between the mean trajectories of models (a) and (b) for postnatal SBP, we perform a sensitivity analysis. We remove the data from day 0 that may be causing an artefact in the model fit, and re-fit model (a) and (b) to the new data set. Figure C.7 shows the comparison between the models fitted to the original data set, and the models fitted to the data with day 0 data removed. The models fit similarly to the two sets of data. The maximum difference between the mean trajectory of model (a) is 1.2 mmHg at day 0, and for model (b) a maximum difference of 1.4 mmHg at day 0. This indicates that the artefact observed in the fractional polynomial model (b) is not due to artefactual data and indicates that this model is not appropriate to represent the data.

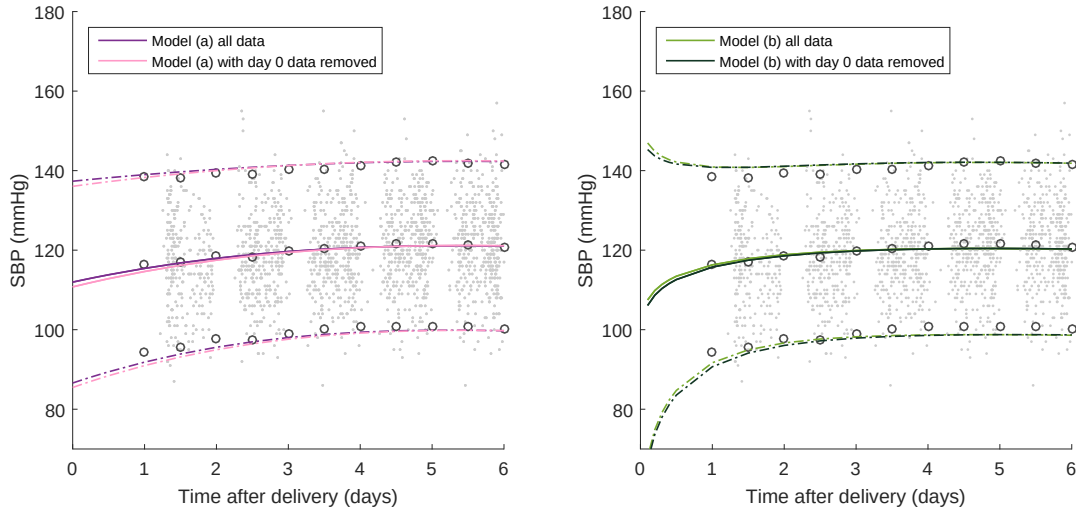


Figure C.7: Sensitivity analysis for model (a) (left) and model (b) (right) postnatal SBP where the data from day zero has been removed. The dashed lines show the 3rd, 50th and 97th empirical centiles for each day after delivery. Only the trajectory from day 0 to 6 is displayed, as the model fit for days 6 to 16 is the same for both models, as in Figure C.6.

C.7.3 Model selection

The model orders chosen for model (a), fractional polynomial powers for model (b), and the WAIC scores for all models, are presented in Table C.1.

	(a) Standard polynomial			(b) Fractional polynomials			(c) Cubic splines
	P	Q	WAIC_a	p_1	p_2	WAIC_b	WAIC_c
<i>Antenatal</i>							
SBP	2	1	25639	-0.5	3	25636	25597
DBP	2	1	23212	0.5	2	23213	23182
HR	2	2	23909	3	3	23889	-
Temp	1	1	3697	-0.5	-	3686	3630
SpO ₂	1	1	11762	1	-	11759	11742
RR	1	1	18151	1	-	18149	18150
<i>Postnatal</i>							
SBP	3	2	41135	0.5	0.5	41210	-
DBP	3	2	37812	0.5	0.5	37910	-
HR	3	2	39960	-1	-1	40210	-
Temp	1	1	6323	1	-	6319	-
SpO ₂	1	1	17902	0.5	0.5	17805	-
RR	1	1	4243	1	-	4224	-

Table C.1: Information about models (a) to (c) fitted to univariate (each vital sign independently) antenatal and postnatal data collected in the 4P study. For model (a), the selected polynomial model orders for the population level (P) and subject-specific level (Q) are presented (see Equation 5.5). For model (b), the selected fractional powers (p_1 and p_2) are presented (see Equation 5.6). WAIC is an assessment of model fit where a lower WAIC value indicates a better fit (Appendix C.5).

Models (a) and (b) were fitted to 4P antenatal and postnatal measurements of each vital sign individually (SBP, DBP, HR, Temp, SpO₂ and RR). Model (c) was fitted to antenatal measurements only. It should be noted that the standard polynomial orders (model (a)) chosen for antenatal SBP and DBP were the same as in previous studies [124, 71, 145].

For the antenatal trends, the cubic spline model (c) fitted the SBP and DBP data slightly better than models (a) or (b), however did it not converge for HR. For SpO₂ and RR, the fractional polynomial powers chosen for model (b) were the same as the

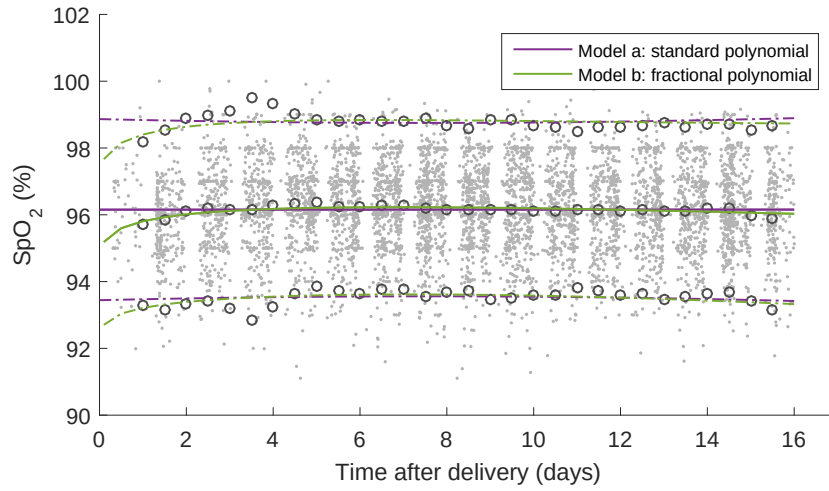


Figure C.8: Models (a) and (b) applied to postnatal SpO_2 measurements from the 4P study. The 3rd, 50th and 97th centiles are shown (solid and dashed lines), along with empirical centiles (round circles) for each day after delivery.

polynomial orders for model (a), and thus the same model was fitted. Overall, models (a) and (b) had very similar fits to the antenatal data.

For the postnatal data, model (a) fitted better than model (b) for SBP, DBP and HR. As discussed in the previous section, for postnatal SBP, the fractional polynomial model (b) appeared to produce an artefact between day 0 and 1. A similar artefact was present for DBP and HR which means model (b) is not appropriate to describe the trajectories of these vital signs. For SpO_2 , again there was a difference between model (a) and (b) between day 0 and 1, however the WAIC information criterion indicated model (b) had the better fit. The trajectories for SpO_2 are shown in Figure C.8. Here the mean trajectory of the standard polynomial model (a) was approximately constant at 96.2% ($P = 1$, linear polynomial slope coefficient $\beta_1 < 0.0001$), whereas the fractional polynomial model (b) started 1% lower at 95.2% and curved upwards to meet model (a) at day 3 after delivery. The rest of the trajectory remained the same. As with the SBP example discussed in the previous section (C.7.2), it is difficult to know whether this represents the true physiological state since there is few data in the 24 hours after delivery. For postnatal Temp and RR, the fractional polynomial powers chosen for model

(b) were the same as the polynomial orders for model (a), and thus the same model was fitted. Overall, model (a) performed better than model (b) for the postnatal data, except for SpO₂ where model (b) gave a better fit.

