Non-invasive vital-sign monitoring and data fusion in acute care



Hamish Tomlinson

Keble College

University of Oxford

Supervised by

Professor Lionel Tarassenko

Dr. Marco Pimentel

Trinity Term 2018

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

Declaration

I declare that the work contained in this thesis is entirely my own, except where explicitly noted.

Hamish Tomlinson
Keble College

Acknowledgements

This DPhil has been the biggest challenge of my life and I am incredibly grateful for the support of those around me. Firstly, thanks goes to my supervisors. Thank you to Professor Lionel Tarassenko for your guidance over these 4 years, for teaching me so much about writing (fewer *Tomlinson-isms* and less *tabloid journalism*), for shaping how I approach engineering and clinical problems, and for keeping me on track. Thank you to Dr Marco Pimentel for your friendship, mentorship and advice. To Dr Mauricio Villaroel, thank you for your patience, technical advice, and for setting such a high standard for the quality of work expected.

My time in the lab has been enriched immensely by the members of my research group. To the camera team – thank you for your support and for teaching me the tricks of the trade. Thanks to the members of *Time-Series Terrace* for adopting a member of a foreign row! Your friendship made the ride up to Headington worth it.

I am grateful to Dr Peter Watkinson, Julie Darbyshire, Jody Ede and others at the Kadoorie Centre. It has been amazing to learn from such a wealth of clinical knowledge and I know it has shaped my work for the better. To Dr Oliver Gibson and the staff at Oxehealth – thank you for your collaboration during our clinical trial.

This DPhil has been supported by a number of institutions, which I must acknowledge. I owe tremendous gratitude to the Rhodes Trust for enabling my time in Oxford. I feel so lucky for the friendships, experiences, lessons and ideas that I have developed through the Rhodes Scholarship community. Thanks to the staff and members of Keble College and the Institute of Biomedical Engineering for the support during this time. A special thanks to Professor Geoff Chase and colleagues at the University of Canterbury for encouraging me and inspiring my pursuit of research.

From the moment I stepped on to the Trans-Siberian train in Beijing, Oxford-bound, I've been surrounded by incredible friends. Thanks to those I have met through Rhodes, Good Lad, Keble, 234 Iffley Rd and otherwise just being from down-under, hanging out has been a blast! A special thanks to Ed and Beans at the Eagles Nest; Canada-Kiwi relations have never been stronger. Thanks also to my friends in New Zealand, who have encouraged me from afar, have bought Kiwi humour on trips to visit, and have kept me grounded.

I am incredibly lucky to have such a loving and supportive family, who have always encouraged me during this DPhil. Thank you Mum for being at the end of many long and one-sided phone calls, which have helped me so much. Dad – thank you for always pushing me to achieve my goals. To Sarah and Chris, thanks for your support while I've been away – I promise I'll come home more often now! To Debbie, Peter, Maddy and Claire, thank you for giving me a second home and family.

Most importantly, thank you Eleanor. It has been an incredible adventure and I'm so glad to have shared it with you. Thank you for encouraging me, believing in me, and for cheering me up with your never-ending supply of silliness. Your love, support and patience has made this possible.

Non-invasive vital-sign monitoring and data fusion in acute care

Hamish Tomlinson
Thesis submitted for the degree of Doctor of Philosophy
Keble College
Trinity 2018

Abstract

Post-operative patients can deteriorate physiologically, leading to adverse events such as cardiac arrest, unexpected intensive care unit (ICU) admission or death. Clinical staff monitor patients by observation of the vital signs. Early warning scores (EWS) are computed from vital-sign values to indicate when a patient requires an escalation in care. The investigations in this thesis assess the use of machine learning and continuous vital-sign monitoring to improve detection of physiological deterioration.

Two post-operative datasets are introduced; vital-sign data collected from 407 patients using manual and continuous methods, and vital-sign data collected from 447 patients using manual methods. A machine learning model is trained on the first manual dataset using vital sign and diurnal vital-sign variability values, and is tested on the second dataset. The model achieves an Area Under the Receiver-Operator Characteristic curve (AUROC) of 0.886 for predicting adverse events within 24 hours, with the National Early Warning Score achieving an AUROC of 0.848. Respiratory rate variability, blood oxygen saturation variability and systolic blood pressure variability are more predictive of adverse events than the corresponding vital signs.

A method comparison study is conducted for manual and continuous vital-sign measurements. There is no evidence of a bias towards normal readings in manual observations, or an arousal effect due to the vital-sign taker. However, differences between manual and continuous data are shown to be frequently large. EWS models should be designed specifically for continuous vital-sign data. Methods for estimating diurnal vital-sign variability from continuous data are described, and are shown to discriminate between physiological stability and deterioration.

A clinical study of video camera-based non-contact monitoring of 15 post-operative patients is described. Heart rate and respiratory rate estimation methods are developed with mean average errors of 0.8 to 2.8 heart-beats/minute and 0.6 to 2.0 breaths/minute during 5 periods of physiological deterioration. Spatial maps of respiratory signal amplitude are shown for a patient with abnormal respiratory physiology.

In summary, EWSs should incorporate features of vital-sign variability, with future work on continuous monitoring implementations. Camera-based methods may provide continuous vital-sign monitoring and spatial physiological maps in post-operative care.

Related Publications

- **Hamish Tomlinson**, Marco AF Pimentel, Stephen Gerry, Peter J Watkinson, David A Clifton, and Lionel Tarassenko. Smoothing Effect in Vital-Sign Recordings: Fact or Fiction? A Retrospective Cohort Analysis of Manual and Continuous Vital-Sign Measurements to Assess Data Smoothing in Post-Operative Care. *Anesthesia & Analgesia*, *accepted for publication*, 2018.
- **Hamish Tomlinson**, Marco AF Pimentel, Peter J Watkinson, and Lionel Tarassenko. Time-series clustering for discovering trajectories of physiological deterioration in post-operative care. *NIPS Workshop on Machine Learning for Health*, 2016.
- **Hamish Tomlinson**, Nicholas Dunkley, Oliver J Gibson, Peter J Watkinson, and Lionel Tarassenko. Assessing the feasibility of non-contact vital-sign monitoring in the clinical setting. *Urgent Care in the Thames Valley*, 2018.
- **Hamish Tomlinson**, Marco AF Pimentel, Peter J Watkinson, and Lionel Tarassenko. Improving post-operative care with continuous vital-sign monitoring, *MEIBioeng 16*, 2016.

Contents

Contents vi

List of Figures x

List of Tables xvi

List of Acronyms xviii

1 Post-operative monitoring 1

1.1 Vital-sign monitoring . . . 1

1.1.1 Physiology of the vital signs . . . 2

1.1.2 Manual measurement methods . . . 4

1.1.3 Electronic measurement methods . . . 6

1.1.4 Summary . . . 12

1.2 Adverse events in post-operative care . . . 13

1.2.1 The evidence and cost of avoidable errors . . . 13

1.2.2 Rapid response systems and new approaches to care . . . 15

1.2.3 Summary . . . 17

1.3 Detecting physiological deterioration . . . 17

1.3.1 Early warning scores . . . 17

1.3.2 Data driven approaches . . . 19

1.4 Continuous monitoring outside of critical care . . . 23

1.4.1 The clinical need for continuous monitoring . . . 24

1.4.2 Challenges in continuous vital-sign monitoring . . . 29

1.4.3 Wearable vital-sign monitors . . . 31

1.4.4 Non-contact vital-sign monitoring . . . 32

1.5 Overview of thesis . . . 34

1.6 Summary . . . 35

2 The Computer Alerting Monitoring System Studies 36

2.1 Study population . . . 36

2.2 The CALMS-2 study . . . 37

2.2.1 Study design . . . 38

2.2.2 The CALMS-2 dataset . . . 40

2.3 The CALMS-3 study . . . 47

2.3.1 Motivation . . . 47
2.3.2 Study design . . . 48
2.3.3 The CALMS-3 dataset . . . 49
2.4 Summary . . . 51

3 Diurnal vital-sign variability in post-operative patients 52
3.1 Previous work . . . 53
3.2 Univariate vital-sign variability analysis . . . 55
3.2.1 Overview . . . 55
3.2.2 Model design . . . 57
3.2.3 Model evaluation . . . 61
3.2.4 Results . . . 63
3.2.5 Discussion . . . 66
3.3 Building an improved machine learning EWS . . . 68
3.3.1 Overview . . . 68
3.3.2 Model design . . . 68
3.3.3 Model evaluation . . . 72
3.3.4 Results . . . 75
3.3.5 Discussion . . . 80
3.4 Conclusions . . . 84

4 Method comparison of continuous and manual vital-sign data 85
4.1 Previous work . . . 86
4.1.1 Clinical errors . . . 86
4.1.2 Methodological issues . . . 87
4.1.3 Summary . . . 88
4.2 Data preparation . . . 88
4.2.1 Overview . . . 88
4.2.2 Data selection . . . 88
4.3 Assessing the “smoothing effect” . . . 91
4.3.1 Methods . . . 91
4.3.2 Results . . . 92
4.3.3 Discussion . . . 95
4.4 Assessing agreement . . . 97
4.4.1 Methods . . . 97
4.4.2 Results . . . 98
4.4.3 Discussion . . . 100

4.5 Assessing the "arousal effect" . . . 101
4.5.1 Methods . . . 101
4.5.2 Results . . . 101
4.5.3 Discussion . . . 103
4.6 Conclusions . . . 103

5 Continuous vital-sign variability 104
5.1 Previous work . . . 106
5.2 The value of continuous vital-sign data . . . 107
5.3 Computing variability from continuous data . . . 113
5.3.1 Overview . . . 113
5.3.2 Pre-processing . . . 113
5.3.3 Variability calculation methods . . . 115
5.3.4 Evaluating variability methods . . . 115
5.4 Results . . . 122
5.5 Discussion . . . 127
5.6 Conclusions . . . 130

6 Non-contact vital-sign monitoring 132
6.1 The Non-Contact Vital-sign Monitoring study . . . 133
6.1.1 Equipment . . . 136
6.1.2 Data Collection . . . 138
6.1.3 Study Summary . . . 142
6.2 Patient Case Studies . . . 144
6.3 Heart rate estimation . . . 151
6.3.1 Overview . . . 151
6.3.2 ROI selection . . . 152
6.3.3 Signal processing . . . 155
6.3.4 Computing heart rate estimates . . . 157
6.3.5 Signal quality index . . . 157
6.4 Respiratory rate estimation . . . 159
6.4.1 Overview . . . 159
6.4.2 ROI selection . . . 160
6.4.3 Signal processing . . . 163
6.4.4 Computing respiratory rate estimates . . . 165
6.5 Results . . . 166
6.6 Discussion . . . 171

6.7 Conclusions 174

7 Localised respiration mapping 176
7.1 Previous work 176
7.2 Patient case study 178
7.2.1 Validating non-contact respiratory rate estimates 180
7.3 Temporal variations in respiration 182
7.3.1 Overview 182
7.3.2 Respiratory signal extraction 182
7.3.3 Respiratory signal visualisation 183
7.3.4 Discussion 186
7.4 Spatial variations in respiration 188
7.4.1 Overview 188
7.4.2 Respiratory signal extraction 188
7.4.3 Visualising spatial variations in respiratory amplitude 192
7.4.4 Discussion 196
7.5 Respiratory amplitude maps 197
7.5.1 Overview 197
7.5.2 Data segments 197
7.5.3 Computing respiratory amplitude 199
7.5.4 Results 200
7.5.5 Discussion 202
7.6 Conclusions 204

8 Conclusions and future work 205
8.1 Thesis review 205
8.2 Future work 207
8.2.1 Predictive modelling 207
8.2.2 Continuous vital-sign monitoring 209
8.2.3 Non-contact vital-sign monitoring 210
8.3 Concluding remarks 211

A The Bland-Altman Phenomenon 213
A.1 Background 213
A.2 The “smoothing effect” 214

References 216

List of Figures

1.1 An electrocardiogram waveform for one heart beat. 7
1.2 An arterial blood pressure waveform for one heart beat. 9
1.3 A typical finger pulse oximetry sensor (a). The absorption spectra of oxygenated and deoxygenated haemoglobin (b). 10
1.4 A 15-second window of photoplethysmogram waveform. 11
1.5 An impedance pneumography waveform, showing the fluctuations in transthoracic impedance over the course of six breaths. 12

2.1 Time of continuous heart rate (a), respiratory rate (b), and oxygen saturation (c) monitoring received during the CALMS-2 study for the "normal" patient group and the "abnormal" patient group. 46

3.1 Example time-series of heart-rate observations before and after an adverse event (unexpected admission to the Intensive Care Unit) showing the data labelling process for the binary classification task of an early warning score. 56
3.2 The analysis pipeline for estimating the importance of each vital-sign and vital-sign variability feature. 57
3.3 Receiver operating characteristic curves for the test set in the five folds, for each vital sign and vital-sign variability feature. 64
3.4 Risk profiles showing the predicted event likelihood against the vital-sign or vital-sign variability feature value for the models fitted on the training set of the five folds. 65
3.5 Receiver-operator characteristic curves plotting true positive rate against false positive rate for the models tested on the CALMS-2 dataset. . . 78
3.6 Precision-recall curves plotting precision (positive predictive value) against recall (sensitivity) for the models tested on the CALMS-2 dataset. 78
3.7 Time-to-event curves plotting the median time of the first alert (positive model prediction) within 24-hours of an adverse event against the false positive rate for the models tested on the CALMS-2 dataset. 79

4.1 The selection process for extracting pairs of contemporaneous manual and continuous vital-sign measurements of respiratory rate, heart rate and oxygen saturation from the CALMS-2 dataset. 89

4.2 An example heart rate observation, for which the "before" window is shown in blue, the "during" window is shown in green, and the manual observation (synchronised to the time of the blood pressure measurement) is shown in red. 90
4.3 Scatter plots of the continuous-manual difference against the continuous-manual average of measurements of respiratory rate (a), heart rate (b) and oxygen saturation (c). 93
4.4 Bland-Altman plots for the comparison of continuous and manual measurements of respiratory rate (a), heart rate (b) and oxygen saturation (c). 99
4.5 Bland-Altman plots for the comparison of continuous measurements of respiratory rate (a), heart rate (b) and oxygen saturation (c) before and during the time of clinical observation. 102

5.1 Vital-sign time-series for CALMS-2 patient 447 in the final 48 hours before discharge from the ward. 108
5.2 Vital-sign time-series for CALMS-2 patient 462 in the final 48 hours before discharge from the ward. 108
5.3 Vital-sign time-series for CALMS-2 patient 471 in the final 48 hours before discharge from the ward. 110
5.4 Vital-sign time-series for CALMS-2 patient 348 in the final 48 hours before an adverse clinical event. 110
5.5 Vital-sign time-series for CALMS-2 patient 362 in the final 48 hours before an adverse clinical event. 112
5.6 Vital-sign time-series for CALMS-2 patient 464 in the final 48 hours before an adverse clinical event. 112
5.7 The analysis pipeline for computing and evaluating diurnal variability features from continuous vital-sign data for heart rate, respiratory rate and oxygen saturation. 113
5.8 Example vital-sign time-series of heart rate (a), respiratory rate (b), and oxygen saturation (c) for a 24-hour period. The raw vital-sign data are shown in blue, and the (5-minute) median-filtered data are shown in orange. 114
5.9 Distributions of diurnal heart rate variability for the "normal" and "abnormal" patient groups in the CALMS-2 dataset. 118

5.10 Cumulative distribution functions of diurnal respiratory rate variability for the "normal" and "abnormal" patient groups in the CALMS-2 dataset. 120
5.11 Average trajectories of diurnal oxygen saturation variability in the 48 hours before discharge from the ward for the "normal" patient group and before an adverse clinical event for the "abnormal" patient group in the CALMS-2 dataset. 121
5.12 Diurnal vital-sign variability time-series for CALMS-2 patient 447 in the final 48 hours before discharge from the ward. 124
5.13 Diurnal vital-sign variability time-series for CALMS-2 patient 462 in the final 48 hours before discharge from the ward. 124
5.14 Diurnal vital-sign variability time-series for CALMS-2 patient 471 in the final 48 hours before discharge from the ward. 125
5.15 Diurnal vital-sign variability time-series for CALMS-2 patient 348 in the final 48 hours before an adverse event. 125
5.16 Diurnal vital-sign variability time-series for CALMS-2 patient 362 in the final 48 hours before an adverse event. 126
5.17 Diurnal vital-sign variability time-series for CALMS-2 patient 464 in the final 48 hours before an adverse event. 126

6.1 Position of the equipment trolley at the end of the patient's bed. . . . 135
6.2 Technical specifications of the monochromatic camera equipment used in the NVSM study. 137
6.3 Set-up of the trolley containing the non-contact vital-sign monitoring equipment. 138
6.4 Example video frames before and after processing with a blurring algorithm. 141
6.5 Timeline of the study segments and important events for patient NVSM15. The times of chest X-rays are shown as black dots. 145
6.6 Reference heart rate (a), respiratory rate (b), and oxygen saturation (c) for segment NVSM15-1. 146
6.7 Reference heart rate (a), respiratory rate (b), and oxygen saturation (c) for segment NVSM15-2. 147
6.8 Reference heart rate (a), respiratory rate (b), and oxygen saturation (c) for segment NVSM15-3. 148
6.9 Reference heart rate (a), respiratory rate (b), and oxygen saturation (c) for segment NVSM19-1. 149

6.10 Reference heart rate (a), respiratory rate (b), and oxygen saturation (c) for segment NVSM23-1. 150
6.11 Overview of the non-contact heart rate estimation algorithm. 152
6.12 Example video frame from segment NVSM23-1, showing (in red) the position of an 80x80 ROI for heart rate estimation on the left cheek. . 153
6.13 Example 1-minute segment of raw ROI time-series data during NVSM15-1, extracted from ROIs of various sizes on the right cheek. 154
6.14 Example 15-second window of ROI time-series taken from segment NVSM15-1 before and after band-pass filtering for heart rate. 156
6.15 Example window of high-quality bandpass filtered ROI time-series (a) with a 1024-point FFT power spectrum (b) 158
6.16 Example window of poor-quality bandpass filtered ROI time-series (a) with a 1024-point FFT power spectrum (b). 158
6.17 Overview of the non-contact respiratory rate estimation algorithm. . . 160
6.18 Example video frame from segment NVSM15-3, demonstrating the position of a 30x30 ROI for respiratory rate estimation, which is drawn with a red border on the middle of the chest. 161
6.19 Example 1-minute segment of raw ROI time-series data during NVSM15-2, extracted from ROIs of various sizes on the chest. 162
6.20 Example window of ROI time-series from segment NVSM15-1 before and after band-pass filtering for respiratory rate. 163
6.21 A 3-minute period of filtered 20x20 ROI time-series from the right side of the chest during segment NVSM15-2. The peaks of the respiratory waveforms are shown as blue dots and the onsets are shown as red dots. 165
6.22 Camera-derived heart rate and respiratory rate estimates compared with reference values for segment NVSM15-1. 168
6.23 Camera-derived heart rate and respiratory rate estimates compared with reference values for segment NVSM15-2. 168
6.24 Camera-derived heart rate and respiratory rate estimates compared with reference values for segment NVSM15-3. Reference continuous respiratory rate values were not available as there was no impedence pneumography signal. 169
6.25 Camera-derived heart rate and respiratory rate estimates compared with reference values for segment NVSM19-1. 169
6.26 Camera-derived heart rate and respiratory rate estimates compared with reference values for segment NVSM23-1. 170

6.27 Distributions of time gaps between successive heart rate (a) and respiratory rate (b) estimates for all 5 study segments. 170

6.28 Camera-derived respiratory rate estimates compared with reference values for segment NVSM15-2. The estimation algorithm used 10-second sliding windows and did not include a smoothness criterion. . 172

7.1 Timeline of the final day of recording for patient NVSM15, showing the timing of the chest X-rays diagnosing the right-side pneumothorax and ruptured diaphragm. 178

7.2 Reference heart rate (a), respiratory rate (b), and oxygen saturation (c) for segment NVSM15-3. Re-plotted from figure 6.8. 179

7.3 Non-contact respiratory rate estimates derived from 30x30 pixel ROIs on the right (a) and left (b) sides of the chest compared to respiratory rate estimates derived from a 30x30 pixel ROI on the forehead, during segment NVSM15-3. 182

7.4 Positions of the 30x30 pixel ROIs from which respiratory signals were extracted for the analysis of temporal variations in respiratory signal from different locations. 183

7.5 Respiratory signals during the last 15 minutes of data segment NVSM15-3 extracted from ROIs located at various horizontal and vertical locations on the chest. 184

7.6 The waveform amplitudes of the respiratory signals during the last 15 minutes of data segment NVSM15-3 extracted from ROIs located at various horizontal and vertical locations on the chest. 185

7.7 Positions of the horizontal rows of overlapping ROIs on the top, middle and bottom of the chest, from which ROI time-series were extracted to assess how respiratory signal amplitude varied with horizontal position. 189

7.8 Positions of the columns of overlapping ROIs on the far-right (a), mid-right (b), mid-left(c), and far-left (d) of the chest, from which ROI time-series were extracted to assess variations in respiratory amplitude with vertical position. 189

7.9 Location of the 150x150 regions on the right (a) and left (b) sides of the chest. 190

7.10 Selection of overlapping ROI locations relative to the original pixels and the 10x10 *superpixels* (outlined) used in the blurred video frames. 191