We selected the standard polynomial model (a) for final analyses. This model performed better than the fractional polynomial model (b) on the postnatal data. Although the cubic spline model (c) performed better than both model (a) and (b) for antenatal SBP and DBP data, model fit was similar from all models and we did not pursue model (c) for the remaining vital signs or the postnatal period.

C.8 Equations for final univariate vital sign trends

The equations to describe the standard polynomial multilevel model for each vital sign are presented in tables C.2 (antenatal models) and C.3 (postnatal models). We present the estimated mean of the variance parameters, $\hat{\sigma}^2$ and $\hat{\mathbf{D}}$ (Equation C.6), and the predicted population mean and variance, $\tilde{\mathbf{y}}$ and $\text{Var}(\tilde{\mathbf{y}})$ (Equations C.9 and C.10). Note that for antenatal models, time is measured in weeks of gestation starting at 10 weeks, $\tilde{\mathbf{t}} = \tilde{\mathbf{x}} - 10$.

Systolic blood pressure (mmHg),
$\hat{\sigma}^2 = 53.29, \hat{\mathbf{D}} = \{86.00, -1.20; -1.20, 0.08\}$
$\tilde{\mathbf{y}} = \beta_0 + \beta_1 \tilde{\mathbf{t}} + \beta_2 \tilde{\mathbf{t}}^2 = 116.18 - 0.44\tilde{\mathbf{t}} + 0.02\tilde{\mathbf{t}}^2$
$\text{Var}(\tilde{\mathbf{y}}) = [1, \tilde{\mathbf{t}}] \hat{\mathbf{D}} [1, \tilde{\mathbf{t}}]^T + \hat{\sigma}^2 = 86 - 2 \times 1.2\tilde{\mathbf{t}} + 0.08\tilde{\mathbf{t}}^2 + 53.29$

Diastolic blood pressure (mmHg),
$\hat{\sigma}^2 = 27.04, \hat{\mathbf{D}} = \{45.00, -0.40; -0.40, 0.03\}$
$\tilde{\mathbf{y}} = \beta_0 + \beta_1 \tilde{\mathbf{t}} + \beta_2 \tilde{\mathbf{t}}^2 = 71.77 - 0.45\tilde{\mathbf{t}} + 0.02\tilde{\mathbf{t}}^2$
$\text{Var}(\tilde{\mathbf{y}}) = [1, \tilde{\mathbf{t}}] \hat{\mathbf{D}} [1, \tilde{\mathbf{t}}]^T + \hat{\sigma}^2 = 45 - 2 \times 0.4\tilde{\mathbf{t}} + 0.03\tilde{\mathbf{t}}^2 + 27.04$

Heart rate (bpm),
$\hat{\sigma}^2 = 74.61, \hat{\mathbf{D}} = \{70.00, 0.08, -0.007; 0.08, 0.13, -0.005; -0.007, -0.005, < 0.001\}$
$\tilde{\mathbf{y}} = \beta_0 + \beta_1 \tilde{\mathbf{t}} + \beta_2 \tilde{\mathbf{t}}^2 = 82.20 + 0.71\tilde{\mathbf{t}} - 0.01\tilde{\mathbf{t}}^2$
$\text{Var}(\tilde{\mathbf{y}}) = [1, \tilde{\mathbf{t}}, \tilde{\mathbf{t}}^2] \hat{\mathbf{D}} [1, \tilde{\mathbf{t}}, \tilde{\mathbf{t}}^2]^T + \hat{\sigma}^2$
$= 70 - 2 \times 0.08\tilde{\mathbf{t}} + (2 \times -0.007 + 0.13)\tilde{\mathbf{t}}^2 - 2 \times 0.005\tilde{\mathbf{t}}^3 + 0.001\tilde{\mathbf{t}}^4 + 74.61$

Temperature (°C),
$\hat{\sigma}^2 = 0.002, \hat{\mathbf{D}} = \{0.10, < 0.001; < 0.001, < 0.001\}$
$\tilde{\mathbf{y}} = \beta_0 + \beta_1 \tilde{\mathbf{t}} = 36.63 - 0.009\tilde{\mathbf{t}}$
$\text{Var}(\tilde{\mathbf{y}}) = [1, \tilde{\mathbf{t}}] \hat{\mathbf{D}} [1, \tilde{\mathbf{t}}]^T + \hat{\sigma}^2 = 0.1 - 2 \times 0.001\tilde{\mathbf{t}} + 0.001\tilde{\mathbf{t}}^2 + 0.002$

Oxygen saturation (%),
$\hat{\sigma}^2 = 1.44, \hat{\mathbf{D}} = \{0.60, -0.003; -0.003, < 0.001\}$
$\tilde{\mathbf{y}} = \beta_0 + \beta_1 \tilde{\mathbf{t}} = 97.58 - 0.04\tilde{\mathbf{t}}$
$\text{Var}(\tilde{\mathbf{y}}) = [1, \tilde{\mathbf{t}}] \hat{\mathbf{D}} [1, \tilde{\mathbf{t}}]^T + \hat{\sigma}^2 = 86 - 2 \times 1.2\tilde{\mathbf{t}} + 0.08\tilde{\mathbf{t}}^2 + 1.44$

Respiratory rate (rpm),
$\hat{\sigma}^2 = 8.41, \hat{\mathbf{D}} = \{10.00, -0.09; -0.09, 0.006\}$
$\tilde{\mathbf{y}} = \beta_0 + \beta_1 \tilde{\mathbf{t}} = 15.64 - 0.002\tilde{\mathbf{t}}$
$\text{Var}(\tilde{\mathbf{y}}) = [1, \tilde{\mathbf{t}}] \hat{\mathbf{D}} [1, \tilde{\mathbf{t}}]^T + \hat{\sigma}^2 = 10 - 2 \times 0.09\tilde{\mathbf{t}} + 0.006\tilde{\mathbf{t}}^2 + 8.41$

Table C.2: Equations for vital sign trajectories in the antenatal period. These describe the antenatal trajectories presented in Figure 5.3 and 5.4. $\tilde{\mathbf{t}} = \tilde{\mathbf{x}} - 10$ measured in weeks of gestation.

Systolic blood pressure (mmHg),

$$\hat{\sigma}^2 = 49, \hat{\mathbf{D}} = \{120, -14, 0.60; -14, 4.20, -0.21; 0.60, -0.21, 0.01\}$$

$$\tilde{\mathbf{y}} = \beta_0 + \beta_1 \tilde{\mathbf{t}} + \beta_2 \tilde{\mathbf{t}}^2 + \beta_3 \tilde{\mathbf{t}}^3 = 111.91 + 3.9\tilde{\mathbf{t}} - 0.51\tilde{\mathbf{t}}^2 + 0.02\tilde{\mathbf{t}}^3$$

$$\text{Var}(\tilde{\mathbf{y}}) = [1, \tilde{\mathbf{t}}, \tilde{\mathbf{t}}^2] \hat{\mathbf{D}} [1, \tilde{\mathbf{t}}, \tilde{\mathbf{t}}^2]^T + \hat{\sigma}^2$$

$$= 120 - 2 \times 14\tilde{\mathbf{t}} + (2 \times -0.60 + 4.20)\tilde{\mathbf{t}}^2 - 2 \times 0.21\tilde{\mathbf{t}}^3 + 0.01\tilde{\mathbf{t}}^4 + 49$$