7.11 The locations of the three time-points extracted during segment NVSM15-3 for analysing the spatial variations in respiratory signal amplitude. . 191
7.12 Spatial variations in respiratory signal amplitude at time-point S1. . . 193
7.13 Spatial variations in respiratory signal amplitude at time-point S2. . . 194
7.14 Spatial variations in respiratory signal amplitude at time-point S3. . . 195
7.15 Reference vital-sign values for heart rate (a), respiratory rate (b) and oxygen saturation (c) for patient NVSM15 10 minutes before and after the time-point S0 198
7.16 Reference vital-sign values for heart rate (a), respiratory rate (b) and oxygen saturation (c) for patient NVSM15 10 minutes before and after the time-point S4. 199
7.17 Colour map of respiratory signal amplitudes from ROIs on the chest of patient NVSM15 during time-point S0. 200
7.18 Colour map of respiratory signal amplitudes from region of interests (ROIs) on the chest of patient NVSM15 during time-points S1 (a), S2 (b) and S3 (c). The patient had a right-sided pneumothorax (confirmed by the X-rays around time-points S0 and S4). It is unclear whether the diaphragm had ruptured by the time the data were recorded. 201
7.19 Colour map of respiratory signal amplitudes from ROIs on the chest of patient NVSM15 during time-point S4. 202

A.1 Bland-Altman plots showing the differences between measurements from two methods against (correctly) the averages of the measurements (a) and against (incorrectly) the measurement values from one of the methods (b). 214
A.2 Incorrect Bland-Altman plots for the comparison of continuous and manual measurements of respiratory rate (a), heart rate (b) and oxygen saturation (c). 215

List of Tables

2.1 Demographic and other clinical information for patients in the CALMS-2 study. . . . 41

2.2 Overview of the number of manual vital-sign observations per patient, and completeness of the observation sets collected during the CALMS-2 study. . . . 43

2.3 Number of monitoring days and completeness of continuous vital-sign data for the "normal" and "abnormal" patient groups in the CALMS-2 study. . . . 44

2.4 Demographic and other clinical information for patients in the CALMS-3 study who had an adverse event ("abnormal" patients) and patients who did not have an adverse event ("normal" patients). . . . 49

2.5 Overview of the number of manual vital-sign observations per patient, and completeness of the observation sets collected during the CALMS-3 study. . . . 50

2.6 Summary of the vital-sign data collected during the CALMS-2 and CALMS-3 clinical studies. . . . 51

3.1 Limits of physiological plausibility for the 5 main vital signs. . . . 57

3.2 Percentiles for computing the knot positions in the restricted cubic splines logistic regression model. . . . 60

3.3 Area under the reciever operating characteristic curve results for the 5-fold feature importance analysis of each vital sign and vital-sign variability feature. . . . 63

3.4 Oxford Early Warning Score . . . 71

3.5 National Early Warning Score . . . 71

3.6 Area under the reciever operating characteristic curve results for the 5-fold cross-validation feature selection analysis on the CALMS-3 dataset. 76

3.7 Results for the performance analysis of the machine learning models and the benchmark early warning scores on the CALMS-2 dataset. . . 77

3.8 Analysis of the sensitivity, specificity and time-to-event of the machine learning models using decision thresholds that achieve the sensitivity, and the specificity, of the National Early Warning Score. . . . 79

4.1 Results for the "smoothing effect" analysis between continuous and manual measurements of respiratory rate, heart rate and oxygen saturation for "during" window sizes of 1 to 10 minutes. 94
4.2 Bias and limits of agreement for the Bland-Altman analysis comparing continuous vital-sign measurements and manual vital-sign measurements of respiratory rate, heart rate and oxygen saturation. 100
4.3 Bias and limits of agreement for the Bland-Altman analysis comparing continuous vital-sign measurements of respiratory rate, heart rate and oxygen saturation before and during the time of clinical observation. . 101

5.1 Statistical distance metrics for different methods of computing diurnal variability for heart rate, respiratory rate and oxygen saturation. . . . 123

6.1 Demographic and monitoring details of patients in the NVSM study. . 142
6.2 Details of the five vital-sign segments chosen from patients in the NVSM study for analysis of non-contact vital-sign monitoring algorithms. . . 145
6.3 ROI size and location for extracting heart rate from each segment. . . 155
6.4 Respiratory rate estimation ROI size and location for each segment. . 162
6.5 Results for the comparison of camera-derived estimates of heart rate and respiratory rate with reference values from the Philips bedside monitor for each segment. 167

7.1 Comparison of the camera-derived respiratory rate estimates from a 30x30 pixel ROI on the forehead (the "reference") with the camera-derived respiratory rate estimates from 30x30 pixel ROIs on the right and left sides of the chest. 181

List of Acronyms

ABP arterial blood pressure.

ADU analog-to-digital unit.

APACHE II Acute Physiology And Chronic Health Evaluation II.

APSF Anesthesia Patient Safety Foundation.

AUCPR area under the precision recall curve.

AUROC area under the receiver operating characteristic curve.

BMI body mass index.

CCD charge-coupled device.

CDF cumulative distribution function.

CI confidence interval.

CRF clinical research form.

DTEWS Decision-tree Early Warning Score.

DTFT Discrete Time Fourier Transform.

ECG electrocardiogram.

EHR electronic health record.

eObs electronic observation.

EWS Early Warning Score.

FDA US Food and Drug Administration.

FFT Fast Fourier Transform.

FIR finite impulse response.

FPR false positive rate.

GCS Glasgow Coma Scale.

HDU High Dependency Unit.

HRV heart rate variability.

ICU Intensive Care Unit.

IIR infinite impulse response.

IMS Integrated Monitoring System.

IP impedance pneumography.

IQR inter-quartile range.

LoA limits of agreement.

MAE Mean Absolute Error.

MET medical emergency team.

MEWS Modified Early Warning Score.

mmHg millimetres of mercury.

NEWS National Early Warning Score.

NICU Neonatal Intensive Care Unit.

pAUROC partial AUROC.

PDA Personal Digital Assistant.

PPG photoplethysmogram.

PPGI photoplethysmographic imaging.

PPV positive predictive value.

PSD power spectral density.

PTT pulse transit time.

RD respiratory depression.

RGB red green blue.

RMSE Root Mean Squared Error.

ROC Receiver operating characteristic.

ROI region of interest.

RRS rapid response system.

SaO_2 arterial blood oxygen saturation.

SNR signal-to-noise ratio.

SpO_2 peripheral blood oxygen saturation.

SQI signal quality index.

TPR true positive rate.

TTE time to event.

ViEWS VitalPac Early Warning Score.

Chapter 1

Post-operative monitoring

In this thesis we investigate improved strategies for monitoring patients in acute post-operative wards. In this chapter we review the background literature on current and emerging practices of post-operative patient monitoring. We first describe the physiology and methods of monitoring the vital signs (section 1.1). We then introduce the problem of adverse clinical events in post-operative care (section 1.2), and previous strategies to address this problem (section 1.3). We describe the use of continuous vital-sign monitoring equipment to improve post-operative patient monitoring (section 1.4). Finally, we provide an overview of the thesis (section 1.5).

1.1 Vital-sign monitoring

Monitoring the clinical biomarkers of patient health is a well-established strategy for detecting physiological deterioration in hospital care (Shoemaker et al. [1983]). Modern medicine possesses many tools to assess patients and their illnesses. Among these, monitoring of vital signs provides a basic insight into the patients overall condition and whether this condition is stable or showing signs of deterioration (Van Leuvan and Mitchell [2008]). Patterns of critical illness will often appear in the collection of diagnostic measurements well before patients show physical signs of distress (Franklin and Mathew [1994]; Hillman et al. [2002]; Schein et al. [1990]). Early identification of clinical deterioration in hospitalised patients is important to prevent subsequent cardiopulmonary arrest and reduce mortality (Beaumont et al. [2008]).

Patient monitoring strategies are designed to detect clinically important changes in physiology so that care can be escalated for deteriorating patients. For over 100 years, nurses have monitored a patient's physiology using the same vital signs: heart rate, respiratory rate, temperature, blood pressure and, in recent years, peripheral blood oxygen saturation (SpO_2) (Ahrens [2008]; Elliott and Coventry [2012]). Nurses measure vital signs intermittently, with a measurement frequency typically between 15 minutes and 12 hours, depending on the acuity level of the ward, and the clinical condition

of the patient (Evans et al. [2001]; Van Leuvan and Mitchell [2008]). High-acuity patients such as those in the Intensive Care Unit (ICU) are monitored continuously with automated measurement devices so that acute changes in physiology can be detected in real time, and specific physiological patterns, such as cardiac arrhythmias (irregular heartbeats), can be detected.

Post-operative patients are typically more acute than general ward patients and less acute than Intensive Care Unit (ICU) patients. This is reflected in the level of monitoring received in post-operative care, with most patients' vital signs measured manually every 1 to 4 hours. When clinical staff judge that a particular patient requires more intensive monitoring, the frequency of vital-sign observations can be increased to multiple measurements per hour, or continuous monitoring equipment can be used temporarily. Specific monitoring practices may vary between wards and hospitals, depending on factors such as staffing numbers and surgical specialities.

1.1.1 Physiology of the vital signs

Heart rate

The human heart beat is controlled by electrical activity localised at specific nodes within the cardiac muscle, which stimulate contractions in a coordinated manner to pump blood. The heart consists of two atria (upper chambers) and two ventricles (lower chambers) – blood enters the heart through the right atrium and passes to the right ventricle, where it is pumped into the pulmonary circulation (the lungs) to exchange carbon dioxide and receive oxygen. Oxygenated blood then enters the left atrium, passes to the left ventricle, and is pumped to the systemic circulation where the oxygen is used in the body. The heart rate measures the number of contractions of the heart per minute. For adults, normal heart rates typically range from 60 to 100 beats/minute. Bradycardia (low heart rate) and tachycardia (high heart rate) can be indicators of physiological abnormality.

Respiratory rate

The respiratory rate is the number of breaths taken per minute. Under normal conditions, the respiratory rate is automatically controlled by homeostatic mechanisms to keep the partial pressures of oxygen and carbon dioxide in the blood constant. For adults, normal respiratory rates range from 12 to 18 breaths/minute. Bradypnoea (low respiratory rate) and tachypnoea (high respiratory rate) can be indicators of physiological abnormality.

Blood oxygen saturation

Blood oxygen saturation is the proportion of oxygen-saturated haemoglobin (the protein that carries oxygen) relative to total haemoglobin in the blood. Normal blood oxygen saturation levels are 95 to 100%. Blood oxygen saturation levels below 90% (hypoxaemia) are considered abnormal, and cause hypoxia, which is when bodily tissues are deprived of adequate oxygen supply.

Blood pressure

Blood pressure is a measure of the force exerted by the blood per unit area of the arterial wall (McGhee and Bridges [2002]), and is measured in millimetres of mercury (mmHg). Blood pressure varies over the cardiac cycle and is expressed in terms of the systolic blood pressure, which is the maximum value occurring during that part of the cardiac cycle when the ventricles contract (systole), and the diastolic blood pressure, which is the minimum value during that part of the cardiac cycle when the heart fills up with blood following systole. Normal blood pressure ranges from 90 to 140 mmHg (systolic) and from 60 to 90 mmHg (diastolic). Hypotension (low blood pressure) and hypertension (high blood pressure) can be indicators of abnormal physiology.

Temperature

Thermoregulation is the part of the homeostatic mechanism that controls the body temperature at its optimal point. Normal body temperatures typically range from 36.5°C to 37.5°C, and individual body temperatures can range depending on factors such as age, sex and time of day. Differences in temperature exist between different parts of the body: 'Core temperature' is the operating temperature within the internal environment of the body including the major organs, and is maintained within narrow range. Temperatures of peripheral parts of the body (such as oral temperature) have more variance, and are around 1°C lower than core temperature. Hypothermia (low body temperature) and hyperthermia (high body temperature) can be indicators of physiological abnormality.

Other vital signs

Heart rate, respiratory rate, peripheral blood oxygen saturation (SpO_2), blood pressure and temperature are the five most common vital signs collected in hospital wards – in this thesis we refer to these as the 5 main vital signs. Neurological status is another vital sign that is commonly recorded. This is an assessment of a patient's level of

consciousness or alertness, ranging from perfectly alert (and therefore normal) to being in a state of coma (unconscious and unable to be awakened). Urine output may also be included as a vital sign to indicate renal function and fluid status (Elliott and Coventry [2012]). Finally, many clinical risk scores (such as those described in section 1.3.1) include a binary variable indicating whether a patient is (or is not) receiving supplemental oxygen support through a mask or cannula, which is typically documented during collection of the main vital signs.

1.1.2 Manual measurement methods

The standard of care for monitoring most hospitalised patients is periodic manual measurement of the vital signs by clinical staff. Below we describe standard methods of manually recording the vital signs. In addition to these methods, it is common for clinical staff to document manual vital-sign estimates by recording the value shown on the patient monitor if the patient is receiving electronic continuous monitoring (see section 1.1.3). Clinical staff may also temporarily attach continuous monitoring equipment to measure the vital signs during their manual observations (e.g. using pulse oximetry to measure heart rate and blood oxygen saturation).

Heart rate

Heart rate can be measured manually by palpating a peripheral artery (such as the radial artery on the forearm) and counting the number of beats that occur over a time window (e.g. 15, 30 or 60 seconds), then applying the appropriate multiple to obtain the average number of beats per minute.

Respiratory rate

Respiratory rate is difficult to measure clinically as methods of measurement often have poor accuracy and repeatability (Hooker et al. [1989]; Lovett et al. [2005]). The standard of care for manual measurement of respiratory rate in the Oxford University Hospitals NHS Foundation Trust (Oxford, UK) is by manually counting the number of breaths taken over 60 seconds. However, shortcuts are often taken by counting the number of breaths over a 15-second window, resulting in unusual distributions of respiratory rate recordings peaked at 12, 16 or 20 breaths/minute, as clinical staff count 3, 4 or 5 breaths and multiply by 4 to extrapolate to 1 minute (Badawy et al. [2017]). This practice exacerbates the inaccuracies in measuring respiratory rate,

as miscounting only a few breaths in a 15-second window can lead to very different recorded values.

Blood oxygen saturation

Blood oxygen saturation can be measured arterially or peripherally. Estimating the arterial blood oxygen saturation (SaO_2) involves extracting a sample of blood from the radial artery and measuring the concentration of oxygen with a blood gas analyser. However, it is far more common to estimate the oxygen saturation peripherally (SpO_2) by temporarily attaching a pulse oximeter probe to the patient's finger or earlobe (see section 1.1.3).

Blood pressure

Arterial blood pressure may be measured manually with a sphygmomanometer, which uses an occluding cuff. The cuff is placed on a patient's upper arm and inflated to occlude the artery. The auscultatory method involves inflating the cuff to fully occlude the artery then gradually deflating the cuff, listening with a stethoscope to hear the pressure at which blood starts to flow (systolic pressure) and at which there is no occlusion (diastolic pressure). The oscillometric method is performed using electronic equipment, which also inflates the cuff to above the systolic pressure and subsequently slowly deflates it, measuring the amplitude of the cardiac-synchronous oscillations in pressure, which are known as the oscillometric waveform. Signal processing algorithms estimate the systolic and diastolic pressures from the amplitude of the oscillometric waveform at various points when the cuff is deflated.

Temperature

Body temperature is difficult to measure accurately on the ward. The gold standard for measuring core temperature is with a catheter in the central circulation, but this is highly invasive and is generally only performed in acute scenarios such as the ICU (McCallum and Higgins [2012]). Rectal thermometers provide good estimates of core temperature, but obtaining measurements from the rectum is time consuming and may be unfavourable for some patients (McCallum and Higgins [2012]). Peripheral thermometers can be used to measure temperature from the mouth (oral measurement), the ear (tympanic measurement) or the armpit (axilla measurement). The accuracy and repeatability of peripheral thermometers is generally poor (Niven et al. [2015]), but they are commonly used for routine measurements due to their ease of use.

Neurological status

Neurological status is commonly measured with categorical scales that encode subjective assessments of a patient's level of consciousness. The Glasgow Coma Scale (GCS) (Teasdale and Jennett [1974]) is a numeric system that assigns scores based on a patient's eye response, verbal response and motor response, with a score of 3 indicating a deep level of unconsciousness (or death) and a score of 15 indicating full consciousness. The AVPU scale is a simplification of the GCS, with which clinical staff measure a patient's response to stimuli (Mackay et al. [2000]). Possible outcomes that are tested for in order of increasing severity are Alert (fully awake), Verbal (responds when spoken to), Pain (responds to pain stimuli) and Unresponsive. A patient's AVPU score is recorded as the letter corresponding to their test outcome (e.g. A for Alert).

1.1.3 Electronic measurement methods

When patients require more intensive monitoring than manual monitoring methods allow, continuous vital-sign monitoring may be achieved with electronic monitoring equipment, using input sensors to collect time-series of physiological measurements. In most cases, the time-series are waveforms representing fluctuations in physiological parameters due to processes such as the breathing or cardiac cycles. Continuous monitors typically use clips and sticky patches to secure the sensors to the patient's body. High-acuity patients may require monitoring with invasive methods, which generally involve inserting a cannula into the bloodstream. Here we describe several standard continuous monitoring methodologies.

Electrocardiography

The electrical activity of the heart can be measured with an electrocardiogram (ECG), using electrodes placed on the patient's limbs and the surface of the chest. electrocardiograms (ECGs) operate by recording the changes in electrical potential generated by the changing polarity of the cells within the heart's atria and ventricles. ECGs can be used to obtain information about the function of the heart, often with the purpose of diagnosing heart problems such as myocardial infarction or cardiac arrhythmia. Continuous ECG monitoring is the standard of care for cardiac patients and patients in the ICU. The ECG waveform consists of several waves, which characterise each part of the heartbeat. The component waves of the ECG waveform are labelled alphabetically from P to T, which describe the following physiological phenomena:

- P-wave: Atrial depolarization that induces atrial contraction, which can be seen as the first wave on the ECG. Depolarization normally begins in the high right atrium and then travels through to the left atrium. Deviations from this mechanism result in P waves with abnormal morphology.
- Q-wave: Septal depolarization, which can be seen as any initial downward deflection after the P-wave, if it is visible at all.
- R-wave: The largest upward deflection on the ECG waveform caused by early ventricular depolarization, which induces ventricular contraction.
- S-wave: The first negative deflection after the R-wave, caused by late ventricular depolarization inducing late contraction of the rim of the ventricular muscle.
- T-wave: The final major deflection on the ECG caused by repolarization of the ventricles, which induces ventricular relaxation.

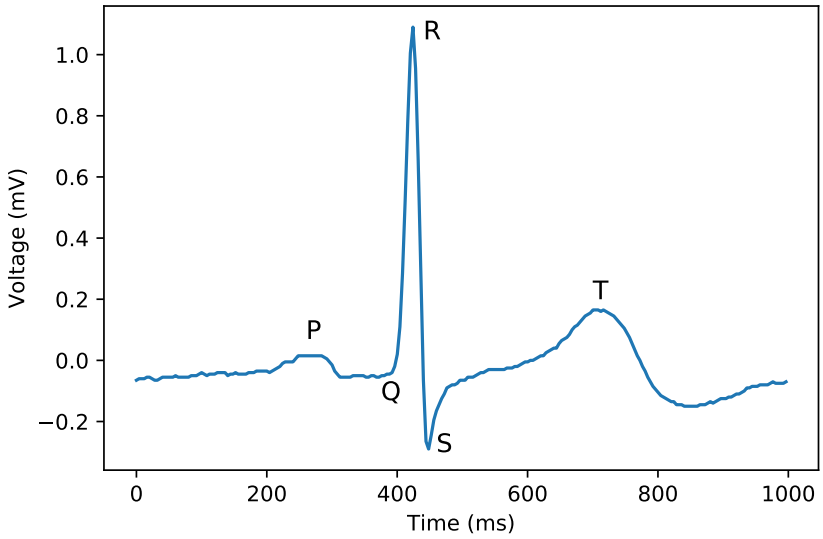


Figure 1.1: An ECG waveform for one heart beat, showing the electrical fluctuations measured by the ECG, with labels (P to T) of the component waves.

The location of the component waves of an example ECG waveform is shown in Figure 1.1. The main waveform deflection is known as the QRS complex, which constitutes the waveform section between the Q and S waves. There are tell-tale waveform morphology abnormalities that characterise arrhythmias, which can be detected automatically or through observation. The heart rate can be estimated from

the ECG waveform using a number of signal processing methodologies, the simplest of which is simply counting the number of R peaks in a 60-second window.

Arterial blood pressure (ABP) monitoring

Accurate measurement of blood pressure is sought during haemodynamic monitoring, which is undertaken by inserting a catheter into an artery, usually on the wrist. The catheter is connected to a fluid-filled tubing system, which is designed to transmit the mechanical pressure signal generated in the arterial system during cardiac contractions. An electronic pressure transducer then converts the mechanical pressure signal into a digital signal, which is displayed as the arterial blood pressure (ABP) waveform.