Diastolic blood pressure (mmHg),

$$\hat{\sigma}^2 = 28, \hat{\mathbf{D}} = \{60, -5, 0.18; -5, 1.40, -0.07; 0.18, -0.07, 0.004\}$$

$$\tilde{\mathbf{y}} = \beta_0 + \beta_1 \tilde{\mathbf{t}} + \beta_2 \tilde{\mathbf{t}}^2 + \beta_3 \tilde{\mathbf{t}}^3 = 71.08 + 3.1\tilde{\mathbf{t}} - 0.39\tilde{\mathbf{t}}^2 + 0.01\tilde{\mathbf{t}}^3$$

$$\text{Var}(\tilde{\mathbf{y}}) = [1, \tilde{\mathbf{t}}, \tilde{\mathbf{t}}^2] \hat{\mathbf{D}} [1, \tilde{\mathbf{t}}, \tilde{\mathbf{t}}^2]^T + \hat{\sigma}^2$$

$$= 60 - 2 \times 5\tilde{\mathbf{t}} + (2 \times -0.18 + 1.40)\tilde{\mathbf{t}}^2 - 2 \times 0.07\tilde{\mathbf{t}}^3 + 0.004\tilde{\mathbf{t}}^4 + 28$$

Heart rate (bpm),

$$\hat{\sigma}^2 = 57.80, \hat{\mathbf{D}} = \{160, -21, 0.89; -21, 6.60, -0.34; 0.89, -0.34, 0.02\}$$

$$\tilde{\mathbf{y}} = \beta_0 + \beta_1 \tilde{\mathbf{t}} + \beta_2 \tilde{\mathbf{t}}^2 + \beta_3 \tilde{\mathbf{t}}^3 = 88.07 + 3.1\tilde{\mathbf{t}} - 0.28\tilde{\mathbf{t}}^2 - 0.008\tilde{\mathbf{t}}^3$$

$$\text{Var}(\tilde{\mathbf{y}}) = [1, \tilde{\mathbf{t}}, \tilde{\mathbf{t}}^2] \hat{\mathbf{D}} [1, \tilde{\mathbf{t}}, \tilde{\mathbf{t}}^2]^T + \hat{\sigma}^2$$

$$= 160 - 2 \times 21\tilde{\mathbf{t}} + (2 \times 0.89 + 6.60)\tilde{\mathbf{t}}^2 - 2 \times 0.34\tilde{\mathbf{t}}^3 + 0.02\tilde{\mathbf{t}}^4 + 57.80$$

Temperature (°C),

$$\hat{\sigma}^2 = 0.002, \hat{\mathbf{D}} = \{0.16, -0.002; -0.002, < 0.001\}$$

$$\tilde{\mathbf{y}} = \beta_0 + \beta_1 \tilde{\mathbf{t}} = 36.68 - 0.004\tilde{\mathbf{t}}$$

$$\text{Var}(\tilde{\mathbf{y}}) = [1, \tilde{\mathbf{t}}] \hat{\mathbf{D}} [1, \tilde{\mathbf{t}}]^T + \hat{\sigma}^2 = 0.16 - 2 \times 0.002\tilde{\mathbf{t}} + 0.001\tilde{\mathbf{t}}^2 + 0.002$$

Oxygen saturation (%),

$$\hat{\sigma}^2 = 1.21, \hat{\mathbf{D}} = \{0.67, -0.02; -0.02, 0.003\}$$

$$\tilde{\mathbf{y}} = \beta_0 + \beta_1 \tilde{\mathbf{t}} = 96.15 + 0.001\tilde{\mathbf{t}}$$

$$\text{Var}(\tilde{\mathbf{y}}) = [1, \tilde{\mathbf{t}}] \hat{\mathbf{D}} [1, \tilde{\mathbf{t}}]^T + \hat{\sigma}^2 = 0.67 - 2 \times 0.02\tilde{\mathbf{t}} + 0.003\tilde{\mathbf{t}}^2 + 1.21$$

Respiratory rate (rpm),

$$\hat{\sigma}^2 = 8.41, \hat{\mathbf{D}} = \{10.00, -0.09; -0.09, 0.006\}$$

$$\tilde{\mathbf{y}} = \beta_0 + \beta_1 \tilde{\mathbf{t}} = 16.34 - 0.002\tilde{\mathbf{t}}$$

$$\text{Var}(\tilde{\mathbf{y}}) = [1, \tilde{\mathbf{t}}] \hat{\mathbf{D}} [1, \tilde{\mathbf{t}}]^T + \hat{\sigma}^2 = 10 - 2 \times 0.09\tilde{\mathbf{t}} + 0.006\tilde{\mathbf{t}}^2 + 8.41$$

Table C.3: Equations for vital sign trajectories in the postnatal period. These describe the postnatal trajectories presented in Figure 5.3 and 5.4. $\tilde{\mathbf{t}} = \tilde{\mathbf{x}}$ measured in days after delivery.

References

- [1] C. Ronsmans and W. J. Graham, “Maternal mortality: who, when, where, and why,” *The Lancet*, vol. 368, no. 9542, pp. 1189–1200, 2006.
- [2] United Nations, “The millennium development goals report,” tech. rep., 2015.
- [3] M. C. Hogan, K. J. Foreman, M. Naghavi, S. Y. Ahn, M. Wang, S. M. Makela, A. D. Lopez, R. Lozano, and C. J. Murray, “Maternal mortality for 181 countries, 1980–2008: A systematic analysis of progress towards Millennium Development Goal 5,” *The Lancet*, vol. 375, no. 9726, pp. 1609–1623, 2010.
- [4] M. Koblinsky, M. E. Chowdhury, A. Moran, and C. Ronsmans, “Maternal morbidity and disability and their consequences: neglected agenda in maternal health,” *Journal of Health, Population and Nutrition*, pp. 124–130, 2012.
- [5] 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.
- [6] M. Knight, S. Kenyon, P. Brocklehurst, J. Neilson, J. Shakespeare, and J. J. Kurinczuk, “Saving lives, improving mothers’ care: Lessons learned to inform future maternity care from the UK and Ireland Confidential Enquiries into Maternal Deaths and Morbidity 2009–2012,” tech. rep., MBRRACE-UK, 2014.
- [7] NICE, “CG50: Acutely ill patients in hospital,” tech. rep., National Institute for Health and Care Excellence (NICE), 2007.
- [8] P. B. Lovett, J. M. Buchwald, K. Stürmann, 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.
- [9] “Maternity services in England,” tech. rep., National Audit Office, 2013.
- [10] K. Butt, K. Lim, S. Bly, Y. Cargill, G. Davies, N. Denis, G. Hazlitt, L. Morin, A. Ouellet, and S. Salem, “Determination of gestational age by ultrasound,” *Journal of Obstetrics and Gynaecology Canada*, vol. 36, no. 2, pp. 171–181, 2014.
- [11] NICE, “CG62: Antenatal care for uncomplicated pregnancies,” tech. rep., National Institute for Health and Care Excellence (NICE), 2008.
- [12] D. Walsh, “A review of evidence around postnatal care and breastfeeding,” *Obstetrics, Gynaecology & Reproductive Medicine*, vol. 21, no. 12, pp. 346–350, 2011.
- [13] NICE, “CG37: routine postnatal care of women and their babies,” tech. rep., National Institute for Health and Care Excellence (NICE), 2006.
- [14] M. Knight, M. Nair, D. Tuffnell, S. Kenyon, J. Shakespeare, P. Brocklehurst, and J. J. Kurinczuk, “Saving lives, improving mothers’ care: Surveillance of maternal deaths in the UK 2012–14 and lessons learned to inform maternity care from the

- UK and Ireland Confidential Enquiries into Maternal Deaths and Morbidity 2009–14,” tech. rep., MBRRACE-UK, 2016.
- [15] W. Ngan Kee, “Confidential enquiries into maternal deaths: 50 years of closing the loop,” *British Journal of Anaesthesia*, vol. 94, no. 4, pp. 413–416, 2005.
- [16] “Births in England and Wales: 2015,” tech. rep., Office for National Statistics, 2015.
- [17] D. Chou, Ö. Tunçalp, T. Firoz, M. Barreix, V. Filippi, P. von Dadelszen, N. van den Broek, J. G. Cecatti, and L. Say, “Constructing maternal morbidity—towards a standard tool to measure and monitor maternal health beyond mortality,” *BMC Pregnancy and Childbirth*, vol. 16, no. 45, pp. 1–10, 2016.
- [18] M. W. O’hara and J. E. McCabe, “Postpartum depression: current status and future directions,” *Annual Review of Clinical Psychology*, vol. 9, pp. 379–407, 2013.
- [19] H. L. Hedriana, S. Wiesner, B. G. Downs, B. Pelletreau, and L. E. Shields, “Baseline assessment of a hospital-specific early warning trigger system for reducing maternal morbidity,” *International Journal of Gynecology & Obstetrics*, vol. 132, no. 3, pp. 337–341, 2016.
- [20] S. Singh, A. McGlennan, A. England, and R. Simons, “A validation study of the CEMACH recommended modified early obstetric warning system (MEOWS),” *Anaesthesia*, vol. 67, no. 1, pp. 12–18, 2012.
- [21] D. A. Harrison, J. A. Penny, S. M. Yentis, S. Fayek, and A. R. Brady, “Case mix, outcome and activity for obstetric admissions to adult, general critical care units: a secondary analysis of the icnarc case mix programme database,” *Critical Care*, vol. 9, no. 3, p. 1, 2005.
- [22] C. AbouZahr, “Global burden of maternal death and disability,” *British Medical Bulletin*, vol. 67, no. 1, pp. 1–11, 2003.
- [23] C. D. Acosta, J. J. Kurinczuk, D. N. Lucas, D. J. Tuffnell, S. Sellers, M. Knight, *et al.*, “Severe maternal sepsis in the UK, 2011–2012: A national case-control study,” *PLoS Medicine*, vol. 11, no. 7, p. e1001672, 2014.
- [24] R. C. Bone, R. A. Balk, F. B. Cerra, R. P. Dellinger, A. M. Fein, W. A. Knaus, R. M. Schein, and W. J. Sibbald, “Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis,” *Chest*, vol. 101, no. 6, pp. 1644–1655, 1992.
- [25] L. Say, D. Chou, A. Gemmill, Ö. Tunçalp, A.-B. Moller, J. Daniels, A. M. Gülmezoglu, M. Temmerman, and L. Alkema, “Global causes of maternal death: a WHO systematic analysis,” *The Lancet Global Health*, vol. 2, no. 6, pp. e323–e333, 2014.
- [26] I. Al-Zirqi, S. Vangen, L. Forsen, and B. Stray-Pedersen, “Prevalence and risk factors of severe obstetric haemorrhage,” *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 115, no. 10, pp. 1265–1272, 2008.
- [27] W.-H. Zhang, S. Alexander, M.-H. Bouvier-Colle, and A. Macfarlane, “Incidence of severe pre-eclampsia, postpartum haemorrhage and sepsis as a surrogate marker for severe maternal morbidity in a european population-based study: the MOMS-B survey,” *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 112, no. 1, pp. 89–96, 2005.
- [28] “Prevention and management of postpartum haemorrhage,” tech. rep., Royal College of Obstetricians and Gynaecologists (RCOG), 2009.

- [29] R. Kerr, L. O. Eckert, B. Winikoff, J. Durocher, S. Meher, S. Fawcus, S. Mundle, B. Mol, S. Arulkumaran, K. Khan, *et al.*, “Postpartum haemorrhage: Case definition and guidelines for data collection, analysis, and presentation of immunization safety data,” *Vaccine*, vol. 34, no. 49, pp. 6102–6109, 2016.
- [30] J. L. Atterbury, L. J. Groome, and C. Hoff, “Blood pressure changes in normotensive women readmitted in the postpartum period with severe preeclampsia/eclampsia,” *Journal of Maternal-Fetal Medicine*, vol. 5, no. 4, pp. 201–205, 1996.
- [31] NICE, “NG107: Hypertension in pregnancy, diagnosis and management,” tech. rep., National Institute for Health and Care Excellence (NICE), 2010.
- [32] L. Duley, “The global impact of pre-eclampsia and eclampsia,” in *Seminars in Perinatology*, vol. 33, pp. 130–137, Elsevier, 2009.
- [33] M. Lazdam, A. de la Horra, J. Diesch, Y. Kenworthy, E. Davis, A. J. Lewandowski, C. Szmigielski, A. Shore, L. Mackillop, R. Kharbanda, *et al.*, “Unique blood pressure characteristics in mother and offspring after early onset preeclampsia,” *Hypertension*, vol. 60, no. 5, pp. 1338–1345, 2012.
- [34] D. Goldhill, S. 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.
- [35] D. Goldhill and A. McNarry, “Physiological abnormalities in early warning scores are related to mortality in adult inpatients,” *British Journal of Anaesthesia*, vol. 92, no. 6, pp. 882–884, 2004.
- [36] 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: CEMACH, 2007.
- [37] R. Morgan and M. Wright, “In defence of early warning scores,” *British Journal of Anaesthesia*, vol. 99, no. 5, pp. 747–748, 2007.
- [38] H. Gao, A. McDonnell, D. A. Harrison, T. Moore, S. Adam, K. Daly, L. Esmonde, D. R. Goldhill, G. J. Parry, A. Rashidian, *et al.*, “Systematic review and evaluation of physiological track and trigger warning systems for identifying at-risk patients on the ward,” *Intensive Care Medicine*, vol. 33, no. 4, pp. 667–679, 2007.
- [39] G. B. Smith, D. R. Prytherch, P. E. Schmidt, P. I. Featherstone, and B. Higgins, “A review, and performance evaluation, of single-parameter “track and trigger” systems,” *Resuscitation*, vol. 79, no. 1, pp. 11–21, 2008.
- [40] 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.
- [41] C. S. Parshuram, J. Hutchison, and K. Middaugh, “Development and initial validation of the bedside paediatric early warning system score,” *Critical Care*, vol. 13, no. 4, p. 1, 2009.
- [42] 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.
- [43] J. A. Hanley and B. J. McNeil, “The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve,” *Radiology*, vol. 143, no. 1, pp. 29–36, 1982. PMID: 7063747.