ABP monitoring is undertaken in the ICU for critically ill patients who require invasive monitoring to provide reliable continuous measures of their haemodynamic status and cardiovascular function. The ABP waveform displays repeating units that characterise the physiological fluctuations that occur over the course of one heart beat. For ABP, the repeating unit is the pressure pulse generated by left ventricular contraction. The pressure pulse morphology begins with an increasing gradient at the onset of systole, and ends with a decreasing gradient during diastole. The component parts of the ABP waveform describe the following physiology:

- Systolic uptake: The pressure wave during ventricular ejection.
- Peak systolic pressure: Maximum pressure generated during ventricular ejection.
- Systolic decline: The decline in arterial pressure as ventricular contraction ends.
- Dicrotic notch: Closure of the aortic valve at the onset of ventricular diastole.
- Diastolic runoff: Further decline in arterial pressure as ventricular contraction ends.
- End-diastolic pressure: The pressure of the blood in the left ventricle at the end of diastole.

The components of an example ABP waveform are shown in figure 1.2. The morphology of the ABP waveform is of clinical interest, as abnormalities in the morphology can indicate issues in haemodynamic status (Syeda et al. [2003]) such as hypertension or atherosclerosis (plaque build up in the arteries). Systolic and diastolic blood pressure are measured from the ABP waveform as the peak systolic pressure and the end diastolic pressure, respectively.

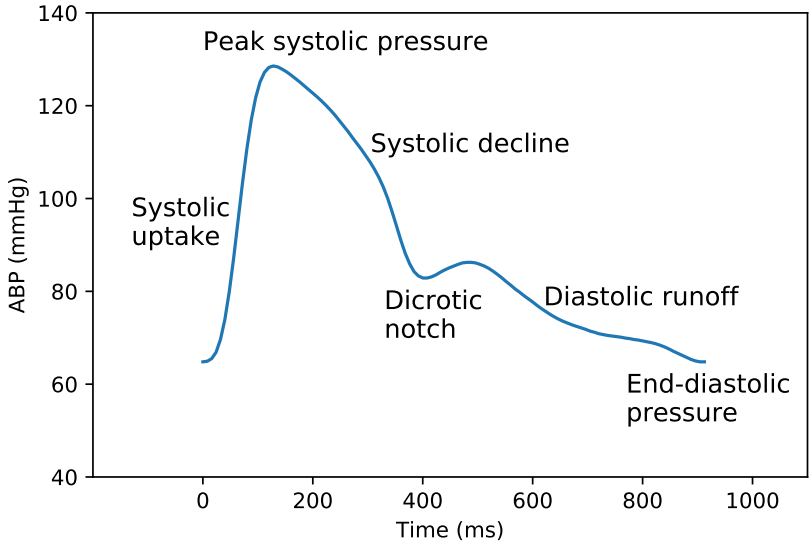
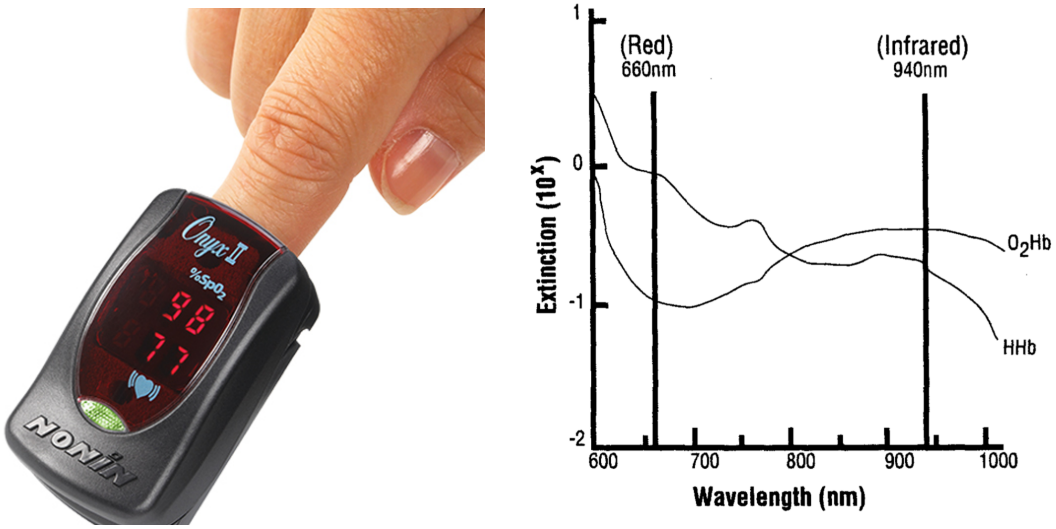


Figure 1.2: An ABP waveform for one heart beat, showing the blood pressure fluctuations measured during haemodynamic monitoring. The component parts of the waveform are labelled.

Pulse oximetry

A pulse oximeter is a medical device used to provide non-invasive continuous estimates of SpO_2. Pulse oximetry attaches a sensor to a well-perfused part of the body, such as a finger, earlobe or the forehead (Maeda et al. [2011]; Tamura et al. [2014]), and uses light to measure the absorption of oxygenated and deoxygenated haemoglobin. In transmission mode (e.g. with a finger sensor) light is passed through the body segment and the transmitted light is measured with a photodetector, while in reflectance mode the skin is illuminated and the reflected light is measured with a photodetector. Two light sources are usually used – one with a red-light wavelength of 600-700 nm, and another with an infrared wavelength of approximately 900 nm. The absorption of oxygenated and deoxygenated haemoglobin significantly differ at these wavelengths, allowing the estimation of SpO_2 from ratios of the amplitude of the cardiac-synchronous signal at the two wavelengths. Transmission based pulse oximetry is most common, using a small probe that clips on to the end of a finger. In figure 1.3a we show a typical pulse oximeter, and in figure 1.3b we show the light absorption curves for oxygenated and deoxygenated haemoglobin.



(a) Source: Nonin **(b)** Source: Sinex [1999]

Figure 1.3: A typical finger pulse oximetry sensor (a). The absorption spectra of oxygenated (O_2Hb) and deoxygenated (HHb) haemoglobin, with lines at typical operating wavelengths of red and infrared light sources (b).

Photoplethysmography

The photoplethysmogram (PPG) is the cardiac-synchronous waveform measured at either of the wavelengths in a pulse oximeter. It measures the volumetric changes in the blood pressure pulse over the cardiac cycle using optical methods. The volumetric changes in the pressure pulse distend the subcutaneous vasculature, giving rise to the cardiac-synchronous changes in light absorption measured with the photodetector. Other methods of measuring the photoplethysmogram (PPG) include non-contact photoplethysmographic imaging (PPGI) with a video-camera, which is described in section 1.4.4.

The heart rate can be derived from the PPG waveform (e.g. by counting the number of beats in a minute). Indeed, pulse oximetry is commonly used to continuously monitor heart rate, or is attached temporarily to a patient to collect a manual reading of heart rate and SpO_2. The PPG waveform is also modulated by other physiological systems, such as the mechanics of breathing (Meredith et al. [2012]). Algorithms have been developed to derive respiratory rate from the PPG waveform (Charlton et al. [2017]; Fleming and Tarassenko [2007]; Karlen et al. [2013]; Nilsson et al. [2000]; Pimentel et al. [2015]), but this use of pulse oximetry has not been sufficiently validated for widespread clinical use. We show an example 15-second window of PPG waveform data in figure 1.4.

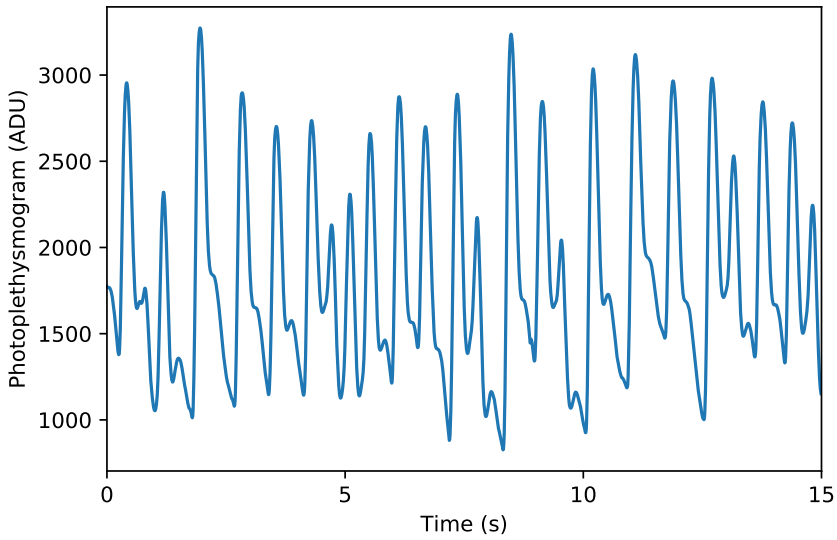


Figure 1.4: A 15-second window of PPG waveform, showing the volumetric changes in the pulse pressure. The dominant waves are cardiac synchronous. The waveform amplitude is modulated by a sinusoid with a much longer period than the cardiac rhythm, which is caused by breathing. The dicrotic notch (see figure 1.2) is the small peak in (most of) the PPG waveforms between the main peak and the trough. analog-to-digital unit (ADU) is an arbitrary unit for an analogue sensor signal that has been converted into a digital format.

Impedance pneumography

Impedance pneumography (IP) is a non-invasive method of continuously measuring respiratory rate. The physiological phenomenon underlying this technique is the relationship between the gas volume of the chest and transthoracic electrical impedance – when one increases, so does the other (Grenvik et al. [1972]). Thus, measuring the frequency of fluctuations in transthoracic electrical impedance yields an estimate of the respiratory frequency. This is achieved by attaching electrodes to the thorax and applying a high-frequency alternating current. The current induces a potential difference between the probes, which is measured by a voltmeter. This measured voltage is directly proportional to the electrical impedance. The IP waveform is therefore calculated using Ohm's law, whereby the impedance is the complex ratio of the measured voltage to the applied current. An example segment of IP waveform over several breaths is shown in figure 1.5.

IP is a popular choice of continuous respiratory rate measurement in critical care because the measurement electrodes are the ECG electrodes, through which

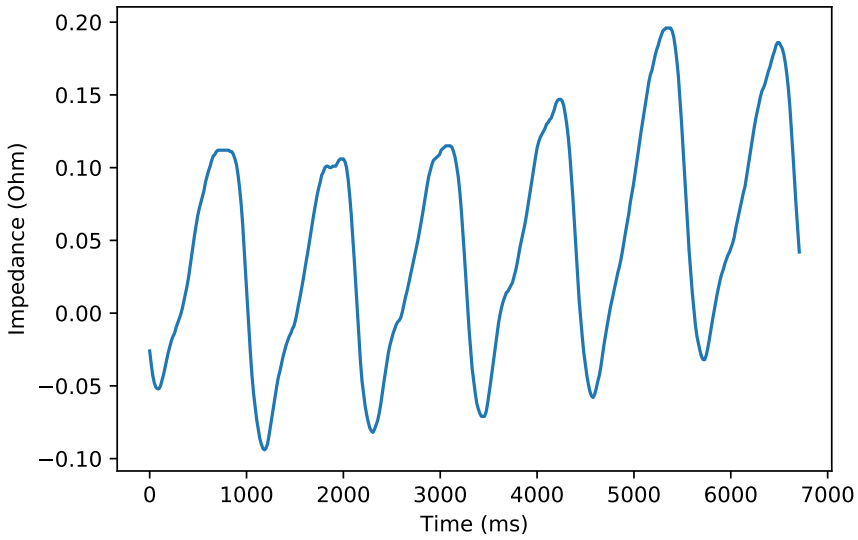


Figure 1.5: An impedance pneumography waveform, showing the fluctuations in transthoracic impedance over the course of six breaths.

the high-frequency current is passed. Furthermore, IP is non-invasive compared to capnography, for example, which requires the patient to wear a face mask to measure the concentration of carbon dioxide in the exhaled air. It is also a direct measure of respiratory rate as the dominant frequency of the waveform is respiratory-synchronous, and thus is more reliable than methods that derive respiratory modulation frequencies indirectly from cardiac-dominant signals such as PPG. Respiratory rate can be derived from the IP waveform by counting the respiratory peaks in a 60-second window.

Other vital signs

As described in section 1.1.2, core body temperature can be measured accurately and continuously using invasive methods by inserting a catheter into a central location such as the pulmonary artery. Unlike other methods described here, this does not generate a waveform as temperature does not vary over short time scales. Neurological status is not measured continuously as categorical scales such as AVPU (described in section 1.1.2) require subjective assessment.

1.1.4 Summary

There are many ways to measure the vital signs using both manual and continuous methods. In post-operative care, the most appropriate method for a particular patient

represents a trade-off between accuracy, invasiveness, cost, comfort and time. Currently, most patients on post-operative wards are monitored manually by clinical staff during routine observations. Continuous technology has traditionally been used in the ICU to monitor critically ill patients, but may also be used for short periods in post-operative care for the most acute patients. New monitoring methods continue to be developed to improve measurement accuracy, reduce invasiveness and measure new physiology (see section 1.4).

1.2 Adverse events in post-operative care

In the last three decades clinical efforts to recognise the deteriorating patient have increased over time (Hillman et al. [2001b]). There is a strong consensus that many in-hospital adverse events could still be prevented with earlier detection of physiological deterioration, followed by appropriate intervention (Adam and Odell [2005]; Berwick et al. [2006]). An adverse event can be defined as any instance of deterioration of the health of a patient, which may occur due to a variety of major complications. For example, Pimentel [2015] defined an adverse event as either a cardiac arrest, an unexpected admission to the ICU, or death on the ward. Patients most at risk of adverse events include patients who have recently had a surgical or medical ICU admission or patients who are recovering from critical illness (Beaumont et al. [2008]).

1.2.1 The evidence and cost of avoidable errors

By assessing vital-sign records of patients who suffered adverse clinical events, researchers have found that symptoms of physiological deterioration have often gone undetected (Kause et al. [2004]; McQuillan et al. [1998]). For example, an international multi-centre study showed that 80% of cardiac arrests, 55% of unanticipated ICU admissions, and 55% of deaths had early signs of abnormal physiology (Kause et al. [2004]). While care pathways since this study are likely to have improved, more recent evidence suggests that abnormal vital signs still precede most adverse clinical events (Andersen et al. [2016]; Bingham et al. [2018]; Fujiwara et al. [2016]; Sprogis et al. [2017]). The consequence of the failure to detect physiological deterioration is significant, as adverse clinical events incur a high cost. There is a human cost with increased morbidity and mortality rates, as well as an economic cost due to increased acuity of care and longer stays on the ward. We describe previous studies assessing the clinical antecedents to adverse events and the cost of adverse events as follows.

Cardiac arrest

Cardiac arrest is the sudden stop in normal blood circulation due to the failure of the ventricles of the heart to contract effectively (Kasper and Harrison [2005]). The resulting lack of blood supply causes cell death from oxygen starvation. The lack of oxygen supply to the brain, known as cerebral hypoxia, causes loss of consciousness and cessation of breathing (Hann [2008]). Clinical antecedents to cardiac arrest are well known, with the following physiological alterations preceding arrest: abnormal respiratory rate, heart rate and temperature, reduced systolic blood pressure, reduced SpO_2, chest pain, poor neurological function, poor metabolic function and nurse or doctor concern (Hillman et al. [2001b]; Hodgetts et al. [2002a]; Schein et al. [1990]; Smith and Wood [1998]). The majority of these abnormalities can be quantified with vital-sign values or neurological status charts (see section 1.1). In a recent study, Andersen et al. [2016] found that out of 7,851 patients who suffered cardiac arrest in a United States clinical registry, 59.4% had abnormal vital signs before the arrest.

Analysis of the UK National Cardiac Arrest Audit database has shown that unexpected in-hospital cardiac arrests on the ward have an occurrence rate of 1.6 arrests per 1000 hospital admissions (Nolan et al. [2014]). If untreated, cardiac arrest will cause death in minutes. Indeed, unadjusted survival to discharge is only 18.4% (Nolan et al. [2014]). Hodgetts et al. [2002b] estimated that up to 23,000 cases of cardiac arrest may be avoided in the UK each year, although it is likely that this number has since reduced with improved processes for identifying physiological deterioration (such as McGregor et al. [2017]).

Unexpected admission to the ICU

Kause et al. [2004] found that up to 54% of patients receive suboptimal care prior to ICU admission, with physiological abnormalities being documented up to 24 hours in advance without inducing an appropriate response. McQuillan et al. [1998] reported that 66% of patients showed visible signs of instability in the 12 hours preceding admission to ICU. More recent studies confirm that abnormal vital signs strongly predict unexpected ICU admissions (Barfod et al. [2012]). Clinical antecedents to unplanned admissions include abnormal heart rate, respiratory rate, and blood pressure, impaired neurological function and nurse or doctor concern (Hillman et al. [2002]).

Patients with unplanned admissions to the ICU tend to have a poor prognosis for survival, with higher mortality rates than patients with planned admission (52% vs 35%) (Goldhill and Sumner [1998]; Hillman et al. [2002]; McGaughey et al. [2007]).

Even patients who survive have a long length of stay, with one study showing that patients admitted from the general ward remain in hospital, on average, for 23.5 days (Hillman et al. [2002]). Several studies have shown the importance of early admission to the ICU, with delays leading to increased ICU mortality (Cardoso et al. [2011]; Chalfin et al. [2007]; Young et al. [2003]), increased hospital mortality (Boniatti et al. [2014]; Johnston et al. [2015]; Phua et al. [2010]), increased length of ICU stay (Chalfin et al. [2007]), and longer duration of mechanical ventilation (O'Callaghan et al. [2012]).

Death on the ward

While most patients are expected to make a recovery during their hospital stay, there is still a high number of unexpected deaths that occur. Estimates in the annual number of preventable in-hospital deaths in the United States range from 100,000 to 400,000 (Barach and Small [2000]; James [2013]). These deaths may be caused by diseases such as sepsis, which occurs due to bacterial infection in the blood stream and can be acquired during treatment. Sepsis has a high burden amongst post-operative patients in particular (Vogel et al. [2010]). Studies have indicated that up to 70% of unexpected deaths are preventable, occurring due to negligence, complications in healthcare management or substandard care, rather than untreatable illness (Brennan et al. [1991]; Wilson et al. [1995]). Hillman et al. [2001b] reported that 30% of unexpected deaths had severely abnormal physiology documented, with the most common antecedents being hypotension and tachypnoea. Sepsis, while commonly diagnosed using laboratory tests, is often preceded by abnormalities in the five main vital signs (Nguyen et al. [2014]).

1.2.2 Rapid response systems and new approaches to care

The problem of adverse clinical events has been summarised with the concept of "failure to rescue", which describes an inability to quickly recognise signs of physiological deterioration and provide treatment to prevent avoidable deaths (Silber et al. [1992]). The rate of "failure to rescue" has been used as an indicator of patient safety and quality of hospital care (Schmid et al. [2007]). To reduce rates of adverse clinical events and therefore the burden of "failure to rescue", new approaches to caring for the deteriorating patient have been developed. These approaches have many different names, including medical emergency team (MET), critical care outreach, and track and trigger system – here we use the term rapid response system (RRS).

RRSs are used in hospital wards to prompt an elevation in acuity of care when

a patient shows signs of physiological deterioration (Armitage et al. [2007]; Gold and Simmons [2006]; NICE [2007]). RRSs are protocolised responses designed to enforce consistency of care, such that all instances of physiological deterioration are promptly detected and treated appropriately. The fundamental structure of RRSs can be disaggregated into an afferent limb, an efferent limb, and administrative and data analysis limbs (DeVita et al. [2006]; Winters et al. [2013]). The afferent limb refers to the aspects of the system associated with the detection of at-risk patients and the triggering of a response (see section 1.3). The efferent limb refers to the implementation of the response, which may include increased monitoring, visits from nurses or clinicians, visits from specialists (such as ICU physicians) or transfers to more acute wards. Policy makers and healthcare providers have recognised the opportunity for quality improvement in the care pathway through targeted response to physiological deterioration, and have advocated for the implementation of RRSs (Adam and Odell [2005]; Berwick et al. [2006]; NICE [2007]). As such, RRSs are now common place in modern hospital wards.

Despite their widespread use, RRSs do not have a strong evidence base (McGaughey et al. [2007]). Early studies on the use of RRSs showed strong results, with a 17-50% reduction in mortality for the intervention group (Bellomo et al. [2004]; Buist et al. [2002]; DeVita et al. [2004]). However, these trials were small, single-centre or unrandomised, and thus lacked the robust methodologies required to provide gold standard evidence. Subsequently, the MERIT study – an international multi-centre randomised control trial – showed no mortality benefit in the use of an RRS (Hillman et al. [2005]). However, the RRS was called in only 30% of cases when RRS activation criteria were fulfilled, suggesting that opportunities for early intervention were missed. Furthermore, the MERIT study was not sufficiently powered to detect a significant change in mortality (Chrysochoou and Gunn [2006]). However, the medical community has formed a consensus in support of RRSs (DeVita et al. [2006]; NICE [2007]), recommending that hospitals implement and locally assess an RRS.

In recent years, investigators have turned to systematic reviews to synthesise the large number of published studies that have collectively reported mixed results on the benefits of RRSs (Chan et al. [2010]; Johnston et al. [2015]; Maharaj et al. [2015]; McNeill and Bryden [2013]; Tirkkonen et al. [2017]; White et al. [2015]; Winters et al. [2013]). All reviews noted that there was significant heterogeneity between studies in terms of patient and hospital characteristics, study designs, clinical processes and outcome measures. There is a consensus that the introduction of an RRS is associated with a reduction in cardiopulmonary arrests (Chan et al. [2010]; Maharaj et al. [2015];

Winters et al. [2013]), but there were mixed conclusions on whether RRSs reduce hospital mortality (Chan et al. [2010]; Maharaj et al. [2015]; Winters et al. [2013]). Trinkle and Flabouris [2011] found that failures in the afferent limb are common and are associated with poor patient outcomes. Likewise, Johnston et al. [2015] reviewed factors associated with "failure to rescue", finding that delays in the escalation of care are associated with poor outcomes.