- [44] “National Early Warning Score (NEWS): Standardising the assessment of acute illness severity in the NHS,” tech. rep., Royal College of Physicians, 2012.
- [45] D. R. Prytherch, G. B. Smith, P. E. Schmidt, and P. I. Featherstone, “ViEWStowards a national early warning score for detecting adult inpatient deterioration,” *Resuscitation*, vol. 81, no. 8, pp. 932–937, 2010.
- [46] D. J. Silcock, A. R. Corfield, P. A. Gowens, and K. D. Rooney, “Validation of the National Early Warning score in the prehospital setting,” *Resuscitation*, vol. 89, pp. 31–35, 2015.
- [47] N. Alam, I. Vegting, E. Houben, B. van Berkel, L. Vaughan, M. Kramer, and P. Nanayakkara, “Exploring the performance of the National Early Warning Score (NEWS) in a European emergency department,” *Resuscitation*, vol. 90, pp. 111–115, 2015.
- [48] A. C. Quinn, T. Meek, and C. Waldmann, “Obstetric early warning systems to prevent bad outcome,” *Current Opinion in Anesthesiology*, vol. 29, no. 3, pp. 268–272, 2016.
- [49] A. Ur and M. Gordon, “Origin of korotkoff sounds,” *American Journal of Physiology*, vol. 218, no. 2, pp. 524–529, 1970.
- [50] L. T. Hersh, B. Friedman, and R. Medero, “Method for oscillometric blood pressure determination employing curve fitting,” 1998. US Patent 5,704,362.
- [51] J. E. Sinex, “Pulse oximetry: Principles and limitations,” *The American Journal of Emergency Medicine*, vol. 17, no. 1, pp. 59–66, 1999.
- [52] D. Dowding, S. Freeman, S. Nimmo, D. Smith, and M. Wisniewski, “An investigation into the accuracy of different types of thermometers,” *Journal of Professional Nursing*, vol. 18, no. 3, pp. 166–8, 2002.
- [53] C. Subbe, R. Davies, E. Williams, P. Rutherford, and L. Gemmell, “Effect of introducing the Modified Early Warning Score on clinical outcomes, cardio-pulmonary arrests and intensive care utilisation in acute medical admissions,” *Anaesthesia*, vol. 58, no. 8, pp. 797–802, 2003.
- [54] M. Folke, L. Cernerud, M. Ekström, and B. Hök, “Critical review of non-invasive respiratory monitoring in medical care,” *Medical and Biological Engineering and Computing*, vol. 41, no. 4, pp. 377–383, 2003.
- [55] 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, no. 1, p. e008769, 2016.
- [56] 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.
- [57] 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.
- [58] N. Ochsenbein-Kölble, M. Roos, T. Gasser, R. Huch, A. Huch, and R. Zimmermann, “Cross sectional study of automated blood pressure measurements throughout pregnancy,” *British Journal of Obstetrics & Gynaecology*, vol. 111, no. 4, pp. 319–325, 2004.
- [59] A. Halligan, E. O’Brien, K. O’Malley, F. Mee, N. Atkins, R. Conroy, J. J. Walshe, and M. Darling, “Twenty-four-hour ambulatory blood pressure measurement in a primigravid population,” *Journal of Hypertension*, vol. 11, no. 8, pp. 869–873, 1993.

- [60] D. Churchill and D. Beevers, "Differences between office and 24-hour ambulatory blood pressure measurement during pregnancy," *Obstetrics & Gynecology*, vol. 88, no. 3, pp. 455–461, 1996.
- [61] 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. 342–350, 2012.
- [62] A. A. Mahendru, T. R. Everett, I. B. Wilkinson, C. C. Lees, and C. M. McEniery, "A longitudinal study of maternal cardiovascular function from preconception to the postpartum period," *Journal of Hypertension*, vol. 32, no. 4, pp. 849–856, 2014.
- [63] T. Hartgill, T. Bergersen, and J. Pirhonen, "Core body temperature and the thermoneutral zone: a longitudinal study of normal human pregnancy," *Acta Physiologica*, vol. 201, no. 4, pp. 467–474, 2011.
- [64] C. Winter, J. Crofts, C. Laxton, S. Barnfield, and T. Draycott, eds., *PRactical Obstetric Multi-Professional Training (PROMPT): Course manual*. Cambridge University Press, 2 ed., 2012.
- [65] S. T. Blackburn, *Maternal, fetal, & neonatal physiology: A clinical perspective*. Elsevier Health Sciences, fourth edition ed., 2013.
- [66] R. Deckardt, P. M. Fembacher, K. Schneider, and H. Graeff, "Maternal arterial oxygen saturation during labor and delivery: pain-dependent alterations and effects on the newborn," *Obstetrics & Gynecology*, vol. 70, no. 1, pp. 21–25, 1987.
- [67] A. Templeton and G. 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.
- [68] E. Ekholm, R. Erkkola, and J. Hartiala, "Comparison of cardiovascular reflex tests and blood pressure measurement in prediction of pregnancy-induced hypertension," *European Journal of Obstetrics & Gynecology and Reproductive Biology*, vol. 54, no. 1, pp. 37–41, 1994.
- [69] U. Magriples, M. H. Boynton, T. S. Kershaw, S. S. Rising, and J. R. Ickovics, "Blood pressure changes during pregnancy: impact of race, body mass index, and weight gain," *American Journal of Perinatology*, vol. 30, no. 5, p. 415, 2013.
- [70] M. L. Thompson, M. A. Williams, and R. S. Miller, "Modelling the association of blood pressure during pregnancy with gestational age and body mass index," *Paediatric and Perinatal Epidemiology*, vol. 23, no. 3, pp. 254–263, 2009.
- [71] M. Ishikuro, T. Obara, H. Metoki, T. Ohkubo, M. Yamamoto, K. Akutsu, K. Sakurai, N. Iwama, M. Katagiri, K. Yagihashi, *et al.*, "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. 141–148, 2013.
- [72] C. Macdonald-Wallis, R. J. Silverwood, A. Fraser, S. M. Nelson, K. Tilling, D. A. Lawlor, and B. L. de Stavola, "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.
- [73] J. W. Samways, S. Vause, E. Kontopantelis, J. Eddleston, S. Ingleby, A. Roberts, and B. Clarke, "Maternal heart rate during the first 48h postpartum: a retrospective cross sectional study," *European Journal of Obstetrics & Gynecology and Reproductive Biology*, vol. 206, pp. 41–47, 2016.

- [74] A. Papageorgiou, S. Kennedy, L. Salomon, E. Ohuma, L. Cheikh Ismail, F. Barros, A. Lambert, M. Carvalho, Y. Jaffer, E. Bertino, *et al.*, “International standards for early fetal size and pregnancy dating based on ultrasound measurement of crown–rump length in the first trimester of pregnancy,” *Ultrasound in Obstetrics & Gynecology*, vol. 44, no. 6, pp. 641–648, 2014.
- [75] A. Reinders, A. Cuckson, J. Lee, and A. Shennan, “An accurate automated blood pressure device for use in pregnancy and pre-eclampsia: The Microlife 3BTO-A,” *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 112, no. 7, pp. 915–920, 2005.
- [76] Y. Chung, A. de Greeff, and A. Shennan, “Validation and compliance of a home monitoring device in pregnancy: Microlife WatchBP Home,” *Hypertension in Pregnancy*, vol. 28, no. 3, pp. 348–359, 2009.
- [77] M. T. Petterson, V. L. Begnoche, and J. M. Graybeal, “The effect of motion on pulse oximetry and its clinical significance,” *Anesthesia & Analgesia*, vol. 105, no. 6, pp. S78–S84, 2007.
- [78] K. H. Shelley, “Photoplethysmography: beyond the calculation of arterial oxygen saturation and heart rate,” *Anesthesia & Analgesia*, vol. 105, no. 6, pp. S31–S36, 2007.
- [79] A. Kamal, J. Harness, G. Irving, and A. Mearns, “Skin photoplethysmography? A review,” *Computer Methods and Programs in Biomedicine*, vol. 28, no. 4, pp. 257–269, 1989.
- [80] M. Raghuram, K. V. Madhav, E. H. Krishna, and K. A. Reddy, “On the performance of wavelets in reducing motion artifacts from photoplethysmographic signals,” in *Bioinformatics and Biomedical Engineering (iCBBE), 2010 4th International Conference on*, pp. 1–4, IEEE, 2010.
- [81] H. Han and J. Kim, “Artifacts in wearable photoplethysmographs during daily life motions and their reduction with least mean square based active noise cancellation method,” *Computers in Biology and Medicine*, vol. 42, no. 4, pp. 387–393, 2012.
- [82] M. R. Ram, K. V. Madhav, and E. H. Krishna, “ICA-based improved DTCWT technique for MA reduction in PPG signals with restored respiratory information,” *IEEE Transactions on Instrumentation and Measurement*, vol. 62, no. 10, 2013.
- [83] K. A. Reddy and V. J. Kumar, “Motion artifact reduction in photoplethysmographic signals using singular value decomposition,” in *2007 IEEE Instrumentation & Measurement Technology Conference IMTC 2007*, pp. 1–4, IEEE, 2007.
- [84] W. Karlen, K. Kobayashi, J. Ansermino, and G. Dumont, “Photoplethysmogram signal quality estimation using repeated Gaussian filters and cross-correlation,” *Physiological Measurement*, vol. 33, no. 10, p. 1617, 2012.
- [85] Q. Li and G. Clifford, “Dynamic time warping and machine learning for signal quality assessment of pulsatile signals,” *Physiological Measurement*, vol. 33, no. 9, p. 1491, 2012.
- [86] 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.
- [87] E. Gil, J. María Vergara, and P. Laguna, “Detection of decreases in the amplitude fluctuation of pulse photoplethysmography signal as indication of obstructive sleep apnea syndrome in children,” *Biomedical Signal Processing and Control*, vol. 3, no. 3, pp. 267–277, 2008.