1.2.3 Summary

Adverse clinical events continue to occur on hospital wards, although the balance of evidence suggests that RRSs have helped to reduce this burden. While RRSs continue to gain acceptance, improvements in their implementation are still clearly required. In this thesis, we focus on the afferent limb of RRSs in terms of the detection of physiological deterioration with patient vital-sign data. The remaining sections of this chapter provide a review of current and emerging methods for detecting the deteriorating patient.

1.3 Detecting physiological deterioration

Critical settings in any RRS are the criteria used to judge when a patient requires a higher acuity of care. A poorly-designed set of RRS calling criteria can let down even the slickest of response teams. Indeed, failure of the afferent limb is recognised as the single most significant contributor of failing to rescue a patient (Adam and Odell [2005]; Findlay et al. [2012]). RRS calling criteria can require both subjective judgement, whereby clinical staff use their expertise to assess patient condition, and objective judgement, where a patient's physiological measurements and quantified behavioural symptoms are assessed against preset criteria. Policy guidelines advocate for a strong emphasis on objective assessment, as institutions have found that RRSs are under-utilised when only subjective criteria are used (DeVita et al. [2006]). Furthermore, improved vital-sign monitoring and recognition of clinical crisis may have outcome benefits (DeVita et al. [2010]). Early Warning Scores (EWSs) have emerged as the canonical form of objective RRS calling criteria, which we describe as follows.

1.3.1 Early warning scores

An EWS is a clinical algorithm that quantifies a patient's level of physiological abnormality (Morgan et al. [1997]). EWSs apply numerical scores based on values of

physiological biomarkers (including the 5 main vital signs) and alert clinical staff to escalate care when the scores reach a specified threshold. Over time, different designs of EWS have emerged, including single-parameter systems, multiple-parameter systems, and data-driven systems. The performance of EWSs can be measured *in-silico* using methods described in chapter 3. High-performing EWSs are desirable, as they allow nursing staff to escalate care in a timely manner for those experiencing physiological deterioration, while avoiding unnecessarily high workloads due to false alarms.

Single-parameter systems

The most simple form of EWS are single-parameter systems, which require only one physiological parameter to depart from normality to prompt an escalation in care (Bellomo et al. [2004]; Buist et al. [2002]; DeVita et al. [2004]; Hillman et al. [2001b]; Lee et al. [1995]). For example, the EWS designed by Hillman et al. [2001a] had escalation criteria such as a respiratory rate greater than 36 breaths/minute or a heart rate less than 40 beats/minute. First introduced three decades ago as the trigger criteria for the original RRS study (Lee et al. [1995]), there are now over 39 unique versions of single-parameter algorithms in current use (Smith et al. [2008b]). All systems identified use heart rate, respiratory rate and blood pressure measurements, with many also using temperature, neurological status and urine flow. More subjective evaluation criteria, such as cause for concern, are also included in some implementations (Parissopoulos and Kotzabassaki [2005]).

An advantage of single-parameter systems lies in their simplicity, as more complex systems can lead to errors in documentation (Subbe et al. [2007]) and thus can affect patient care (Clifton et al. [2015]). However, the growing use of computer documentation systems, which are shown to reduce errors (Prytherch et al. [2006]), may nullify the case for simplicity. The performance of single parameter systems is both variable and poor, despite their widespread usage in modern hospitals (Cretikos et al. [2007]; Smith et al. [2008b]).

Multiple-parameter systems

Multiple-parameter EWSs require two or more parameters to exceed predetermined limits in order to trigger a response. For example, the National Early Warning Score (NEWS) (McGinley and Pearse [2012]) assigns a score of 3 for heart rate values below 40 beats/minute and a score of 2 for heart rate values between 40 and 50 beats/minute. Similar scores are assigned to all vital signs, with scores associated with vital-sign

ranges according to their level of physiological abnormality. For a given set of vital-sign observations, the scores for each vital sign are found and are summed to form the overall EWS value. The escalation criterion is therefore a threshold on the total EWS value. Over thirty-three different implementations of multiple-parameter EWSs have been described in the literature (Smith et al. [2008a]), with differing structure and performance across all variants. EWSs are recommended to include heart rate, respiratory rate, systolic blood pressure, neurological status, SpO_2 and temperature (NICE [2007]), and recent implementations of EWSs have included an additional score (of 2 or 3) for patients receiving supplemental oxygen support (McGinley and Pearse [2012]; Prytherch et al. [2010]).

Multi-parameter EWSs are often used in combination with single-parameter systems such that there will be escalation criteria for combinations of vital signs as well as separate criteria for individual vital signs (Sharpley and Holden [2004]). EWSs also commonly have graded responses, whereby lower scores may indicate a need for increased monitoring frequency by nursing staff, while a higher score may indicate a need for assessment by a consulting critical care physician (McGinley and Pearse [2012]). The use of multiple parameter EWSs with a graded response is common in the UK (Gao et al. [2007]), and is recommended by policy groups (NICE [2007]). The UK has a National Early Warning Score (NEWS) that is recommended by the Royal College of Physicians (McGinley and Pearse [2012]).

Multiple-parameter EWSs have been shown to positively discriminate between normal and abnormal physiology (Cuthbertson et al. [2007]; Gardner-Thorpe et al. [2006]; Priestley et al. [2004]). Their introduction has had other benefits as well, such as increasing the frequency of respiratory rate measurement in a variety of settings(McBride et al. [2005]; Odell et al. [2007]). However, the problem of avoidable in-hospital deaths persists (Curry and Jungquist [2014]; Galhotra et al. [2007]) and improvements in the afferent limb of RRSs are necessary to meet clinical needs (Bobay et al. [2008]; DeVita et al. [2006, 2010]; Gao et al. [2007]; McGaughey et al. [2007]; Romero-Brufau et al. [2014]).

1.3.2 Data driven approaches

Single and multiple parameter EWS systems are designed heuristically with clinical experts specifying the ranges of increasing abnormality for each vital sign, as well as the scores specifying the importance of the range (Prytherch et al. [2010]). However, this method of designing an EWS is limited (Kyriacos et al. [2011]; Taenzer et al.

[2011]) as the discrete vital-sign ranges and integer scores will not be optimal, and different populations and hospital wards may have different EWS requirements. The development of data-driven approaches to EWSs addresses these limitations, with the potential to deliver improved early warning of physiological deterioration. Data-driven approaches use physiological data collected from patients to derive or learn the optimal coefficients for an EWS rather than using rules based on clinical experience. Here we describe basic methods of using data to derive ranges and scores for multiple-parameter EWSs, as well as methods based on machine learning.

Basic methods

The most basic method of designing an EWS based on patient data is to retain the fundamental structure of a traditional EWS, whilst deriving the vital-sign ranges and associated scores from the dataset. Tarassenko et al. [2011] developed a "centile-based" EWS by deriving vital-sign ranges for heart rate, respiratory rate, SpO_2 and systolic blood pressure based on percentiles of a large dataset acquired from 863 acutely ill patients. Ranges based on the 1st and 99th percentiles were assigned a score of 3, the 1st to the 5th and 95th to 99th percentiles were assigned a score of 2, and the 5th to the 10th and 90th to 95th percentiles were assigned a score of 1, with all other percentiles assigned a score of 0. A modified version of this EWS is used as the standard of care in the Oxford University Hospitals NHS Trust.

Badriyah et al. [2014] used a decision tree algorithm to automate the design of an EWS from a database of 198,755 vital-sign observation sets, which they called Decision-tree Early Warning Score (DTEWS). A decision tree algorithm is a type of machine learning model (see below) that splits the data into nested trees of conditional statements that dictate the predicted outcome for a set of values. In the case of DTEWS, decision tree analysis was used to find the optimal vital-sign ranges (tree branches) for predicting whether a vital-sign observation set preceded an adverse clinical event, and the odds ratio at each tree node was used to specify the scores associated with the vital-sign ranges. Analysis of DTEWS found that it provided good discrimination for predicting adverse clinical events. Interestingly, the vital-sign ranges in DTEWS are similar to those in NEWS, which were designed heuristically.

There are several limitations in the use of basic data-driven EWSs, such as step changes in the predicted likelihood of an adverse event at the boundaries of the vital-sign ranges. For example, in NEWS heart rates below 40 beats/minute have a score of 3, while heart rates above 40 beats/minute have a score of 2. This suggests that heart rates of 41 and 43 beats/minute are equally likely to lead to an adverse

event, while heart rates of 39 beats/minute are 1.5 times as likely. This is clearly a flaw, and it would be more principled if there were a smooth relationship between vital-sign values and the likelihood of an adverse event. Furthermore, basic methodologies do not effectively model interactions between vital signs, such as the inverse correlation between respiratory rate and SpO_2, during times of cardiorespiratory problems. Basic methodologies retain the simplicity of EWSs, which can be calculated manually at the bedside, but this is of limited benefit as the introduction of digitised vital-sign documentation systems enable complex calculations to be performed automatically. Machine learning based EWSs can be designed to detect physiological deterioration without these limitations, thus potentially performing better than traditional EWSs and basic data-driven EWSs.

Machine learning

Machine learning is the study of algorithms that learn from, and make predictions on, data. When used for pattern recognition, machine learning algorithms are trained to associate the inputs and outputs of a set of training examples, thus learning to recognise patterns in the data. The structure of a machine learning algorithm is a mathematical model, with behaviour specified by a set of free parameters. During training, the free parameters of the machine learning model are adjusted to map features of the input set to the required value of the corresponding output set, in a process known as supervised learning. Supervised machine learning algorithms may be used for classification, where there is a discrete set of possible outcomes, known as classes, or for regression, where the output is continuous. The performance of a supervised machine learning algorithm is tested with exposure to new data, whereby the outcomes of the dataset are predicted from the inputs, and are compared with the true outcomes. Machine learning has been applied to speech recognition, computer vision, bioinformatics, finance, and medical problems (Bishop [2006]), and will continue to improve in performance with the increasing availability of large datasets.

EWSs can be framed as binary classification tasks, with the goal of predicting whether a set of vital-sign observations precedes an adverse clinical event or not. In this framework, standard classification algorithms can be used to train an EWS, using the vital signs as input features and labelling the data as “abnormal” if it preceded an adverse event or “normal” otherwise. For example, Churpek et al. [2014] used a multivariate classification algorithm to develop an EWS for general ward patients. Binary classification algorithms perform well in this setting when there are sufficient examples of both “normal” and “abnormal” data. However, examples of specific

diseases are relatively rare in most medical datasets (Pimentel et al. [2014]) and this can indeed be the case in the context of EWSs, as most patients (80 to 90%) recover on the ward without suffering an adverse event. In the absence of large datasets there may be insufficient examples of "abnormal" data to effectively train an algorithm to recognise patterns of both "normality" and "abnormality", and thus methods such as "novelty detection" may be more appropriate (Pimentel et al. [2014]).

Novelty detection methods, also known as "anomaly detection" or "one-class classification", model the distribution of the most common data class, and perform classification by recognising that data in the less common class are "novel" and is unlikely to have been generated from the modelled distribution (Markou and Singh [2003]). Tarassenko et al. [2006] designed an EWS using novelty detection by modelling the probability distribution of heart rate, respiratory rate, SpO_2, temperature and blood pressure in an acute care population, using data acquired from physiological monitors. The model, built using a kernel density estimator (Parzen [1962]), recognised abnormal physiology as novel data in the limits of the distribution. Further work has extended this modelling strategy to build more sophisticated kernel density estimation based EWSs (Hann [2008]; Pimentel [2015]; Pimentel et al. [2013a]).

Machine learning has been a popular tool for developing risk prediction algorithms in wards such as the ICU where patients are monitored with high-frequency continuous monitoring equipment. Large databases collected in the ICU such as the MIMIC databases (Johnson et al. [2016]; Saeed et al. [2011]) allow sophisticated models to be used. For example, hidden Markov models have been used to model physiology as a generative linear dynamical system, with state properties that are dependent on the underlying patient condition, and outputs as measured during physiological monitoring (Quinn and Williams [2007]). This framework allows the detection of abnormal physiology, as the conditional probability of the sequential data streams will be low when the generative process is unknown to the system, or, in other words, when "novel" data are presented. Variants of this modelling strategy have been used including factorial switching linear dynamical systems models to infer the presence of signal artefact (Quinn et al. [2009]). Bayesian time series modelling methods have been deployed to identify recurring dynamic motifs in physiological data streams (Fox et al. [2009]), including the use of hierarchical dirichlet processes to identify heart rate dynamics Saria et al. [2010a], and autoregressive hidden Markov models (Li-wei et al. [2012]) to identify blood pressure dynamics. Autoregressive hidden Markov models have been used for the early detection of neonatal sepsis, whereby the Markovian system dynamics are modelled differently for all possible latent states, including sepsis,

stable health, and periods of sensor artefact (Stanculescu et al. [2014]).

Machine learning with high-dimensional neural networks, known as deep learning, uses a model architecture inspired by the human brain to achieve state of the art performance in many benchmark computer science tasks (Goodfellow et al. [2016]; LeCun et al. [2015]). Recently deep learning has been applied to derive highly accurate models for medical risk prediction tasks (Beaulieu-Jones et al. [2018]; Harutyunyan et al. [2017]; Purushotham et al. [2018]). To avoid overfitting model parameters to the variance in a dataset, deep learning models need to be trained on large databases so that they generalise well to out-of-sample data.

A major disadvantage of complex models (e.g. deep learning) is that these models are "black boxes" that do not easily allow interpretation of the reasons behind the model's decision (Miotto et al. [2017]). This presents a challenge for clinical decision support tasks, as it may be difficult for clinical staff to design appropriate interventions in situations when it is unclear why the model has estimated that a patient is experiencing physiological deterioration. Future work on EWSs should focus on the addition of complexity while maintaining clinical interpretability.

1.4 Continuous monitoring outside of critical care

While machine learning may improve the prediction of adverse clinical events from vital-sign measurements, the timeliness of EWSs is still limited by the use of intermittent vital-sign observations. With measurement intervals of 4 hours, patients can deteriorate significantly before clinical staff become aware of the patient's condition. Even an algorithm that could perfectly identify abnormal vital signs would miss instances of deterioration in this time between measurements. Thus, to continue improving the performance of EWSs, the monitoring coverage of patients must be increased.

The monitoring coverage of patients could be improved by requiring nursing staff to measure vital signs more frequently. By reducing the time between observations, clinical staff could be alerted earlier in the progression of physiological deterioration, potentially enabling the prevention of some adverse events. Yet it is hard to define the optimal measurement frequency for any patient or population (DeVita et al. [2010]). For example, Petersen et al. [2016] showed no significant benefit to increasing the monitoring frequency of acute medical and surgical patients from twice to three times per day. If four or more vital-sign observations were taken daily, rather than three times, would this have changed the results of this study? This hypothesis could be tested with any arbitrary number of comparative studies, yet none are likely to

find sufficient evidence to justify the increase in clinical workload that is required to record more vital-sign observations. Indeed, many nursing activities are already neglected due to high clinical workloads, despite the fact that inadequate nursing cover is associated with worse patient outcomes (Ball et al. [2014]; Needleman et al. [2002, 2011]). Improvements in patient monitoring must be found within available clinical resources, which motivates the use of continuous monitoring technology.

1.4.1 The clinical need for continuous monitoring

Surveillance monitoring

Continuous vital-sign monitoring has been proposed as a potential "safety net" for patients at risk of deterioration outside of the ICU. Traditionally, continuous monitoring equipment is reserved for critical care patients, who require near-constant attention due to their physiological instability. However, there is a growing consensus that some wards below the acuity of the ICU would also benefit from continuous vital-sign monitoring technology. Step-down wards are ideal candidates for this level of monitoring, as patients are admitted after major surgery or ICU stay and typically require greater care than general ward patients. Indeed, continuous monitoring on wards with higher rates of adverse events is a valid concept, as increasing the care and support for high-acuity patients forms the basis for ICUs and high-dependency units worldwide (Bellomo et al. [2012]). Some even believe that all hospitalised patients should receive continuous vital-sign monitoring – an idea known as surveillance monitoring. As distinct from condition monitoring, whereby clinicians prescribe monitoring based on their assessment of need, surveillance monitoring involves the continuous observation of all patients regardless of their health status (Pyke et al. [2012]).

Continuous monitoring during anaesthesia

A high proportion of published studies assessing the use of continuous surveillance monitoring have been related to anaesthesia. Opioid analgesics are commonly used as an effective treatment of post-operative pain, however a rare yet fatal side-effect can be a catastrophic reduction in ventilatory drive leading to respiratory depression (RD). RD is characterised by bradypnoea and hypoxaemia, with breathing rates below 10 breaths/minute and SpO_2 less than 80% (Catley et al. [1985]; Dahan et al. [2010]). While intermittent vital-sign observations may reveal these symptoms, the Anesthesia Patient Safety Foundation (APSF) has proposed that this level of care is inadequate for reliably recognising the clinical evolution of opioid induced RD in

the post-operative period (Taenzer et al. [2014]). The APSF, alongside other authors, have advocated the routine use of continuous post-operative respiratory monitoring in at-risk patients receiving patient controlled analgesia or neuraxial opioids in post-operative care (Maddox et al. [2008]; Overdyk et al. [2007]; Stoelting [2015]; Taenzer et al. [2012, 2014]; Weinger [2006]; Weinger and Lee [2011]).

The position held by the APSF in advocating continuous surveillance monitoring is supported by data published in the anaesthesia literature. Two observational studies of continuous respiratory monitoring have revealed that the prevalence rate of post-operative hypoxaemia and bradypnoea may be be severely underestimated in vital-sign records of manual observations (Overdyk et al. [2007]; Sun et al. [2015]). In a retrospective cohort study, Maddox et al. [2008] noted that at least 16 patients with declining physiological status were identified by continuous respiratory monitoring during a 33-month period, thus avoiding unwarranted outcomes and possible transfer to the ICU. A before-and-after concurrence study testing the impact of continuous heart rate and SpO_2 monitoring in post-operative care was undertaken by Taenzer et al. [2010], who reported fewer rescue events and intensive care unit transfers in the monitored group. The authors noted in a later publication that they have since implemented a hospital-wide policy of continuous heart rate and SpO_2 monitoring for all medical and surgical patients, and have seen institutional reductions in rescue events (0-65%) and in ICU transfers (0-50%) (Taenzer et al. [2012]). Ochroch et al. [2006] implemented continuous pulse oximetry monitoring in a post-operative ward, showing that whilst continuous monitoring was not associated with overall decreased transfer to ICU, there was a reduction in ICU admissions for pulmonary complications.

While continuous SpO_2 monitoring is a positive step for reducing the incidence of opioid-induced RD, it may be insufficient for monitoring patients receiving supplemental oxygen (Fu et al. [2004]; Maddox et al. [2008]; Stoelting [2015]; Weinger [2006]; Weinger and Lee [2011]). Under most circumstances, if ventilation were cut by 50% then SpO_2 will decrease significantly in only 2-3 minutes, however, SpO_2 may not decrease if the patient is breathing as little as 25% supplemental oxygen (Fu et al. [2004]; Weinger [2006]). Thus, supplemental oxygen can mask hypoventilation, making pulse oximetry a very late detector of RD (Weinger and Lee [2011]). It has been recommended that continuous respiratory rate monitoring is provided for patients receiving patient-controlled analgesia and supplemental oxygenation after surgery (Maddox et al. [2008]; Overdyk et al. [2007]; Stoelting [2015]; Weinger [2006]; Weinger and Lee [2011]). However, there are challenges in implementing this recommendation, most notably the lack of a reliable method of continuous respiratory rate measurement

(Lovett et al. [2005]), and a poor level of clinical understanding that pulse oximetry is not a reliable indicator of adequate ventilation (Cretikos et al. [2008]).

Other non-ICU use cases of continuous monitoring

Although post-operative RD has been the focus of many trials of surveillance monitoring, there have been studies of continuous vital-sign monitoring in other non-ICU settings. A trial of routine ECG monitoring in 8,932 cardiac telemetry patients showed almost no benefit to continuous ECG surveillance monitoring, as only 1 of the 3 arrests that were survived during telemetry monitoring was clearly signalled by the monitor before the arrest (Schull and Redelmeier [2000]). While continuous ECG monitoring has been suggested as a precautionary measure for some patients (even with a non-cardiac primary diagnosis), there is little evidence to justify the additional expense and discomfort of continuous surveillance over more selective monitoring programmes (Cusack and Coutts [2003]). Watkinson et al. [2006] conducted a randomised control trial of continuous vital-sign monitoring for preventing adverse clinical events in high-risk medical and surgical patients, with results indicating that mandated continuous monitoring in the study population had no effect on adverse events or mortality. However, the study had several limitations – no EWS system was used during the study, and the ward had a low staff:patient ratio, so clinical staff may not have had a chance to address physiological deterioration despite the use of continuous monitoring technology. Furthermore, few patients (16%) received monitoring for more than 48 hours out of the desired 72-hour monitoring period.

Positive outcomes have been reported in studies combining continuous surveillance monitoring with an EWS, many of which have used the Visensia (previously BioSign) patient monitoring system (OBS Medical, Abingdon, UK) developed by Tarassenko et al. [2005]. Visensia combines continuous vital-sign data in a probabilistic model of patient physiological status. In a trial of high-risk elective/emergency surgery or medical patients, surveillance monitoring with Visensia generated alerts in 105 patients every 8 hours (on average), 95% of which were deemed to be true by clinical staff (Tarassenko et al. [2006]). In an observational study of 326 step-down unit patients receiving continuous vital-sign monitoring, the retrospective application of Visensia to the data of the 7 patients who triggered MET activation criteria was shown to provide 6.3 hours earlier warning, on average, of physiological deterioration (Hravnak et al. [2008]). A follow-up study of 303 patients in the same ward trialled Visensia for automatically generating audible early warning alarms (Hravnak et al. [2011]). This intervention was associated with a decrease in cardiorespiratory instability concern

criteria, as well as a decrease in instability duration and the number of occurrences of serious instability episodes per admission. Subbe et al. [2017] described the use of cableless sensors for continuous vital-sign monitoring with automated EWS calculation in two general clinical wards, reporting a significant reduction in cardiac arrests, mortality, and severity of illness on admission to the ICU.

Secondary benefits of continuous monitoring

Beyond the primary motivation of improving patient outcomes through increased vital-sign measurement frequency, there may be additional benefits to surveillance monitoring. Manual vital-sign measurements are known to be error-prone and poorly documented (Hooker et al. [1989]; Lovett et al. [2005]), EWS values are often calculated incorrectly (Clifton et al. [2015]; Prytherch et al. [2006]; Subbe et al. [2007]), and clinical observation policies can be poorly adhered to (Cardona-Morrell et al. [2016a]; Clifton et al. [2015]; Hands et al. [2013]; Van Leuvan and Mitchell [2008]). Electronic observation (eObs) systems have been developed to improve manual vital-sign documentation and EWS calculation (Smith et al. [2006]; Wong et al. [2015]). Clinical trials of eObs systems have shown improvements in the time needed to take vital-sign observations (Bellomo et al. [2012]; Wong et al. [2017]) and the accuracy of EWS calculations (Pullinger et al. [2017]), as well as showing reductions in mortality (Schmidt et al. [2015]) and length of hospital stay (Bellomo et al. [2012]). Many of the advantages of eObs systems could also apply to continuous vital-sign monitoring systems, which can measure vital signs in real-time without any human input, and can also interface with software for automated EWS calculation and documentation.

Continuous surveillance monitoring may improve the consistency of monitoring throughout a patient's stay on the ward. Several studies have shown that RRS activations from manual vital-sign observations are more frequent during daytime and weekdays, compared to night-time and weekends (Churpek et al. [2017]; Galhotra et al. [2006]; Jones et al. [2013]; Mullins and Psirides [2016]). This phenomenon has partly been attributed to the low frequency of vital-sign observations taken at night during periods of reduced staffing levels (Hands et al. [2013]; Sundararajan et al. [2016]). There is a spike in the occurrence of cardiac arrest around 2 to 3 am (Jones et al. [2006]) and an increased mortality rate in patients admitted to hospital after hours and during the weekend (Bhonagiri et al. [2011]; Sundararajan et al. [2016]), which coincide with periods of reduced clinical staffing. Furthermore, the occurrence of RRS activations and cardiac arrests increase at around 6 to 8 am (Hands et al. [2013]; Jones et al. [2005, 2006]; Mullins and Psirides [2016]), which is the typical time of

nursing shift changeover. These RRS activations are associated with an increased mortality rate compared to activations at other times (Churpek et al. [2017]). It may be that patients deteriorate overnight unnoticed, only to be found with highly abnormal physiology as nursing shifts change (DeVita [2005]). There is also a higher likelihood for RRS alerts in the hours closest to ward admission (Hravnak et al. [2015]). It may therefore be beneficial to improve the frequency of vital-sign measurements during these times of increased risk. However, it could be difficult to increase the frequency of manual vital-sign observations as this could increase staffing costs and interrupt patients' sleep. Introducing continuous surveillance monitoring could be a viable method of improving the consistency of patient monitoring and smoothing the diurnal variation in RRS activations, cardiac arrest and mortality (Churpek et al. [2017]; DeVita [2005]; Galhotra et al. [2006]; Jones et al. [2005]; Yoder et al. [2013]).

There may also be an economic benefit to the use of continuous surveillance monitoring. Slight et al. [2014] constructed models to estimate the total cost savings attributable to the implementation of a continuous monitoring system in a medical-surgical unit. This system used a piezoelectric sensor under the patient's mattress to measure heart rate and respiratory rate. The base economic model, which considered the total cost savings of intervention effects, predicted a 120.3 - 347.9% annualised return of investment, while a more conservative model predicted a 25.4 - 125.5% annualised return of investment. In a clinical study of continuous pulse oximetry monitoring in postoperative care, Ochroch et al. [2006] reported lower estimated care costs from enrolment to censoring in patients receiving continuous monitoring compared with unmonitored patients ($15,481 versus $18,713; P = 0.038). Other authors have also predicted that the cost of instituting and maintaining continuous monitoring technology would be more than offset by cost savings associated with reduced ICU admissions and length of stay (Stoelting [2015]; Taenzer et al. [2012]).

Summary

The introduction of continuous surveillance monitoring to wards outside of the ICU may benefit patients through improved monitoring coverage, leading to better outcomes and lower costs. Indeed, clinical staff are optimistic that continuous monitoring can improve care. In a survey by Watkins et al. [2016], nurses agreed that continuous vital-sign surveillance will help enhance patient safety. A separate survey by Prgomet et al. [2016] showed that nurses and doctors believe that continuous vital-sign surveillance would allow earlier identification of patient deterioration and provide reassurance to patients. However, amongst the many attempts to address "failure to rescue",

and despite the clinical willingness to embrace new technology, patient surveillance remains an under-explored area of research (Taenzer et al. [2011]). There is currently insufficient evidence for continuous vital-sign monitoring to be recommended for routine clinical practice in general wards (Cardona-Morrell et al. [2016b]; van Loon et al. [2015]). Improvements in monitoring technology, clinical study design, and the reporting of statistical testing are required before surveillance monitoring can become more widely clinically accepted (Cardona-Morrell et al. [2016b]; van Loon et al. [2015]).

1.4.2 Challenges in continuous vital-sign monitoring

There may be potential for continuous surveillance monitoring to improve clinical outcomes, but there are important technical challenges that must be addressed before this potential can be realised. Continuous vital-sign monitoring technology must be accurate and reliable, while EWS algorithms must have sufficiently high sensitivity and specificity for predicting physiological deterioration. Furthermore, continuous monitoring sensors must be comfortable for the patient to wear, and there must be an appropriate user interface for clinical staff to access the vital-sign and EWS data. The performance of technology in these key areas will influence the ability of surveillance monitoring to improve patient outcomes. Indeed, it may be the case that previously reported clinical trials of continuous surveillance monitoring have been limited by poor technical execution. Considering the mixed results from previous clinical trials of continuous surveillance monitoring, it is reasonable to conclude that improvements in vital-sign sensor technology and signal processing algorithms are required in order to maximise the impact of continuous surveillance monitoring.

False alarms

Continuous vital-sign monitoring equipment in the ICU are programmed to trigger alarms when vital-sign values exceed pre-programmed thresholds, or when algorithms detect abnormal patterns such as arrhythmias or apnoeas. These alarm systems are designed to improve patient safety by providing full monitoring coverage regardless of whether or not clinical staff are present at the bedside. Adverse clinical events can be detected and treated without delay, to the benefit of the patient's recovery.

Alerting algorithms for continuous monitoring systems must be sensitive to detect as many instances of physiological abnormality as possible (Imhoff and Kuhls [2006]). A drawback of high algorithm sensitivity is that periods of normal physiology may

be incorrectly interpreted as abnormal, thus generating false alarms (Bonnici et al. [2013]). The rate of false and/or clinically insignificant alarms, also known as nuisance alarms, may range from 80 to 99% in the ICU (Cvach [2012]). The overabundance of monitoring alarms desensitizes clinicians to the apparent urgency of the alerts, often leading to staff disabling alarm systems. This is sometimes referred to as the "cry wolf" effect, which is a problem in critical care, as the majority of all threshold alarms in the ICU have no real clinical impact (Lawless [1994]; Sendelbach and Funk [2013]).

False alarms in hospital wards outside of the ICU would place an unnecessarily high burden on the workload of clinical staff as the staff-to-patient ratios are much lower outside of the ICU, so there are fewer staff to respond to alerts. However, there may be strategies to reduce the occurrence of false alarms in wards outside of the ICU. These wards may tolerate alarm delays that are much greater than those in critical care, as patients are less acute and thus are less likely to require an urgent clinical intervention. This allows the use of signal processing strategies such as signal averaging and alarm delays to reduce the occurrence of false alarms due to erroneous signal artefacts from the vital-sign monitoring equipment (Li and Clifford [2012]; Orphanidou et al. [2015]). Advanced machine learning techniques such as time series models, Bayesian methods and deep learning may also improve the performance of alerting algorithms, thus reducing the burden of false alarms. Improvements in the sensitivity and specificity of alerting algorithms are required before automated alarms can be incorporated into surveillance monitoring systems outside of the ICU.

Ambulation after surgery

Post-operative ambulation is an important part of the care pathway during recovery from major surgery (Dorcaratto et al. [2013]). Patients are encouraged and enabled to mobilise for periods of time during post-surgical recovery to avoid hazards of bed rest (Fearon et al. [2005]). Traditional bedside continuous monitoring systems require wired sensors to be attached to patients, thus tethering patients to the bedside and hindering the ease of ambulation. Bedside equipment is necessary in the ICU as patients require highly accurate monitoring of physiological waveforms, while the requirements for physiological monitoring are much less stringent outside of the ICU. Indeed, since the standard of care for most non-ICU wards is manual vital-sign observations, any level of continuous monitoring would improve monitoring coverage. Wireless or non-contact vital-sign monitoring equipment may be beneficial in post-operative care, allowing patients to ambulate freely while receiving the benefits of continuous monitoring.

1.4.3 Wearable vital-sign monitors

In recent years there has been considerable growth in the use of wearable sensors (wearables) for non-invasive physiological monitoring. New devices have been developed by manufacturers including traditional medical technology companies such as Philips and Nonin Medical, and consumer technology companies like Apple and Fitbit. The most popular application of wearable technology is fitness tracking. Wearables are commonly equipped with heart rate sensors, GPS sensors and accelerometers to measure physical exertion and duration of exercise. Clinical applications of wearable technology have had less attention, which is likely due to the difficulty of achieving clinical levels of monitoring accuracy and obtaining regulatory approval for clinical use. However, wearables may be useful for clinical care since they can provide continuous non-invasive physiological monitoring for extended periods of time.

Commercial wearable systems have been developed to provide clinical measurements of a number of physiological parameters. Traditionally, ambulatory ECG monitoring is conducted with a Holter monitor (Yilmaz et al. [2010]), consisting of several ECG electrodes recording waveform data written to a neck-worn device. Wearable patches are now available that combine an adhesive electrode with data storage and transmission chips to wirelessly monitor and record ECG signals (Hernandez-Silveira et al. [2015]; Lobodzinski [2013]). Wearable garments with integrated ECG sensors have also been tested in several clinical settings (Jeffs et al. [2016]). Ambulatory PPG monitoring can be conducted with pulse oximeter sensors attached to several anatomical sites, including the finger, the wrist, the upper arm and the ear (Maeda et al. [2011]; Tamura et al. [2014]). While heart rate and SpO_2 are the most commonly measured vital signs with wearable sensors (as these are easily derived from ECG and PPG waveforms), various systems also estimate temperature, respiratory rate and blood pressure with wireless arm bands, chest sensors, wrist bands or thumb sensors (Budinger [2003]; Klingeberg and Schilling [2012]; Weenk et al. [2017]). Accelerometers and gyroscopes are often integrated into wearable systems to detect movement for applications such as falls detection (Budinger [2003]) or to suppress vital-sign signals contaminated with movement artefact (Salehizadeh et al. [2016]).

There are many potential clinical applications for wearable systems, which can be deployed in a number of settings. Wearables have been trialled to measure vital signs from patients in step-down wards (Chipara et al. [2010]), patients discharged from the ICU (Jeffs et al. [2016]), patients undergoing surgery and patients in the general wards (Hernandez-Silveira et al. [2015]). Wearable systems have also been deployed remotely

in disaster scenarios, for example using electronic tags equipped with a pulse oximeter and ECG leads to assist with triaging patients based on the severity of their condition (Gao et al. [2006]). There is also increasing interest in the use of wearable sensors for monitoring patients in their own home to improve the management of chronic health conditions and reduce hospitalisation (Baig et al. [2013]; Clifford and Clifton [2012]). Many of the studies conducted to date have reported the use of a wireless communication protocol for remotely transmitting vital-sign signals to a central base station, including Wi-Fi (Orphanidou et al. [2009]), radio networks (Chipara et al. [2010]), and the Zigbee protocol (Patel et al. [2012]).

The clinical utility of wearable sensors is currently limited by a number of technical issues that have been repeatedly identified in the literature. Patients find current wearable systems uncomfortable and will typically remove them after a period of time, resulting in low levels of compliance in clinical studies and significant episodes of missing vital-sign measurements (Bergmann and McGregor [2011]; Bonnici et al. [2012]; Fensli et al. [2008]; Jeffs et al. [2016]). Wearable sensors often produce unreliable measurements during periods of patient movement and may be sensitive to sensor placement (Chipara et al. [2010]; Maeda et al. [2011]; Orphanidou et al. [2009]; Tamura et al. [2014]). Perhaps the biggest issue with wearables is the short battery life of most systems (Baig et al. [2013]; Budinger [2003]; Nangalia et al. [2010]; Orphanidou et al. [2009]; Pantelopoulos and Bourbakis [2010]). While compromises can be made to improve battery life, including reducing the frequency at which measurements are transmitted or using low power transmission (Budinger [2003]), wearables will only be practical clinically when they can last several days on one battery charge. Finally, well-validated multi-parameter wearables must be developed to meet the requirements of many clinical wards, most of which have a minimum set of parameters (e.g. heart rate, respiratory rate and SpO_2) that a wearable must be able to measure.

1.4.4 Non-contact vital-sign monitoring

Camera-based non-contact monitoring is an emerging field in physiological measurement, in which visible light cameras are used to capture a photoplethysmographic imaging (PPGI) waveform from a targeted region of interest (ROI) on a patient's skin (Kamshilin et al. [2015]). The PPGI waveform was first extracted by Wu et al. [2000] using infrared light, and by Verkruysse et al. [2008] using ambient visible light. The most common way of capturing the PPGI waveform uses the average ROI pixel intensity from the green channel of a red green blue (RGB) camera sensor to create a

univariate signal, which is cardiac-synchronous due to the changes in skin colour and volume as blood pulses through the peripheral vasculature (Verkruysse et al. [2008]).

Signal processing algorithms continue to be developed to improve the quality of the PPGI waveform data extracted from the video. For example, blind source separation has been used to derive the latent PPGI signal from each of the RGB camera channels (Poh et al. [2010]). McDuff et al. [2014b] used orange, green and cyan colour channels rather than RGB channels, finding that their method improved the ability to locate systolic and diastolic peaks in the PPGI signal. Other work has focussed on the development of methods to improve the robustness of non-contact monitoring algorithms to motion artefact (de Haan and Jeanne [2013]; Feng et al. [2014]). In addition, new methods of estimating vital-sign values from PPGI data have been developed, such as the use of autoregressive models to derive robust estimates of heart rate and respiratory rate (Takano and Ohta [2007]; Tarassenko et al. [2014]).

Most of the previous literature in the field of non-contact vital-sign monitoring has reported experiments in controlled laboratory environments with healthy subjects (McDuff et al. [2014a]; Monkaresi et al. [2014]; Poh et al. [2011]). While there are clear advantages to conducting experiments in a controlled environment, this practice cannot emulate the challenging conditions of a clinical ward such as varying light levels or high patient movement. Previous work by members of our research group has focused on conducting non-contact monitoring studies in clinical environments. A clinical trial of non-contact vital-sign monitoring has been conducted in the high-dependency care area of the Neonatal Intensive Care Unit (NICU) at the John Radcliffe Hospital in Oxford, UK (Villarroel et al. [2014]). In this study, pre-term infants were monitored with a video camera and standard patient monitoring equipment during the daytime for at least 4 consecutive days of being nursed in an incubator. The camera was secured by a bespoke arm mechanism and positioned above a hole cut at the top of the incubator canopy. Other contributions by our research group in this field include a trial of non-contact vital-sign monitoring in patients undergoing haemodialysis in the Oxford Kidney Unit at the Churchill hospital in Oxford, UK (Tarassenko et al. [2014]). An optical video camera was attached to an arm mechanism by the patient's bedside, and was positioned approximately 1 metre away from the patient, with the patient's face in the focal plane of the camera image. Patients were monitored for the duration of their dialysis session, up to 4 hours. We present a clinical study of non-contact monitoring of acute post-operative patients in chapter 6.

Non-contact monitoring allows several different physiological parameters to be derived simultaneously. Studies have demonstrated that heart rate and respiratory

rate can be derived from video data, and in recent years published work has shown that it may be possible to derive SpO_2 from video data (Guazzi et al. [2015]; Shao et al. [2016]; Verkruysse et al. [2017]). Non-contact SpO_2 algorithms operate similarly to pulse oximetry (see section 1.1.3) by computing ratios between measurements from RGB channels of different wavelengths (Guazzi et al. [2015]) or by computing ratios between camera measurements with light sources of different wavelengths (Shao et al. [2016]; Verkruysse et al. [2017]). Further work is needed before non-contact estimation of SpO_2 becomes practical as the derived values still require external calibration (Guazzi et al. [2015]; Verkruysse et al. [2017]).

A non-contact monitoring system would have many benefits over traditional continuous monitoring systems for post-operative patients. Patients are not tethered to the bedside and can therefore ambulate freely, thus this type of system is potentially more appropriate in post-operative care than wired bedside monitoring systems. Furthermore, the issues that plague ambulatory monitoring solutions, such as patient discomfort, lack of mobility, poor battery life and unreliable networking, are easily resolved with a system that is remote, with no requirements for being wireless. Finally, contact sensors may damage fragile skin and must be cleaned to avoid passing on infection, which is not an issue with non-contact monitoring.

The integration of an EWS with non-contact vital-sign monitoring could be an innovative method of improving the afferent limb on an RRS. However, several technical improvements are required before this could be possible. For example, non-contact estimation algorithms must be robust to patient motion and variable lighting conditions, as well as different patient physiologies and skin tones. Non-contact algorithms must also be able to accurately detect changes in vital-sign values. Most previous studies have provided accuracy analyses conducted with vital-sign values that hardly change during the course of the experiment. EWSs are designed to alert clinical staff when important changes in vital-sign values occur, so a non-contact monitoring based system must reliably track changes in physiology.

1.5 Overview of thesis

In this thesis we analyse the use of machine learning and continuous vital-sign monitoring for improving the detection of physiological deterioration in post-operative care. We begin in chapter 2 by describing datasets of vital signs collected manually and from continuous monitoring equipment during two clinical studies in an acute post-operative population. We use this dataset for the analyses in chapters 3 to 5.

In chapter 3 we use machine learning to assess the value of features of diurnal vital-sign variability for improving the performance of an EWS for detecting physiological deterioration. Chapter 4 provides a method comparison study of continuous and manual vital-sign data. We derive features of diurnal variability from continuous vital-sign data in chapter 5 for heart rate, respiratory rate and SpO_2, and assess their ability to discriminate between physiological deterioration and physiological stability.

In chapter 6 we present a clinical study that was conducted to assess extended non-contact monitoring with a video camera in post-operative patients during elective stay in the ICU. We extract periods of video data from the study dataset, which we use for analyses in chapters 6 and 7. We present methodologies to derive heart rate and respiratory rate estimates from video camera data, and compare the non-contact estimates to contemporaneous values from reference vital-sign monitoring equipment. In chapter 7 we describe a method for visualising respiratory amplitude in different regions over a patient's chest, and present a case study of this methodology with a patient that experienced significant respiratory distress. Finally, concluding remarks and thesis contributions are summarised in chapter 8.

1.6 Summary

In this chapter we provided an overview of patient monitoring in post-operative care. We described the main vital signs in section 1.1, and both manual (section 1.1.2) and continuous (section 1.1.3) methods of measuring them. We introduced the problem of adverse clinical events (section 1.3) and the use of EWSs to detect physiological deterioration (section 1.3.1) including methods based on machine learning (section 1.3.2). We described the potential of continuous vital-sign monitoring technology for detecting physiological deterioration in post-operative care (section 1.4) and how this relates to recent developments in non-contact vital-sign monitoring (section 1.4.4). Finally, we provided an overview of the thesis in section 1.5.