- [88] M. A. Cretikos, R. Bellomo, K. Hillman, J. Chen, S. Finfer, and A. Flabouris, "Respiratory rate: the neglected vital sign," *Medical Journal of Australia*, vol. 188, no. 11, p. 657, 2008.
- [89] World Health Organisation, *Programme for control of acute respiratory infections: fourth programme report 1988-1989*, 1989.
- [90] 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.
- [91] D. Meredith, D. Clifton, P. Charlton, J. Brooks, C. 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.
- [92] B. M. Sayers, "Analysis of heart rate variability," *Ergonomics*, vol. 16, no. 1, pp. 17–32, 1973.
- [93] J. Pumpura, 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.
- [94] 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.
- [95] F. Rahman, S. Pechnik, D. Gross, L. Sewell, and D. S. Goldstein, "Low frequency power of heart rate variability reflects baroreflex function, not cardiac sympathetic innervation," *Clinical Autonomic Research*, vol. 21, no. 3, pp. 133–141, 2011.
- [96] L. A. Fleisher, "Heart rate variability as an assessment of cardiovascular status," *Journal of Cardiothoracic and Vascular Anesthesia*, vol. 10, no. 5, pp. 659–671, 1996.
- [97] P. Hung, S. Bonnet, R. Guillemaud, E. Castelli, and P. T. N. Yen, "Estimation of respiratory waveform using an accelerometer," in *Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on*, pp. 1493–1496, IEEE, 2008.
- [98] G.-Z. Liu, B.-Y. Huang, and L. Wang, "A wearable respiratory biofeedback system based on generalized body sensor network," *Telemedicine and e-Health*, vol. 17, no. 5, pp. 348–357, 2011.
- [99] A. Bates, M. J. Ling, J. Mann, and D. Arvind, "Respiratory rate and flow waveform estimation from tri-axial accelerometer data," in *2010 International Conference on Body Sensor Networks (BSN)*, pp. 144–150, IEEE, 2010.
- [100] J. Hernandez, D. McDuff, and R. W. Picard, "BioWatch: estimation of heart and breathing rates from wrist motions," in *Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare*, pp. 169–176, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2015.
- [101] L. Estrada, A. Torres, L. Sarlabous, and R. Jané, "Respiratory signal derived from the smartphone built-in accelerometer during a respiratory load protocol," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 6768–6771, IEEE, 2015.

- [102] A. R. Fekr, K. Radecka, and Z. Zilic, "Tidal volume variability and respiration rate estimation using a wearable accelerometer sensor," in *Wireless Mobile Communication and Healthcare (Mobihealth), 2014 EAI 4th International Conference on*, pp. 1–6, IEEE, 2014.
- [103] 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," in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, pp. 5677–5680, IEEE, 2009.
- [104] A. M. Chan, N. Ferdosi, and R. Narasimhan, "Ambulatory respiratory rate detection using ECG and a triaxial accelerometer," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4058–4061, IEEE, 2013.
- [105] M. Haescher, D. J. Matthies, J. Trimpop, and B. Urban, "A study on measuring heart-and respiration-rate via wrist-worn accelerometer-based seismocardiography (SCG) in comparison to commonly applied technologies," in *Proceedings of the 2nd International Workshop on Sensor-based Activity Recognition and Interaction*, p. 2, ACM, 2015.
- [106] P. H. Charlton, T. Bonnici, L. Tarassenko, D. A. Clifton, R. Beale, and P. J. Watkinson, "An assessment of algorithms to estimate respiratory rate from the electrocardiogram and photoplethysmogram," *Physiological Measurement*, vol. 37, no. 4, p. 610, 2016.
- [107] S. G. Fleming and L. Tarassenko, "A comparison of signal processing techniques for the extraction of breathing rate from the photoplethysmogram," *World Academy of Science, Engineering and Technology*, vol. 2, no. 4, pp. 232–236, 2007.
- [108] S. Fleming, L. Tarassenko, M. Thompson, and D. Mant, "Non-invasive measurement of respiratory rate in children using the photoplethysmogram," in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, pp. 1886–1889, IEEE, 2008.
- [109] A. Johansson, "Neural network for photoplethysmographic respiratory rate monitoring," *Medical and Biological Engineering and Computing*, vol. 41, no. 3, pp. 242–248, 2003.
- [110] J. Lee and K. 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.
- [111] J. Lee and K. H. Chon, "Time-varying autoregressive model-based multiple modes particle filtering algorithm for respiratory rate extraction from pulse oximeter," *Biomedical Engineering, IEEE Transactions on*, vol. 58, no. 3, pp. 790–794, 2011.
- [112] K. V. Madhav, M. R. Ram, E. H. Krishna, N. R. Komalla, and K. A. Reddy, "Estimation of respiration rate from ECG, BP and PPG signals using empirical mode decomposition," in *2011 IEEE Instrumentation and Measurement Technology Conference (I2MTC)*, pp. 1–4, IEEE, 2011.
- [113] 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.
- [114] L. Nilsson, A. Johansson, and S. Kalman, "Monitoring of respiratory rate in post-operative care using a new photoplethysmographic technique," *Journal of Clinical Monitoring and Computing*, vol. 16, no. 4, pp. 309–315, 2000.