Chapter 2

The Computer Alerting Monitoring System Studies

2.1 Study population

The investigations conducted in chapters 3 to 5 of this thesis involved the analysis of vital-sign data acquired from patients recovering from major upper gastrointestinal cancer surgery in the Oxford University Hospitals NHS Trust (Oxford, UK). Upper-gastrointestinal cancer refers to malignancies of the upper-gastrointestinal tract, including cancers of the oesophagus, stomach, and pancreas, as well as gastrointestinal stromal tumours (Shah and Kurtz [2010]). Of the 14 million worldwide cancer diagnoses annually, over a quarter are gastrointestinal cancers, with more than 4 million new cases each year (Meyerhardt et al. [2015]). Gastrointestinal diseases have a heavy burden in the UK as the third most common cause of death, the leading cause of cancer death, and the most common cause of hospital admission (Williams et al. [2007]). Treatment options for upper gastrointestinal cancer include surgery, radiotherapy and chemotherapy, which are considered individually or in combination depending on the stage of disease and patient fitness.

Surgery for the removal of gastrointestinal malignancies is a major undertaking, requiring considerable recovery time (typically about a week) for the body to restore homeostasis (Bond-Smith et al. [2016]; Grade et al. [2011]). The outcome of patients who undergo gastrointestinal surgery varies greatly, depending on such factors as age and co-morbidities, the complexity of the surgical procedure and the management of postoperative recovery (Grade et al. [2011]; Jakobson et al. [2014]). Postoperative complications can be common, particularly for those patients considered to be at high risk (Grade et al. [2011]; Jakobson et al. [2014]; Martos-Benítez et al. [2016]). Effective management of post-operative recovery is critical for improving the outcomes of major surgery, as complications can have immediate effects during the postoperative period and long-term effects on quality of life and mortality (Martos-Benítez et al. [2016]).

Post-operative recovery following gastrointestinal cancer surgery in the Oxford University Hospitals NHS Trust is conducted in a specialist step-down post-operative ward. High-risk patients are admitted after a period of elective ICU stay, while patients with a lower risk of complication are admitted to the ward immediately after surgery. High Dependency Unit (HDU) bed spaces are also available in the step-down ward for the first 2-48 hours of post-operative recovery for the provision of increased monitoring, equivalent to a level 2 ward. Patients in HDU care undergo conventional bedside monitoring, while clinical staff also make manual measurements, typically at hourly intervals. After the initial period of increased monitoring, patients receive general-ward level care (level 1) for the duration of their recovery, consisting of manual vital-sign recording typically at four-hourly intervals. A track-and-trigger system is used such that care is escalated when vital-sign recordings show patterns of physiological abnormality (Tarassenko et al. [2011]).

2.2 The CALMS-2 study

As described in section 1.4, the prevention of adverse events in post-operative care may be limited by the practice of monitoring patients manually, as unobserved deteriorations between successive observations cannot be addressed. It has been hypothesised that continuous vital-sign monitoring might reduce adverse events by improving the frequency of observation. The CALMS-2 trial tested this hypothesis by providing patients with an Integrated Monitoring System (IMS) consisting of ambulatory wearable monitoring equipment combined with an automated alerting system during the recovery period after upper gastrointestinal surgery. The CALMS-2 trial was conducted in a step-down post-operative ward of the Oxford University Hospitals NHS Trust, Oxford, between May 2009 and December 2013. Ethical approval was granted (Mid and South Buckinghamshire ethics committee REC: 08/H0604/79, 09/12/2008 and Leeds [West] ethics committee REC:11/YH/0056, 20/05/2011) and the trial was registered in the ISRCTN database (principal investigator: Dr Peter Watkinson, ISRCTN No: ISRCTN58660550, 11/08/17). Details of the study design, methods and results are summarised here having first been described by Pimentel et al. [2013b] and Watkinson et al. [In preparation 2018].

2.2.1 Study design

The CALMS-2 trial was a prospective, before-and-after interventional study primarily assessing whether the use of an ambulatory IMS could reduce hospital length of stay. Secondary outcome measures included the number of major adverse events, number of unplanned returns to the operating theatre; time to intensive care readmission from the ward; number of cardiac arrest calls; in-hospital mortality and 30-day mortality post-ward admission; and assessment of the performance of the IMS in comparison to the intermittent paper-based EWS used on the ward. An adverse event was defined as a cardiac arrest, death or unplanned admission to the ICU. The study was conducted in two phases, allowing a "baseline" period of data to be acquired in phase I against which the outcomes from phase II could be compared.

Participants

Potential participants were screened during pre-operative assessment and were deemed eligible if they were scheduled to undergo major upper gastrointestinal surgery. This was defined as: oesophagectomy, oesophagogastrectomy, gastrectomy, Whipples operation, liver resection, pancreatectomy, gastric bypass, biliary reconstruction and splenectomy. Participants were excluded based on the following criteria: participants under 16 years of age, pregnant women, participants unable to wear the required monitoring, participants without capacity to consent, and participants who could not understand written English for whom no translator could be found. There were 200 patients recruited in phase I and 207 patients in phase II.

Equipment and set-up

Upon arrival to the ward after surgery, patients were connected to a conventional bedside monitor (Philips M3046A/Intellivue MP50 clinical monitor). The monitors were connected to an external tablet computer (Toughbook Getac CA27), allowing the transmission of vital-sign data to a secure server via the hospital wireless network. The bedside monitor measured respiratory rate via impedance pneumography, heart rate was derived from the ECG and SpO_2 was measured by pulse oximetry. Non-invasive blood pressure measurements were recorded periodically by the bedside monitor and tympanic temperature recordings were entered manually.

When patients were deemed safe to move from HDU care to normal ward level care, they were transitioned from the bedside monitoring equipment to the ambulatory system, which they wore until discharge from the ward or until they requested to

stop wearing the equipment. The ambulatory system consisted of a wearable pulse oximeter (Avant 4100 Wrist-Worn Participant Module Nonin) and a wearable ECG monitor (Corscience BT 3/6 wireless PC ECG), which were connected via Bluetooth to Personal Digital Assistants (PDAs) available to clinical staff (HTC Touch 3G devices). The use of ambulatory ECG devices was discontinued after 25 patients as the devices were poorly tolerated. Heart rate and SpO_2 measurements were provided continuously by the ambulatory equipment, while clinical staff input respiratory rate, temperature and blood pressure into the PDAs during routine observations. PPG waveforms from the pulse oximeter, ECG waveforms (for the first 25 patients), continuous vital-sign measurements and manual observations were transferred via the hospital Wi-Fi to a central server where the IMS was installed.

The alerting system used in this study was the Visensia IMS, a US Food and Drug Administration (FDA) approved patient monitoring system. Visensia integrates up to five vital signs (heart rate, respiratory rate, SpO_2, blood pressure and temperature) to produce a single physiological score of abnormality, the VSI, using a probabilistic model of normal physiology (Tarassenko et al. [2005]). The VSI was normalised between 0.0 (normal physiology) to 5.0 (abnormal physiology), with a threshold of 3.0 chosen to represent a clinically-significant departure from normality. Alerts were generated when the VSI exceeded the threshold for more than 80% of vital-sign values within the preceding 5 minutes. Missing data were imputed with the population mean from the model of normality.

Study phases

Phase I of the CALMS-2 study occurred between May 2009 and June 2011. Consented patients received standard ward care, including the use of a paper-based EWS. During their stay, patients were actively monitored with the IMS, which recorded continuous vital-sign data and calculated VSI values. While the ambulatory vital signs could be seen by clinical staff, the VSI values were not displayed and thus were not used for clinical decisions. Reviews were conducted regularly, allowing the development of standard operating procedures for phase-II to maximise the effectiveness of the monitoring system.

In phase II, which occurred between June 2011 and December 2012, the VSI was displayed on the tablet by the participants' bedside, or on the PDA if they had been transferred to ambulatory monitoring. The VSI value was also displayed with the patient's details on large screens opposite the nursing workstations. The VSI value was used explicitly in clinical decision making, although the escalation protocol for

the VSI was at the discretion of ward staff. Intensive care staff from the adjacent ICU were available to review patients as requested by ward staff. As well as the IMS, patients received normal ward care and the paper-based track-and-trigger system remained in place.

For the analyses in this thesis, we used the CALMS-2 dataset without specifying the study phase during which the data were collected, as this was not relevant to our analysis goals. Instead, for the analysis in chapter 4 we used the entire CALMS-2 dataset and in chapters 3 and 5 we separated patients according to whether or not they suffered a major adverse clinical event during their post-operative recovery. In section 2.2.2 below we compute data summaries separately for those patients who suffered an adverse event ("abnormal" group) and those that did not ("normal" group).

2.2.2 The CALMS-2 dataset

CALMS-2 followed a formal data collection and reconciliation process to obtain the patient data required to complete the primary and secondary analyses. These data were sourced from electronic records and paper notes, and subsequently amalgamated as the CALMS-2 dataset. Patient-sensitive information such as name, hospital ID and birth date were removed during this process, and each patient record was assigned a unique study ID (e.g. "001" for the first patient). Study IDs were generated sequentially, and the mapping of hospital ID to study ID was only available to research nurses and senior members of the clinical team.

Patient and operation data descriptors as well as clinical events were prospectively transcribed from the clinical notes in a standardised manner. Data were double-entered into the study database by two research staff members who were blinded to the other entry. Differences were reconciled by a third researcher with access to the original charts, to ensure that all events had been captured. Hospital IDs were used to match CALMS-2 patients to their clinical information stored in the hospital's electronic records. Relevant information to the CALMS-2 study, as approved by the ethics committee, was copied to the study database. The types of clinical data collected included demographics and types of operations, whether the patient received epidural anaesthesia and the patient's ASA[1] severity score. Furthermore, important dates were collected such as the time of operation, ward admission and discharge, as well as

[1]The American Society of Anaesthesiologists physical status classification (ASA) is a scoring system for assessing the fitness of a patient before surgery (Daabiss [2011]). The score encodes the subjective clinical assessment of a patient's health as an integer value between 0 (completely healthy) and 5 (moribund, unlikely to survive).

elective and emergency ICU admission and discharge.

Table 2.1: Demographic and other clinical information for patients in the CALMS-2 study who had an adverse event ("abnormal" group) and patients who did not have an adverse event ("normal" group). LOS is the length of time the patient stayed on a particular ward and IQR is the interquartile range.

	N (%) Patients		
	Normal (N = 357)	**Abnormal (N = 50)**	**P-value**
Gender (Male)	207 (56.0)	26 (52.0)	0.52
Elective ICU stay	124 (34.7)	34 (68.0)	<0.01
Epidural use	281 (78.7)	43 (86.0)	0.31
ASA Grade			
ASA I	46 (12.9)	3 (6.0)	0.24
ASA II	199 (55.7)	27 (54.0)	0.94
ASA III	72 (20.2)	14 (28.0)	0.28
ASA IV	3 (0.8)	1 (2.0)	0.99
Surgery Type			
Gastric-bypass	20 (5.6)	0 (0.0)	0.17
Oesophagectomy	55 (15.4)	24 (48.0)	<0.01
Hepatectomy	75 (21.0)	6 (12.0)	0.19
Gastrectomy	39 (10.9)	7 (14.0)	0.69
Pancreatectomy	145 (40.6)	10 (20.0)	<0.01
Splenectomy	18 (5.0)	3 (6.0)	0.96
Others	5 (1.4)	0 (0.0)	0.88
Adverse events			
Emergency ICU	-	46 (92.0)	-
Hospital mortality	-	7 (14.0)	-
	Median (IQR)		
Age, years	63 (53 - 69)	63 (52 - 72)	0.24
Ward LOS, days	8.4 (5.8 - 11.9)	19.2 (13.0 - 37.3)	<0.01
Hospital LOS, days	9.3 (6.9 - 13.4)	23.3 (16.2 - 45.0)	<0.01
Emergency ICU LOS, days	-	3.4 (2.0 - 6.3)	-
Time to first event, days	-	4.9 (3.1 - 9.7)	-

Demographics for the CALMS-2 study population are given in table 2.1. Statistical analyses were conducted to test for differences in the demographic values between the "normal" and the "abnormal" patients. Chi-squared tests for equal proportions were used for categorical variables, and Mann-Whitney U tests were used for continuous variables. There was no significant difference in age or gender between "normal" and "abnormal" patients. The use of epidural anaesthesia was equally prevalent in both groups, while patients who had an elective post-operative stay in the ICU

were significantly more likely to suffer an adverse event ($P<0.01$). There was no difference in the ASA score between groups, although there were differences in the types of surgery. For example, patients who had an oesophagectomy were more likely to be from the "abnormal" group ($P<0.01$), while those who had a pancreatectomy were more likely to be from the "normal" group ($P<0.01$). No patients who had gastric-bypass surgery suffered an adverse event. Ward length of stay and hospital length of stay were significantly longer for the "abnormal" group than the "normal" group ($P<0.01$).

Physiological data from paper records

Physiological measurements recorded during routine bedside observations were documented on paper-based charts by clinical staff. Vital signs, including heart rate, respiratory rate, SpO_2, blood pressure, temperature and neurological status, were measured using methods described in section 1.1.2. Neurological status was measured exclusively using the AVPU scale. If patients were receiving supplemental oxygen through an oxygen mask or a cannula, this was also documented. Physiological data from bedside charts were double-entered into the electronic database by two research staff members who were blinded to the other entry, with differences reconciled by a third researcher with access to the original charts.

Table 2.2 summarises the vital-sign data collected by manual observation. The "abnormal" group had approximately twice as many manual vital-sign observations taken as the "normal" group, which is a reflection of the longer average length of stay for "abnormal" patients and also the clinical practice of increasing the frequency of observation during times of abnormal physiology. The percentage of missing data was similar for patients in both groups, for all vital signs. Most vital signs were consistently recorded during clinical observations, except for temperature, which was only measured during 89% of clinical observations in the "normal" group and 87% of clinical observations in the "abnormal" group.

Clifton et al. [2015] analysed the vital-sign data taken during clinical observations in phase I of the CALMS-2 study. While similar rates of vital-sign data missingness (85.2%) were reported as in table 2.2, the authors also had access to the recorded aggregate EWS values, reporting that only 77.9% of the clinical observations had complete vital-sign recordings and completed EWS calculations. Furthermore, the authors found that incomplete observation sets (the set of vital signs recorded during clinical observation) were more likely to contain values that should have led to an EWS alert than complete observation sets.

Table 2.2: Overview of the number of manual vital-sign observations per patient, and completeness of the observation sets collected during the CALMS-2 study. We present data for patients who suffered an adverse clinical event ("abnormal" patients) and for patients who did not suffer such an event ("normal" patients). Note that the main vital signs here are heart rate, respiratory rate, oxygen saturation, systolic blood pressure and temperature.

	Normal	**Abnormal**
Measurements per patient	**Median (IQR)**	
Total measurements (N)	59 (43 - 81)	119 (74 - 222)
Completed (%)		
All main vital signs	87 (81 - 91)	86 (80 - 91)
Heart rate	99 (97 - 100)	99 (98 - 100)
Respiratory rate	98 (97 - 100)	99 (98 - 100)
Oxygen saturation	100 (99 - 100)	100 (99 - 100)
Systolic blood pressure	100 (99 - 100)	100 (99 - 100)
Temperature	89 (84 - 94)	87 (84 - 92)
Neurological status	96 (0 - 100)	95 (19 - 99)
Oxygen support	100 (98 - 100)	100 (98 - 100)

Physiological data from electronic records

Continuous vital-sign measurements from the bedside and ambulatory vital-sign monitors were identified on the central server and assigned to the correct patient records in the CALMS-2 dataset. The data sources available from the bedside monitor were limited to manufacturer-derived vital-sign measurements, as the raw waveform signals were not transmitted. Continuous measurements of heart rate, respiratory rate and SpO_2 were retrieved, with a sampling frequency of 1 Hz. Non-invasive blood pressure measurements were also transmitted by the bedside monitor, which were sampled every 30 minutes. Data from the ambulatory monitors included manufacturer-derived measurements of heart rate and SpO_2, sampled at 1 Hz, and raw PPG waveform data, sampled at 75 Hz.

The total amount of continuous monitoring received by each patient varied throughout the study, mostly due to differences in the length of stay on the ward. There were also periods when patients could have received continuous monitoring but did not, which were considered to be missing data. As reported in Watkinson et al. [In preparation 2018] the most common reason for time without monitoring was due to patients choosing to remove the equipment as they neared discharge from the ward. Documented gaps in recording were attributable to equipment removal during patient care temporarily (40.6%) or until the next research nurse check (23.3%), with 12.7%

of gaps caused by equipment failure such as sensor malfunction or battery depletion of the ambulatory system.

For each patient, we computed the duration of continuous vital-sign monitoring from the bedside monitor and the wearable monitor, which we present in table 2.3. The duration of monitoring time was calculated as the time between the final timestamp and the first timestamp of the data record for the monitoring source, minus any time spent in the ICU between these timestamps. The data time was calculated as the duration of monitoring minus the sum of time gaps longer than 30 minutes. We then computed the data time as a percentage of the monitoring time, which is a measure of continuous monitoring compliance. The "abnormal" group received more continuous monitoring from both sources than the "normal" group. However, the compliance (data time percentage) was lower in the "abnormal" group. This can partly be explained by the longer length of patient stay on the ward for the "abnormal" group. It may be harder to achieve long term monitoring as patients become less tolerant of wearing sensors.

Table 2.3: Number of monitoring days and completeness of continuous vital-sign data for the "normal" and "abnormal" patient groups in the CALMS-2 study. The monitoring time was calculated as the time between the first and last continuous readings for each patient, minus any time spent in the ICU. Data time was calculated from the monitoring time with gaps longer than 30 minutes removed.

	Normal	**Abormal**
	Median (IQR)	
Bedside monitor		
Monitoring time, days	1.27 (0.73 - 2.45)	2.14 (0.94 - 4.93)
Data time, days	1.02 (0.69 - 1.68)	1.70 (0.87 - 2.44)
Data time, %	95.9 (76.9 - 100.0)	75.1 (44.6 - 97.3)
Ambulatory monitor		
Monitoring time, days	3.74 (1.92 - 5.86)	4.43 (1.91 - 10.37)
Data time, days	1.81 (0.84 - 3.30)	2.45 (1.47 - 4.31)
Data time, %	65.6 (42.4 - 80.6)	59.1 (34.2 - 83.9)

Figure 2.1 shows the data time of continuous monitoring from bedside and wearable sources for heart rate, respiratory rate and SpO_2. We ordered the patients by the total duration of monitoring time (i.e. the elapsed time between the beginning and end of all continuous monitoring), and presented the data separately for the "normal" and "abnormal" patient groups. Respiratory rate was not measured by the wearable monitoring equipment, hence there were less continuous respiratory rate data compared to heart rate and SpO_2. The duration of continuous monitoring ranged from over 40

days (for 1 "normal" patient) to less than a day. Most "normal" patients received a similar duration of bedside monitoring but differed in the duration of wearable monitoring received, which was likely to be due to the different lengths of time that the patients stayed on the ward. As discussed in the previous paragraph, monitoring compliance was much lower for "abnormal" patients than for "normal" patients. Considering that patients are most in need of monitoring during times of physiological deterioration, the poor compliance prior to an adverse event is a concerning finding.

Use of the CALMS-2 dataset

Several analyses have been conducted using the CALMS-2 dataset prior to the investigations reported in this thesis. Watkinson et al. [In preparation 2018] reports statistical analyses of the primary and secondary outcomes of the CALMS-2 study. The main finding of this was that the CALMS-2 study failed to demonstrate that an IMS with ambulatory monitoring can reduce the length of stay on the ward after major surgery. There was no significant difference in the length of ward stay between the two phases (the primary outcome). Fewer deaths were reported between phases, although the study was underpowered for measuring mortality with statistical significance. The CALMS-2 study was perhaps limited in its implementation, as patients were not monitored for the full duration of their time on the ward. Furthermore, many patients in CALMS-2 were unlikely to suffer serious complications regardless of their monitoring level, thus potentially diluting the intervention effects in the statistical analyses. We note that the lack of a difference between phases with respect to study outcomes justifies the practice of combining the two phases as one dataset, as we have done in this thesis.

Data collected during phase I of the CALMS-2 study has been used in previous analyses, such as the work of Clifton et al. [2015] in assessing errors and omissions in clinical observation data (discussed in section 2.2.2). In addition, these data have been used to explore the development of EWS models using Bayesian machine learning techniques (Khalid [2014]; Khalid et al. [2014]) and one-class support vector machines (Clifton et al. [2014]). A kernel density estimation based EWS model was developed by Pimentel et al. [2013b], which was then extended to include both phases of the study (Pimentel [2015]), achieving state-of-the-art performance *in silico* compared to traditional EWS systems.