- [115] K. H. Chon, S. Dash, and K. Ju, "Estimation of respiratory rate from photoplethysmogram data using time–frequency spectral estimation," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 8, pp. 2054–2063, 2009.
- [116] W. Karlen, C. Brouse, E. Cooke, J. Ansermino, and G. Dumont, "Respiratory rate estimation using respiratory sinus arrhythmia from photoplethysmography," in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pp. 1201–1204, IEEE, 2011.
- [117] P. S. Addison, J. N. Watson, M. L. Mestek, and R. S. 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.
- [118] M. Orini, M. Peláez-Coca, R. Bailón, and E. Gil, "Estimation of spontaneous respiratory rate from photoplethysmography by cross time-frequency analysis," *Computing in Cardiology*, vol. 38, pp. 661–664, 2011.
- [119] 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.
- [120] 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.
- [121] 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.
- [122] K. V. Madhav, M. R. Ram, E. H. Krishna, N. R. Komalla, and K. A. Reddy, "Robust extraction of respiratory activity from PPG signals using modified MSPCA," *IEEE Transactions on Instrumentation and Measurement*, vol. 62, no. 5, pp. 1094–1106, 2013.
- [123] T. Zhu, M. A. Pimentel, G. D. Clifford, and D. A. Clifton, "Bayesian fusion of algorithms for the robust estimation of respiratory rate from the photoplethysmogram," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 6138–6141, IEEE, 2015.
- [124] L. Loerup, *Physiological Patterns in Pregnancy*. PhD thesis, University of Oxford, 2016.
- [125] 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.
- [126] H. Gan, W. Karlen, D. Dunsmuir, G. Zhou, M. Chiu, G. A. Dumont, and J. M. Ansermino, "The performance of a mobile phone respiratory rate counter compared to the WHO ARI timer," *Journal of Healthcare Engineering*, vol. 6, no. 4, pp. 691–704, 2015.
- [127] J. Shlens, "A tutorial on principal component analysis," *arXiv preprint arXiv:1404.1100*, 2014.
- [128] K. Roth, I. Kauppinen, P. A. Esquef, and V. Valimaki, "Frequency warped burg's method for ar-modeling," in *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.*, pp. 5–8, IEEE, 2003.

- [129] 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.
- [130] A. Guazzi, S. A. Shah, and L. Tarassenko, "Signal processing method and apparatus," 2016. WO patent 2016092291.
- [131] J. M. Bland and D. G. Altman, "Multiple significance tests: the Bonferroni method," *British Medical Journal*, vol. 310, no. 6973, p. 170, 1995.
- [132] D. G. Altman and J. M. Bland, "Measurement in medicine: the analysis of method comparison studies," *The Statistician*, pp. 307–317, 1983.
- [133] R. A. Cernat, C. Ungureanu, G. M. Ungureanu, R. Aarts, and J. Arends, "Real-time extraction of the respiratory rate from photoplethysmographic signals using wearable devices," in *Proceedings of the European Conference on Ambient Intelligence, Eindhoven, Netherlands*, pp. 1–17, 2014.
- [134] J.-W. Yoon, Y.-S. Noh, Y.-S. Kwon, 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.
- [135] P. MacKinnon and J. Morris, *Oxford Textbook of Functional Anatomy*, vol. 1. Oxford University Press, 2005.
- [136] A. T. Papageorgiou, E. O. Ohuma, D. G. Altman, T. Todros, L. C. Ismail, A. Lambert, Y. A. Jaffer, E. Bertino, M. G. Gravett, M. Purwar, *et al.*, "International standards for fetal growth based on serial ultrasound measurements: the Fetal Growth Longitudinal Study of the INTERGROWTH-21st project," *The Lancet*, vol. 384, no. 9946, pp. 869–879, 2014.
- [137] 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, no. 4, pp. 459–464, 2009.
- [138] S. Y. Chu, W. M. Callaghan, S. Y. Kim, C. H. Schmid, J. Lau, L. J. England, and P. M. Dietz, "Maternal obesity and risk of gestational diabetes mellitus," *Diabetes care*, vol. 30, no. 8, pp. 2070–2076, 2007.
- [139] A. Gelman and J. Hill, *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2007.
- [140] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. The MIT Press, 2006.
- [141] S. Fleming, M. Thompson, R. Stevens, C. Heneghan, A. Plüddemann, I. Macconochie, 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.
- [142] T. J. Cole and P. J. Green, "Smoothing reference centile curves: the lms method and penalized likelihood," *Statistics in Medicine*, vol. 11, no. 10, pp. 1305–1319, 1992.
- [143] A. Wade and J. Kurmanavicius, "Creating unbiased cross-sectional covariate-related reference ranges from serial correlated measurements," *Biostatistics*, vol. 10, no. 1, pp. 147–154, 2009.
- [144] R. Bakker, E. A. Steegers, J. P. Mackenbach, A. Hofman, and V. W. Jaddoe, "Maternal smoking and blood pressure in different trimesters of pregnancy: The Generation R study," *Journal of Hypertension*, vol. 28, no. 11, pp. 2210–2218, 2010.

- [145] G. Kac, R. H. Mendes, D. R. Farias, I. Eshriqui, F. Rebelo, C. Benaim, A. A. F. Vilela, N. S. Lima, W. A. Peres, and G. F. Salles, "Hepatic, renal and inflammatory biomarkers are positively associated with blood pressure changes in healthy pregnant women: A prospective cohort," *Medicine*, vol. 94, no. 20, 2015.
- [146] E. Borghi, M. De Onis, C. Garza, J. Van den Broeck, E. Frongillo, L. Grummer-Strawn, S. Van Buuren, H. Pan, L. Molinari, R. Martorell, *et al.*, "Construction of the World Health Organization child growth standards: selection of methods for attained growth curves," *Statistics in Medicine*, vol. 25, no. 2, pp. 247–265, 2006.
- [147] P. Royston and D. G. Altman, "Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling," *Applied Statistics*, pp. 429–467, 1994.
- [148] A. Gelman, J. Hwang, and A. Vehtari, "Understanding predictive information criteria for Bayesian models," *Statistics and Computing*, vol. 24, no. 6, pp. 997–1016, 2014.
- [149] H. Gu, S. Zhang, Y. Qiao, Z. Luo, Y. Zeng, and Q. Wang, "A study of maternal hemodynamic change during healthy pregnancy and women with gestation hypertension," *Biomedical Materials and Engineering*, vol. 16, no. 1, pp. 77–82, 2006.
- [150] 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," vol. 11, no. s2, pp. 543–562, 1993.
- [151] R. Swanton, S. Al-Rawi, and M. 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.
- [152] P. J. Maguire, A. O'Higgins, K. Power, and M. Turner, "The Irish Maternity Early Warning System (IMEWS)," *Irish Medical Journal*, 2014.
- [153] NICE, "NG25: Preterm labour and birth," tech. rep., National Institute for Health and Care Excellence (NICE), 2015.
- [154] C. K. Shapiro-Mendoza and E. M. Lackritz, "Epidemiology of late and moderate preterm birth," in *Seminars in Fetal and Neonatal Medicine*, vol. 17, pp. 120–125, Elsevier, 2012.
- [155] J. Bai, F. W. Wong, A. Bauman, and M. Mohsin, "Parity and pregnancy outcomes," *American Journal of Obstetrics and Gynecology*, vol. 186, no. 2, pp. 274–278, 2002.
- [156] A. Mantakas and T. Farrell, "The influence of increasing BMI in nulliparous women on pregnancy outcome," *European Journal of Obstetrics & Gynecology and Reproductive Biology*, vol. 153, no. 1, pp. 43–46, 2010.
- [157] M. Blondon, L. Harrington, F. Boehlen, H. Robert-Ebadi, M. Righini, and N. Smith, "Pre-pregnancy BMI, delivery BMI, gestational weight gain and the risk of postpartum venous thrombosis," *Thrombosis Research*, vol. 145, pp. 151–156, 2016.
- [158] NICE, "CG189 obesity: identification, assessment and management," tech. rep., National Institute for Health and Care Excellence (NICE), 2014.
- [159] NICE, "NG7: Preventing excess weight gain," tech. rep., National Institute for Health and Care Excellence (NICE), 2015.
- [160] B. T. West, A. T. Galecki, and K. B. Welch, *Linear mixed models: A practical guide using statistical software*. Taylor and Francis, 2007.