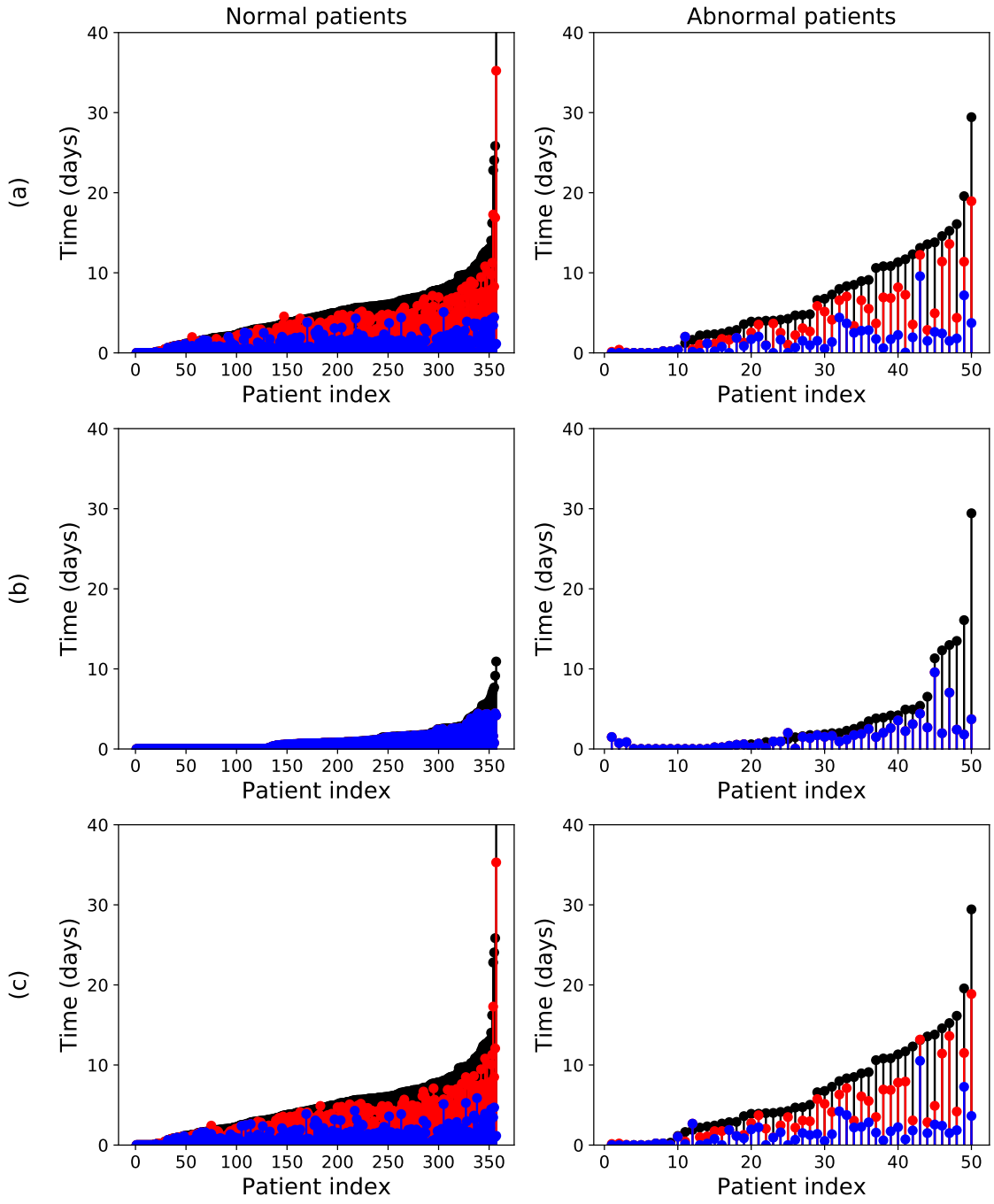


Figure 2.1: Time of continuous heart rate (a), respiratory rate (b), and oxygen saturation (c) monitoring received during the CALMS-2 study for the "normal" patient group and the "abnormal" patient group. Bedside monitoring time is shown in blue and wearable monitoring time is shown in red, ignoring time gaps over 30 minutes. The time between the start and end of continuous monitoring (both sources) is shown in black, ignoring any time spent in the ICU. The wearable monitor did not measure respiratory rate. Patients are ranked by total monitoring time.

2.3 The CALMS-3 study

2.3.1 Motivation

The machine learning based EWS developed by Pimentel [2015] for post-operative patients was limited by the vital-sign data available in the CALMS-2 dataset. Of the 407 patients monitored, only 50 had an adverse clinical event. While this event-rate was consistent with the prevalence of adverse clinical events after major cancer surgery (10-15%, Sukumar et al. [2013]), the resulting dataset was highly unbalanced. The consequence of this class-imbalance can be understood by considering the prediction of adverse events as a binary classification task, in which all vital signs recorded within 24 hours of an event are the "abnormal" class, and all other vital-sign observations (recorded at any time) are the "normal" class. In this framework, only 2-3% of the vital-sign observations in CALMS-2 were "abnormal".

The class-imbalance in CALMS-2 motivated the use of novelty detection methods (described in section 1.3.2). Such methodologies were used by Pimentel [2015] to outperform the current EWS standard of care (NEWS). However, it was recognised that further improvements could be made if a sufficiently balanced dataset was available such that two-class machine learning algorithms could be trained effectively. Indeed, Pimentel [2015] observed that patterns in vital-sign time-series may change during physiological deterioration, warranting further exploration in an improved dataset.

The CALMS-3 study was therefore set-up to further explore the potential to include features of vital-sign time-series patterns in post-operative risk prediction models. The primary objective of CALMS-3 was to collect a case-matched research database of vital-sign data from upper gastrointestinal surgical patients. Secondary objectives included the derivation of a post-operative risk prediction model combining both time-variant and time-invariant clinical parameters, and the determination of whether different patterns in physiological data can be identified in patients that deteriorate after surgery. A further secondary objective was to determine whether patterns in routinely-recorded physiological measurements can improve the performance of alerting systems (EWS models) following upper gastrointestinal surgery. Details of the CALMS-3 study are described here, including the data collection protocol and a summary of the acquired dataset.

2.3.2 Study design

Clinical hospital datasets were searched to identify patients who underwent upper gastrointestinal surgery and subsequently experienced post-operative complications that resulted in emergency ICU admission, emergency theatre admission, or death on the ward. These patients were then case-matched to find a cohort of patients who did not require emergency ICU or theatre during their hospital stay and survived to discharge. Patients were matched by surgery type. Eligibility criteria included all patients above 16 years admitted to the Oxford Radcliffe Hospitals NHS Trust for upper gastrointestinal surgery between January 2005 and May 2011, and those operated on during or after January 2013. These criteria included patients undergoing the following procedures: oesophagectomy, oesophagogastrectomy, gastrectomy, Whipples, liver resection, pancreatectomy, gastric bypass and biliary reconstruction.

Physiological data were extracted from clinical databases and compiled to form the CALMS-3 study dataset. The databases included the hospital surgical clinical (JANUS) system, the hospital Patient Administration System (PAS) or other legacy Electronic patient Record (EPR) successors, the track and trigger bedside observation charts and clinical data from patient notes, and the hospital laboratory database (LIMS) system. Manual vital-sign observation data and post-ICU laboratory data were extracted from these sources, as well as times of admissions, discharges and other relevant clinical events. Other variables likely to be associated with adverse clinical events were identified by expert panels and literature reviews, and were extracted from the databases. Data were then checked and validated before the dataset was anonymised by removing identifiable patient data and assigning each patient record with a unique study ID (e.g. “001” for the first patient).

The CALMS-3 data collection protocol specified that data from 200 patients who suffered an adverse clinical event (“abnormal” patients) would be extracted, and that this data would be matched 1:1 with an additional 200 patients who did not suffer an adverse event (“normal” patients). However, the CALMS-3 research team could not find appropriate data for more than 149 “abnormal” patients. Thus, the research team decided to use a 1:2 matching to ensure that a sufficiently large dataset would be collected. The final study dataset was compiled with data from 149 “abnormal” patients and 298 “normal” patients.

Table 2.4: Demographic and other clinical information for patients in the CALMS-3 study who had an adverse event ("abnormal" patients) and patients who did not have an adverse event ("normal" patients).

	N (%) Patients		P-value
	Normal (N = 298)	**Abnormal (N = 149)**	
Gender (Male)	211 (70.8)	105 (70.5)	0.97
Elective ICU	59 (19.8)	47 (31.5)	<0.01
Epidural Use	280 (94.0)	139 (93.3)	0.92
ASA Grade			
ASA I	16 (5.4)	3 (3.4)	0.48
ASA II	133 (44.6)	66 (44.3)	0.97
ASA III	60 (20.1)	35 (23.5)	0.49
ASA IV	2 (0.7)	0 (0.0)	0.80
Surgery Type			
Whipples/bypass	110 (36.9)	55 (36.9)	1.00
Oesophagectomy	129 (43.3)	64 (43.0)	0.97
Hepatectomy	16 (5.37)	8 (5.37)	1.00
Gastrectomy	43 (14.4)	22 (14.8)	0.96
Adverse events			
Emergency ICU	-	131 (87.9)	-
Emergency theatre	-	81 (54.4)	-
Hospital mortality	-	39 (26.2)	-
	Median (IQR)		
Age, years	67 (61 - 73)	66 (62 - 73)	0.46
Ward LOS	9.91 (7.86 - 13.66)	4.25 (1.77 - 8.90)	<0.01
Hospital LOS	12 (10 - 16)	32 (17 - 54)	<0.01
Emergency ICU LOS	-	3.8 (2.0 - 11.9)	-
Time to first event	-	5.7 (3.0 - 10.8)	-

2.3.3 The CALMS-3 dataset

As in section 2.2.2, we present data separately for the "normal" and "abnormal" patient groups. Demographics for the CALMS-3 study population are given in table 2.4. When appropriate, statistical analyses were conducted to test the null hypothesis that there were no differences in demographics between phases. Chi-squared tests for equal proportions were used for categorical variables, and Mann-Whitney U tests were used for continuous variables. The process used to case-match "normal" patients with "abnormal patients" was successful, as there were no significant differences between patient genders, age, ASA grade and surgery type between these two groups. 87.9% of adverse events were emergency ICU admissions, 54.4% were emergency theatre

admissions and 26.2% were deaths (there was cross-over with multiple patients suffering more than one event), with the first event happening after a median of 5.7 (CI, 3.0 - 10.8) days on the ward. "Normal" patients spent significantly more time on the ward than "abnormal" patients, but significantly less time in hospital (both $P<0.01$).

The physiological parameters collected during the CALMS-3 study were similar to those collected during the CALMS-2 study, with the notable exclusion of continuous vital-sign data from patient monitors. Manual vital-sign observation data, including heart rate, respiratory rate, SpO_2, blood pressure and temperature were compiled from clinical databases. The methods used to collect these data were likely to be the standard manual methods of vital-sign observation as described in section 1.1.2. However, unlike in the CALMS-2 study this could not be confirmed as the data were compiled retrospectively. Neurological status was documented in AVPU status format and it was noted in each vital-sign observation record whether the patients received supplemental oxygen.

Table 2.5: Overview of the number of manual vital-sign observations per patient, and completeness of the observation sets collected during the CALMS-3 study. We present data for patients who suffered an adverse clinical event ("abnormal" patients) and for patients who did not suffer such an event ("normal" patients). Note that the main vital signs here are heart rate, respiratory rate, oxygen saturation, systolic blood pressure and temperature.

	Normal	**Abnormal**
Measurements per patient	**Median (IQR)**	
Total measurements (N)	68 (52 - 89)	54 (31 - 81)
Completed (%)		
All main vital signs	85 (79 - 91)	77 (64 - 86)
Heart rate	98 (96 - 99)	98 (96 - 100)
Respiratory rate	97 (94 - 99)	97 (95 - 100)
Oxygen saturation	100 (98 - 100)	100 (97 - 100)
Systolic blood pressure	100 (98 - 100)	100 (97 - 100)
Temperature	90 (84 - 95)	82 (68 - 88)
Neurological status	93 (85 - 98)	91 (69 - 97)
Oxygen support	99 (98 - 100)	99 (96 - 100)

Table 2.5 summarises the manual vital-sign data collected during CALMS-3. The number of manual observations were similar between the "normal" and "abnormal" patient groups. Data completeness was much lower in the "abnormal" patient group than the "normal" patient group, which was not the case in the CALMS-2 study. This was mostly due to temperature being unreliably recorded in the "abnormal" group. Clinical observations of neurological status were also documented unreliably.

2.4 Summary

Table 2.6: Summary of the vital-sign data collected during the CALMS-2 and CALMS-3 clinical studies. Here, vital signs recorded within 24 hours of an adverse clinical event are considered to be "abnormal" and all other vital-sign observations are considered to be "normal".

	CALMS-2	**CALMS-3**	**Total**
Patient Numbers			
Total	407	447	854
Recovered without an event	357	298	655
Suffered an event	50	149	199
Manual vital-sign data			
Total observations	32961	30955	63916
"Normal" observations	32087	28938	61025
"Abnormal" observations	874	2017	2891
Continuous vital-sign data	Yes	No	-

An overview of the CALMS-2 and CALMS-3 vital-sign data is given in table 2.6. The major differences between the data in the two studies is that the CALMS-2 study collected both continuous and manual vital-sign data, while the CALMS-3 study only collected manual data from clinical observations. Furthermore, the number of patients who suffered an adverse clinical event was significantly higher in the CALMS-3 study compared to the CALMS-2 study.

The CALMS-2 and CALMS-3 datasets were both collected from patients recovering from major upper-gastrointestinal cancer surgery. The manual vital-sign observations from the two datasets can therefore be used in combination for conducting analyses and deriving risk prediction models of acute post-operative patients. Indeed, several of the analyses in this thesis use the data in this way. For example, in chapter 3 a machine learning algorithm is trained and validated on data from the CALMS-3 dataset and tested on data from the CALMS-2 dataset, thus using the data sources independently.

Chapter 3

Diurnal vital-sign variability in post-operative patients

The complexity of EWSs has increased since their introduction in the 1990's, moving from single parameter scores to multi-parameter scores (see section 1.3.1), and then to data-driven approaches with machine learning (see section 1.3.2). However, most EWSs are still calculated from physiological parameters measured at a single point in time. Such an approach may have been a practical necessity in the past, as the complexity of calculating EWSs from previous data points on paper charts may have led to errors. With the widespread adoption of digital systems such as electronic health records (EHRs), computerized EWS calculations allow the design of EWSs with unrestricted computational complexity. The inclusion of longitudinal trends from physiological measurements at multiple time-points could improve post-operative EWS performance (Petit et al. [2018]).

The importance of longitudinal data is intuitive in the context of EWSs. Indeed, the commonly-used term "physiological deterioration" (Escobar et al. [2012]) denotes an interest in the changes of patient condition over time. Clinicians often consider the trends in vital-sign values when assessing the physiological condition of a patient (Churpek et al. [2016]). Furthermore, trends in risk-score values are predictive of patient outcomes (Cabré et al. [2005]; Ferreira et al. [2001]; Kim et al. [2015]; Wong et al. [2011]). Rather than relying on clinicians to interpret the trends in physiological values, the incorporation of longitudinal data into an EWS may improve its ability to identify clinically-important changes over time.

The importance of longitudinal data also has a physiological basis. Most variables of clinical interest show predictable changes with the body's natural circadian (24-hour) rhythm. For example, blood pressure presents an increase in the morning, a small postprandial trough and a deeper descent during nocturnal rest[1] (Fernández et al.

[1]This is true for the young and healthy; with age the nocturnal trough in blood pressure disappears, becoming a nocturnal rise (towards the end of the night) for those over 60 (Ayala et al. [2013])

[2009]; Millar-Craig et al. [1978]). Likewise, body temperature undergoes circadian variations associated with the sleep-wake cycle (Refinetti and Menaker [1992]). Thus, the value of a physiological measurement should ideally be interpreted in the context of its natural cycle. It may even be informative to model the divergence of a physiological variable from its usual cycle as a predictor of physiological deterioration (Fernández et al. [2009]), in addition to the usual practice of comparing individual measurements to known limits of normality. Indeed, patterns of variability in physiological signals over both shorter (ultradian) and longer (circadian and infradian) time scales can be important markers of disease (Goldberger [2001]; Seely and Macklem [2004]).

In this chapter we compute features of diurnal[1] vital-sign variability and assess their value for improving the predictive performance of EWS models. We use machine learning methods to develop an EWS model, using the CALMS-3 dataset for training and validation, before testing the resultant model on the CALMS-2 dataset. We begin by describing previous work using diurnal trend and variability features to predict clinical outcomes (section 3.1). We then present methods for computing vital-sign variability features and assess their importance as predictors of physiological deterioration in the CALMS-3 dataset (section 3.2). The best features are then used in a machine learning model, which is trained and validated on the CALMS-3 dataset, before being tested and compared to previous benchmark EWS systems on the CALMS-2 dataset (section 3.3).

3.1 Previous work

Measures of short term vital-sign variability are known to be biomarkers for predicting disease. Decreases in heart rate variability (HRV), which measures the variability in the time between successive heart beats, are strongly correlated with mortality after acute myocardial infarction (Kleiger et al. [1987]). Clinical risk scores have been developed using features of HRV to predict conditions such as neonatal sepsis (Xiao et al. [2010]) and complex mathematical functions have been proposed to better encode the characteristic changes in variability that occur during critical illness (Bravi et al. [2011]; Richman and Moorman [2000]; Seely and Macklem [2004]). Notably, these short-term variability biomarkers require high-frequency continuous waveform data and are typically used to monitor the critically ill, or specific populations such as cardiac patients. Biomarkers of longer-term vital-sign variability (e.g. diurnal

[1]Processes with 24-hour oscillations are generally called diurnal rhythms, while circadian rhythms (which may also have a 24-hour cycle) must be endogenously generated.

variability) in wards outside of critical care remain relatively unexplored.

Measures of variability and temporal trends in clinical biomarkers have been related to clinical outcomes in previous studies. For example, trends in the daily computations of an organ failure risk score have been shown to predict mortality (Cabré et al. [2005]; Ferreira et al. [2001]). Significant trends in Modified Early Warning Score (MEWS) values were seen prior to cardiac arrests in a general ward (Kim et al. [2015]). Chen et al. [2009] found that vital-sign trends did not predict haemorrhage in a trauma ward, but the vital signs used to compute the trend values were only measured over a 20-minute period prior to hospitalisation. Wong et al. [2011] explored the value of adding risk score trends as a predictor in a survival model, finding that although the trend values were independently associated with time-to-hospital mortality, the inclusion of the trend indicators in the model did not improve predictive performance.

Long-term vital-sign variability and trend features have been included as predictive features in clinical risk scores. Previous studies have used a "bucketing" technique, in which predictive features are computed within temporally-distributed time "buckets", thus encoding the temporal effects in the risk model (Bailey et al. [2013]; Mao et al. [2011]). Alvarez et al. [2013] included the most "abnormal" vital-sign value during the previous 24 hours as a predictive feature in a risk model, which did not explicitly model vital-sign variability but may have captured some predictive diurnal variability effects. Escobar et al. [2012] designed an EWS that included features of 24-hour vital-sign trends, but did not report the importance of these features within the model. Churpek et al. [2016] also designed an EWS model with vital-sign trend features, showing that the inclusion of vital-sign trends significantly improved predictive performance, while also reporting the importance of each vital-sign trend feature. The trend features in this study were computed over the previous six vital-sign values, which the author claimed would approximate a 24-hour time-window due to the typical measurement interval of 4 hours. However, vital-sign measurement frequency commonly increases during physiological deterioration (Ghosh et al. [2018]) so the trend features may have been computed over different time windows prior to an adverse event. Thus, the trend features in this study may not have been a true measure of diurnal effects.

Overall, vital-sign trends have been shown to be predictive of clinical outcomes and there is evidence that the inclusion of vital-sign trends in risk scores has predictive value. However, the methods used have mostly tried to capture a general temporal effect rather than explicitly modelling vital-sign variability within a diurnal cycle. Further work is needed to understand the value of diurnal vital-sign variability in predicting adverse clinical events in post-operative care.

3.2 Univariate vital-sign variability analysis

3.2.1 Overview

We began our analysis by assessing the importance of the five main vital signs and their diurnal variability for predicting physiological deterioration. To do so, we considered each vital sign and their diurnal variability as features within a machine learning model. We then used the model performance as a measure of the predictive importance of the feature used as the model input. This process is described in detail in sections 3.2.2 and 3.2.3. For this analysis we used the CALMS-3 dataset, and applied the framework described below for training and testing machine learning algorithms as EWSs.

Machine learning approach

As described in section 1.3.2, a common methodology used in the design and evaluation of EWSs is to frame the problem of predicting an adverse clinical event as a binary classification task. In this framework, all vital-sign observations recorded within a given time-interval before an event are labelled as "positive" and all vital-sign observations recorded at other times are labelled as "negative". Hence, the task is to classify each vital-sign observation as positive or negative. In this chapter, we assess the predictive importance of the vital signs and the diurnal variability of the vital signs as features within a binary classification machine learning algorithm.

We used a time interval of 24 hours, assuming that the onset of physiological deterioration begins 24 hours in advance of each adverse event. In reality, the diagnosis of each adverse event may have occurred very shortly after the onset of physiological deterioration, or it may even have occurred after a much larger interval (which may even be greater than 24 hours), so a limitation of this approach is that some vital-sign observations will have been mislabelled. However, it is impossible to label vital-sign observations perfectly because the underlying health status of the patient at each point in time is unknown. 24 hours is a common choice of time interval (Romero-Brufau et al. [2014]; Smith et al. [2013]; Tarassenko et al. [2006]) that allows a consistent benchmark for comparing different EWS models.

In summary, the machine learning task is to correctly label each vital-sign observation as positive (an adverse event occurs within the next 24 hours) or negative (an adverse event does not occur within the next 24 hours). For this analysis, we define an adverse event as an unexpected admission to the ICU or death on the ward. The vital-sign observation labelling process is shown in figure 3.1.