- [161] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian data analysis*. Taylor & Francis, 3 ed., 2014.
- [162] G. Lewis, R. Cantwell, T. Clutton-Brock, *et al.*, “Saving mothers’ lives: Reviewing maternal deaths to make motherhood safer: 2006-2008,” vol. 118, no. s1, pp. 1–203, 2011.
- [163] M. Knight, D. Tuffnell, S. Kenyon, J. Shakespeare, R. Gray, and J. J. Kurinczuk, “Saving lives, improving mothers’ care: Surveillance of maternal deaths in the UK 2011–13 and lessons learned to inform maternity care from the UK and Ireland Confidential Enquiries into Maternal Deaths and Morbidity 2009 – 13,” tech. rep., MBRRACE-UK, 2015.
- [164] D. E. Bick, J. Sandall, M. Furuta, M. Y. Wee, R. Isaacs, G. B. Smith, S. Beake, *et al.*, “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.
- [165] R. Isaacs, M. Wee, D. Bick, S. Beake, Z. Sheppard, S. Thomas, V. Hundley, G. Smith, E. Teijlingen, and P. Thomas, “A national survey of obstetric early warning systems in the United Kingdom: Five years on,” *Anaesthesia*, vol. 69, no. 7, pp. 687–692, 2014.
- [166] S. Paterson-Brown and C. Howell, eds., *Managing Obstetric Emergencies and Trauma (MOET): Course manual*. Cambridge University Press, 3 ed., 2014.
- [167] 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. 68, no. 4, pp. 354–367, 2013.
- [168] J. M. Mhyre, R. D’Oria, A. B. Hameed, J. R. Lappen, S. L. Holley, S. K. Hunter, R. L. Jones, J. C. King, and M. E. D’Alton, “The maternal early warning criteria: a proposal from the national partnership for maternal safety,” *Journal of Obstetric, Gynecologic, & Neonatal Nursing*, vol. 43, no. 6, pp. 771–779, 2014.
- [169] G. B. Smith and D. Prytherch, “Obstetric early warning scores: Much more work required,” *Anaesthesia*, vol. 68, no. 7, pp. 778–779, 2013.
- [170] P. J. Maguire, A. C. OHiggins, K. A. Power, N. Daly, A. McKeating, and M. J. Turner, “Maternal bacteremia and the Irish maternity early warning system,” *International Journal of Gynecology & Obstetrics*, vol. 129, no. 2, pp. 142–145, 2015.
- [171] 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 and Gynecology*, vol. 212, no. 4, pp. 536–e1, 2015.
- [172] “Providing equity of critical and maternity care for the critically ill pregnant or recently pregnant woman,” tech. rep., The Royal College of Anaesthetists, 2011.
- [173] 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, no. 2, pp. 152–155, 2014.
- [174] 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, no. 8, pp. 1013–1018, 2011.

- [175] K. Challen and D. Roland, “Early warning scores: a health warning,” *Emergency Medicine Journal*, 2016. Ahead of print.
- [176] M. E. Bauer, S. T. Bauer, B. Rajala, M. P. MacEachern, L. S. Polley, D. Childers, and D. M. Aronoff, “Maternal physiologic parameters in relationship to systemic inflammatory response syndrome criteria: A systematic review and meta-analysis,” *Obstetrics & Gynecology*, vol. 124, no. 3, pp. 535–541, 2014.
- [177] P. J. Maguire, K. A. Power, A. F. Downey, A. C. OHiggins, S. R. Sheehan, and M. J. Turner, “Evaluation of the systemic inflammatory response syndrome criteria for the diagnosis of sepsis due to maternal bacteremia,” *International Journal of Gynecology & Obstetrics*, vol. 133, no. 1, pp. 116–119, 2016.
- [178] G. B. Smith, D. R. Prytherch, P. Schmidt, P. I. Featherstone, D. Knight, G. Clements, and M. A. Mohammed, “Hospital-wide physiological surveillance—a new approach to the early identification and management of the sick patient,” *Resuscitation*, vol. 71, no. 1, pp. 19–28, 2006.
- [179] D. Wong, T. Bonnici, J. Knight, L. Morgan, P. Coombes, and P. Watkinson, “SEND: A system for electronic notification and documentation of vital sign observations,” *BMC Medical Informatics and Decision Making*, vol. 15, no. 1, p. 1, 2015.
- [180] A. Le Bas, E. Chandrachan, A. Addei, and S. Arulkumaran, “Use of the obstetric shock index as an adjunct in identifying significant blood loss in patients with massive postpartum hemorrhage,” *International Journal of Gynecology & Obstetrics*, vol. 124, no. 3, pp. 253–255, 2014.
- [181] A. M. El Ayadi, H. L. Nathan, P. T. Seed, E. A. Butrick, N. L. Hezelgrave, A. H. Shennan, and S. Miller, “Vital sign prediction of adverse maternal outcomes in women with hypovolemic shock: The role of shock index,” *PLoS One*, vol. 11, no. 2, p. e0148729, 2016.
- [182] R. H. Birkhahn, T. J. Gaeta, D. Terry, J. J. Bove, and J. Tloczkowski, “Shock index in diagnosing early acute hypovolemia,” *The American Journal of Emergency Medicine*, vol. 23, no. 3, pp. 323–326, 2005.
- [183] L. Clifton, D. A. Clifton, M. A. Pimentel, P. J. Watkinson, and L. Tarassenko, “Gaussian Process regression in vital-sign early warning systems,” in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 6161–6164, IEEE, 2012.
- [184] G. Verbeke, S. Fieuws, G. Molenberghs, and M. Davidian, “The analysis of multivariate longitudinal data: A review,” *Statistical Methods in Medical Research*, vol. 23, no. 1, pp. 42–59, 2014.
- [185] S. Fieuws and G. Verbeke, “Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles,” *Biometrics*, vol. 62, no. 2, pp. 424–431, 2006.
- [186] K. Hillman, J. Chen, M. Cretikos, R. Bellomo, D. Brown, G. Doig, S. Finfer, A. Flabouris, MERIT Study Investigators, *et al.*, “Introduction of the medical emergency team (MET) system: a cluster-randomised controlled trial,” *The Lancet*, vol. 365, no. 9477, pp. 2091–2097, 2005.
- [187] T. Bonnici, S. Gerry, D. Wong, J. Knight, and P. Watkinson, “Evaluation of the effects of implementing an electronic early warning score system: protocol for a stepped wedge study,” *BMC Medical Informatics and Decision Making*, vol. 16, no. 1, p. 1, 2016.
- [188] 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*, pp. emermed–2011, 2012.

- [189] A. Koski and M. Juhola, “Segmentation of digital signals based on estimated compression ratio,” *IEEE transactions on Biomedical Engineering*, vol. 43, no. 9, pp. 928–938, 1996.
- [190] B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell, “Stan: A probabilistic programming language,” *Journal of Statistical Software*, 2016.
- [191] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.
- [192] M. D. Hoffman and A. Gelman, “The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1593–1623, 2014.
- [193] T. Sorensen and S. Vasisht, “A tutorial on fitting Bayesian linear mixed models using Stan,” *Journal of Mathematical Psychology*, [In Press].
- [194] T. Sorensen and S. Vasisht, “Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists,” [Pre-print].
- [195] A. Vehtari and A. Gelman, “WAIC and cross-validation in Stan,” [Pre-Print], 2014.