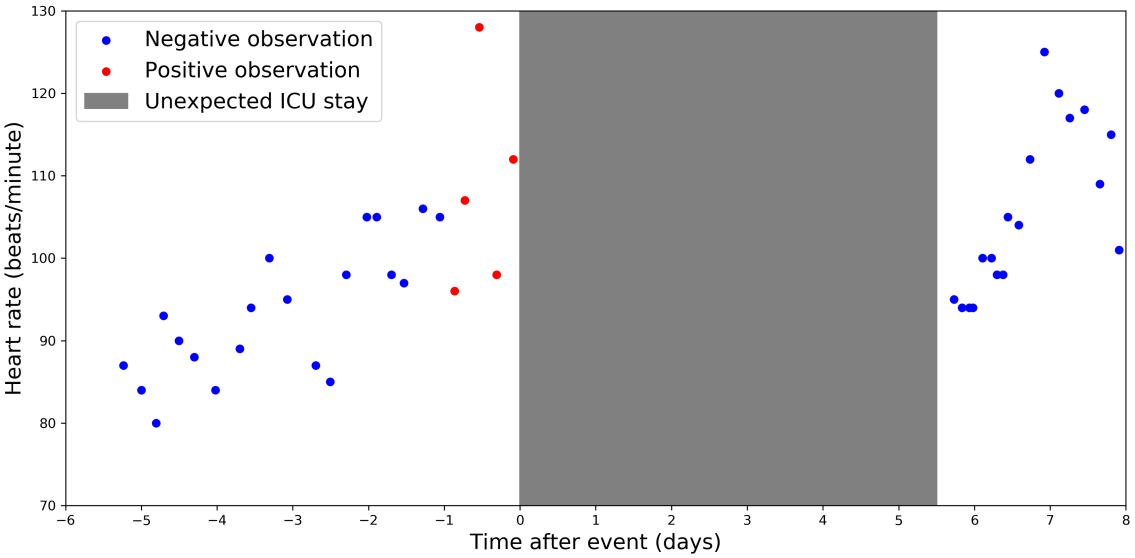


Figure 3.1: Example time-series of heart-rate observations before and after an adverse event (unexpected admission to the ICU). The data labelling process for the binary classification task is shown, with heart-rate observations within 24 hours before the event shown in red (positive label) and all other heart-rate observations shown in blue (negative label). Time in the ICU is shaded in grey, when no vital-sign observation data were available as the patient was not on the ward.

Analysis pipeline

A number of methodological steps were used to estimate the importance of each vital-sign and vital-sign variability feature for predicting adverse clinical events. The steps of the analysis are shown schematically in figure 3.2 and are listed as follows. The steps are then described in detail in sections 3.2.2 and 3.2.3.

- Remove erroneous vital-sign observations from the CALMS-3 dataset.
- Compute features of diurnal vital-sign variability by subtracting the minimum vital-sign value from the maximum vital sign within the previous 24-hour window, for each vital-sign observation.
- Split the dataset into 5 different independent training and testing sets such that vital-sign observations from a given patient are in either the training set or the test set for each fold.
- For each feature, train a logistic regression model with restricted cubic splines and compute the area under the receiver operating characteristic curve (AUROC) (see section 3.2.3) – repeat for each fold. Feature importance is computed as the "conservative mean" (Khosla et al. [2010]) AUROC over the folds.

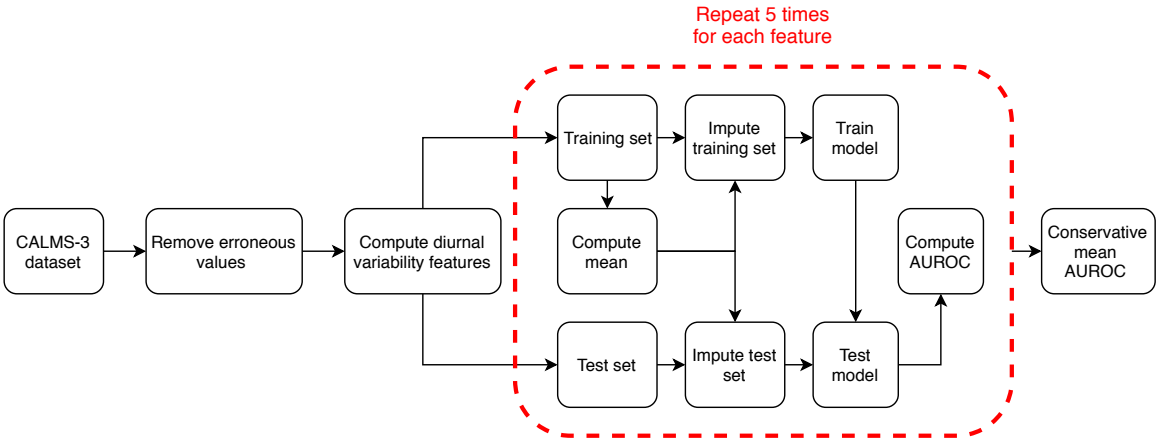


Figure 3.2: The analysis pipeline for estimating the importance of each vital-sign and vital-sign variability feature. Erroneous vital-sign values were removed based on limits of plausible physiological values. Variability features were computed as the difference between the maximum and minimum vital-sign values over the previous 24 hours. A 5-fold data partition scheme was used to split the dataset into training and test sets, with data stratified by patient. Mean imputation and model training and testing were conducted within each fold, using a logistic regression model with restricted cubic-spline interpolation for each feature. Feature importance was the "conservative mean" AUROC over the five folds.

3.2.2 Model design

For the analysis of vital-sign and vital-sign variability feature importance, we used a logistic regression model with restricted cubic-spline interpolation (see below) to allow the probability of the event to increase non-linearly with the feature value. As described in section 3.2.1, we built a separate model for each vital-sign and variability feature. The steps to build each model are described as follows.

Pre-processing

Table 3.1: Limits of physiological plausibility for the 5 main vital signs. Vital-sign observation values outside of these limits were removed from the CALMS-3 dataset.

Vital sign	Lower limit	Upper limit
Heart rate (beats/minute)	30	300
Respiratory rate (breaths/minute)	3	45
SpO_2 (%)	60	101
Temperature (°C)	32	42
Systolic blood pressure (mmHg)	40	300

We began our analysis by ensuring that all vital-sign observation values are physiologically plausible. To do so, we removed all vital-sign observations from the CALMS-3 dataset that were outside the physiological limits shown in table 3.1.

Calculating vital-sign variability features

Vital-sign variability values were computed for each observation set by taking all vital-sign observations in the preceding 24 hours (from the time of the given observation set) and computing the difference between the maximum and minimum values. If fewer than 4 vital-sign observations were recorded within the preceding 24-hour window, then no diurnal variability feature was computed for that observation set and the variability from the previous observation set was used. Variability values were only forward-imputed once, i.e. if two or more consecutive variability values could not be computed, then only the first missing variability value was imputed from the previous variability computation. Furthermore, no variability values were computed for the vital-sign observation sets from the first 24 hours of the patient's stay on the ward.

We used this methodology to compute the variability features as it encodes whether a patient shows more or less diurnal changes in their vital-sign values than normal. For example, blood pressure normally has a peak and trough during day-time, so this method captures whether the trough-to-peak range is more or less pronounced. This variability estimation methodology also captures trends of deteriorating physiology when there is a large difference between the baseline vital-sign value and any subsequently large increases or decreases in the value of that vital sign. This method of estimating variability has not been used in previous EWS models to our knowledge. Previous methods have included separate features for the maximum and minimum values, the difference between the first and last value, and the slope of previous values within a given time-window (Churpek et al. [2016]). Our method is designed to explicitly encode differences in the peak-to-trough range in diurnal vital-sign values, which are of primary interest in this analysis.

Post-processing

Following the removal of physiologically implausible vital-sign values and the computation of diurnal vital-sign variability, there were missing feature values. We used mean imputation as a post-processing step to impute the missing values, conducting this step separately for the training set and the held-out test set in every fold (see section 3.2.3). To do so, the sample mean value of each feature was computed from

the training set, which was then used to fill the missing values in both the training set and the test set. An alternative method would have been to use the mean values from the entire dataset for all patients (in both training and testing). However, this method would introduce information in the test set based on data from the patients in the training set, and this would bias the results. The training set and the test set must remain independent.

Cubic-spline logistic regression

We used the methodology of Harrell [2015] to compute logistic regression models with restricted cubic-spline interpolation for each vital-sign and vital-sign variability feature. Restricted cubic splines are variable transformations that allow the relationship between the predictor and outcome variables to be non-linear, and are appropriate for vital signs that can vary above and below limits of normality. For example, both low and high values of heart rate may be predictive of physiological deterioration, so there is a non-linear relationship between heart rate values and risk of an adverse event.

Cubic splines fit piecewise local cubic polynomials between knots (transition points). Parameters are fitted to ensure smooth transitions between local polynomials such that the polynomial values, and first and second derivatives, match at the knots. Restricted cubic splines operate similarly, except with restrictions to make the fitted function linear at the tails by omitting the non-linear terms for the first and last spline functions. The number of knots dictates the number of local polynomials used, and therefore the complexity of the overall fitted function. Restricted cubic spline functions are calculated as follows.

$$g(x) = \beta_0 + \sum_{i=1}^{k-1} \beta_i X_i \tag{3.1}$$

$$X_{j+1} = \begin{cases} x & j = 0 \\ (x - t_j)_+^3 - (x - t_{k-1})_+^3 (t_k - t_j)/(t_k - t_{k-1}) + \\ \qquad (x - t_k)_+^3 (t_{k-1} - t_j)/(t_k - t_{k-1}) & j = 1, ..., k-2 \end{cases} \tag{3.2}$$

where x is the feature value, $g(x)$ is the fitted cubic spline function, and X_j are the polynomial terms in the spline function. t_k are the positions of the k knots. β are the spline parameters. The plus subscript denotes that the value is 0 when the term is negative (equivalent to multiplication by the Heaviside function).

To compute the likelihood of an adverse event from the cubic spline interpolation of the vital-sign or vital-sign variability feature, we used logistic regression. This was calculated using the sigmoid function, as follows.

$$P(x) = \frac{1}{1 + e^{-g(x)}} \tag{3.3}$$

where $P(x)$ is the likelihood of an adverse event for the feature value x.

We computed the positions of the knots from percentiles of the training set. Knot positions were recomputed for each fold of the dataset. We used the percentiles recommended by Harrell [2015], which are given table 3.2. We chose the number of knots for each feature as described in section 3.2.3.

Table 3.2: Percentiles for computing the knot positions in the restricted cubic splines logistic regression model (Harrell [2015]).

knots	**Percentiles (%)**						
3			10	50	90		
4			5	35	65	95	
5		5	27.5	50	72.5	95	
6	5	23	41	59	77	95	
7	2.5	18.33	34.17	50	65.83	81.67	97.5

Optimisation

To derive the spline parameters β in the logistic regression model, we used the LogisticRegression function from the Scikit-learn (Pedregosa et al. [2011]) package in Python. Our configuration found a maximum likelihood solution to the parameter values by solving the primal formulation of the unconstrained optimisation problem given in equation (3.4) with the LIBLINEAR library (Fan et al. [2008]), which uses a trust-region Newton optimisation method (Lin et al. [2007]).

$$\underset{minimise}{w} \quad \frac{1}{2} w^T w + C \sum_{i=1}^{N} log(1 + e^{-y_i w^T x_i}) \tag{3.4}$$

where w is the vector of parameters β, x_i are the feature values, of which there are N, and y_i are the ground-truth labels (0 for negative and 1 for positive). C is the inverse-regularisation parameter – we used a large value (10^{10}) for no regularisation.

3.2.3 Model evaluation

We used the classification performance of the model for each feature as the measure of feature importance. We assigned the data randomly into training and test sets (as described below) to compute the mean and standard deviation of the AUROC over each fold, repeating this for numbers of cubic spline knots ranging from 3 to 7. We computed the "conservative mean" for each number of knots by subtracting the standard deviation of the AUROCs from the mean of the AUROC. This methodology was recommended by Khosla et al. [2010] for selecting features in situations with unbalanced datasets, to account for variations due to sampling over a small number of non-homogeneous examples. For each feature, the number of knots that resulted in the highest conservative mean AUROC was chosen, and this value was used as a measure of feature importance.

Receiver-operating characteristics

Receiver operating characteristic (ROC) curves are used to illustrate the diagnostic ability of a binary classification algorithm for different decision threshold values. The decision threshold is the value above which the model output (for example the EWS score or predictive likelihood) is deemed to be a positive prediction. During evaluation of a binary classifier, model outcomes can be true positives, true negatives, false positives or false negatives depending on whether the predicted outcome matches the ground-truth labels. The proportion of true or false, positive or negative model outcomes depends on the decision threshold used. ROC curves allow the model performance to be assessed for all possible decision thresholds at once.

ROC curves are created by plotting the true positive rate (TPR), which is also known as the sensitivity, against the false positive rate (FPR), which is the same as (1 - specificity). TPR is the proportion of true positives that are identified by the classifier; FPR is the proportion of negatives that were incorrectly labelled by the classifier, and specificity is the proportion of true negatives that are identified by the classifier (Altman and Bland [1994]). In clinical terms, a high sensitivity is important to ensure that most instances of physiological deterioration are detected, while a high specificity is important to ensure that false alarms do not have an unmanageable burden on clinical staff. Equation (3.5) shows how TPR and FPR are computed (Bishop [2006]). The AUROC is a commonly used metric for summarising the performance of a binary classification algorithm independently of the decision threshold value (Hanley and McNeil [1982]). It can be viewed as the discriminatory power of the model over all

possible operating points (decision thresholds). The AUROC is a value between 0 and 1, with 0.5 or below equivalent to random guessing, and 1 representing perfect TPR and FPR across all operating points. The AUROC can be calculated using standard numerical methods (e.g. the trapezoidal rule) for computing a definite integral.

$$TPR = \frac{TP}{TP + FN} \ , \qquad FPR = \frac{FP}{FP + TN} \tag{3.5}$$

where TP, FP, TN, FN are the number of true positive, false positive, true negative and false negative predictions, respectively.

Creating folds for training and testing

We used a scheme to create random splits of the dataset such that the data from a given patient were either present in the training set or the test set, but not both, for each fold. We also ensured that the training and test sets would be approximately balanced in terms of the proportion of positive or negative ground-truth labels (observation sets). Our implementation used the StratifiedKFold function in the Scikit-learn (Pedregosa et al. [2011]) package in Python to generate the five folds as follows.

We first grouped all patients as "abnormal" (suffered an adverse event) or "normal" (no adverse event). For each group of patients, we created 5 random and unique subsets of patients, which were approximately equal in length (for example, of the 149 "abnormal" patients, there were 4 groups of 30 and 1 group of 29). By random we mean that the data were shuffled using a random number generator rather than being grouped chronologically. The data subsets for each group were then combined in a pairwise manner (i.e. the first subset of "abnormal" patients and the first subset of the "normal" patients were combined, and so on) to create 5 unique subsets of all of the patients. These groupings were the test sets of the 5 folds. The training set for each fold was created as the data from all patients not in the test set. Thus, patients appeared in multiple training sets but only once in the test set.

This methodology ensured that an equal proportion of "abnormal" and "normal" patients appeared in each fold. However, it did not ensure that the various sets contained equal proportions of positive or negative-labelled vital-sign observations. This was not an issue, as there were a large number of patients in each fold such that any per-patient differences in the proportion of positive to negative labels (caused by differences in length of stay on the ward) were similarly distributed in each set.

Risk curves

To visualize the fitted models for each feature, we created risk curves – plots of the predicted likelihood of an adverse event across the range of values in the dataset. We created risk curves for the data in each fold by fitting the model to the training set, creating a vector spanning the range of values in the dataset, and computing the predicted likelihood of an event $P(x)$ (equations (3.1) to (3.3)) for each value in the vector. The resultant vectors of predicted likelihoods (of which there was one for each of the five folds) were then plotted against the vector of feature values. This allowed us to examine the model-fit, and assess whether feature values with high risk of an adverse event were consistent with prior clinical knowledge. As described before, the number of knots used in the models for each feature were chosen based on the highest conservative mean AUROC across the five folds.

3.2.4 Results

Table 3.3 shows the number of knots and feature-importance results for each vital-sign and vital-sign variability feature. Figure 3.3 shows the ROC curves for the five folds, for each feature. Likewise, figure 3.4 shows the risk curves for each feature.

Table 3.3: AUROC results for the 5-fold feature importance analysis of each vital sign and vital-sign variability feature. The model was logistic regression with restricted cubic spline interpolation. The number of spline knots is shown for each feature, which were chosen to optimise the conservative mean AUROC (the mean minus the standard deviation) across the five folds. Features are listed in order of decreasing conservative mean AUROC. Here "var." refers to variability.

Feature	Knots	AUROC	
		Mean (Standard deviation)	Conservative mean
Respiratory rate var.	3	0.764 (0.025)	0.739
SpO_2 var.	5	0.749 (0.027)	0.723
Respiratory rate	7	0.744 (0.035)	0.709
Systolic blood pressure var.	3	0.716 (0.016)	0.700
Heart rate	7	0.738 (0.046)	0.692
Heart rate var.	3	0.715 (0.029)	0.686
Temperature	6	0.607 (0.005)	0.602
SpO_2	4	0.613 (0.032)	0.581
Systolic blood pressure	3	0.607 (0.032)	0.575
Temperature var.	5	0.577 (0.042)	0.536

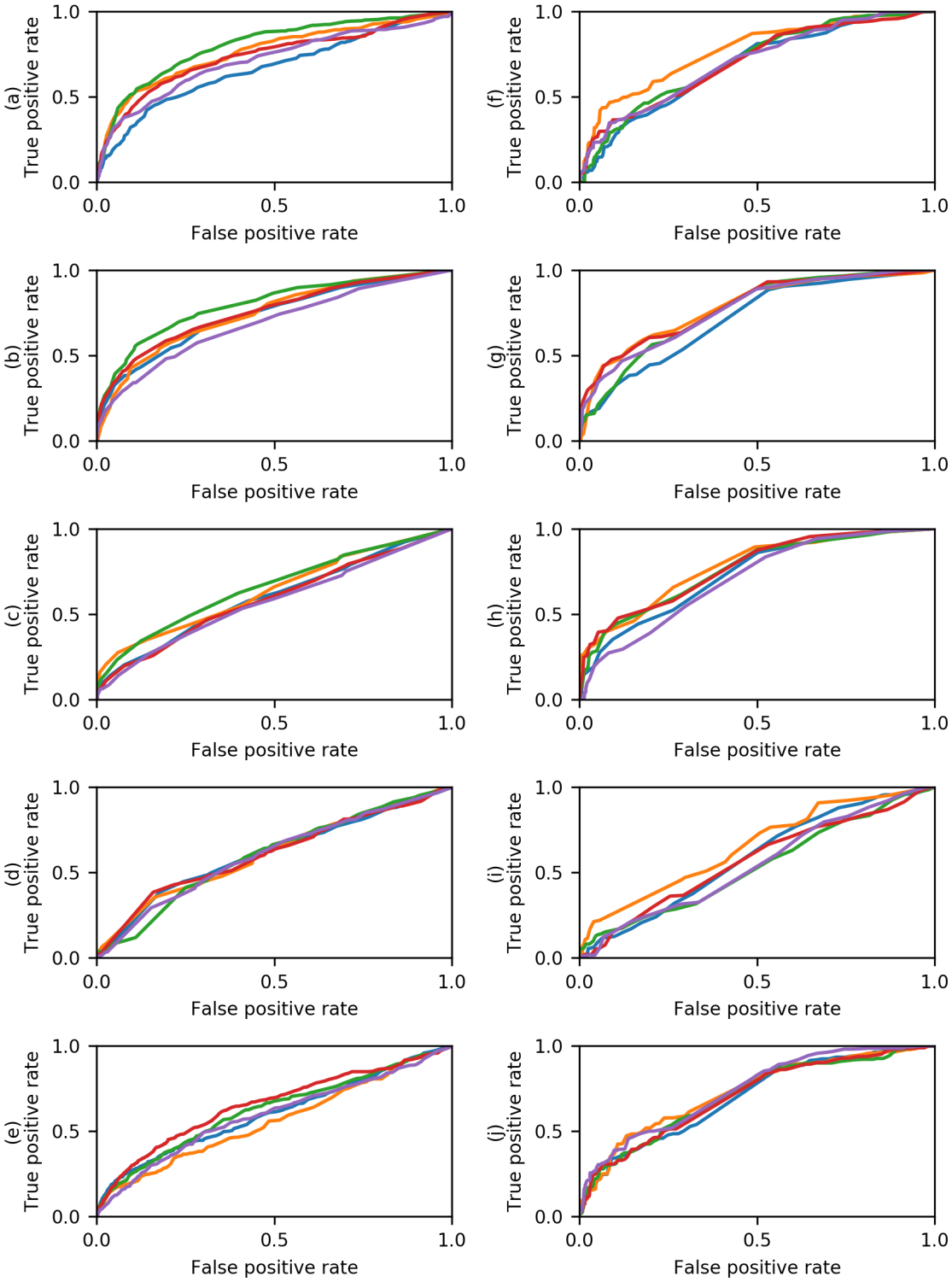


Figure 3.3: ROC curves for the test set in the five folds, for each vital sign and vital-sign variability feature. Curves are shown for heart rate (a), respiratory rate (b), SpO_2 (c), temperature (d), systolic blood pressure (e), heart rate variability (f), respiratory rate variability (g), SpO_2 variability (h), temperature variability (i), and systolic blood pressure variability (j).

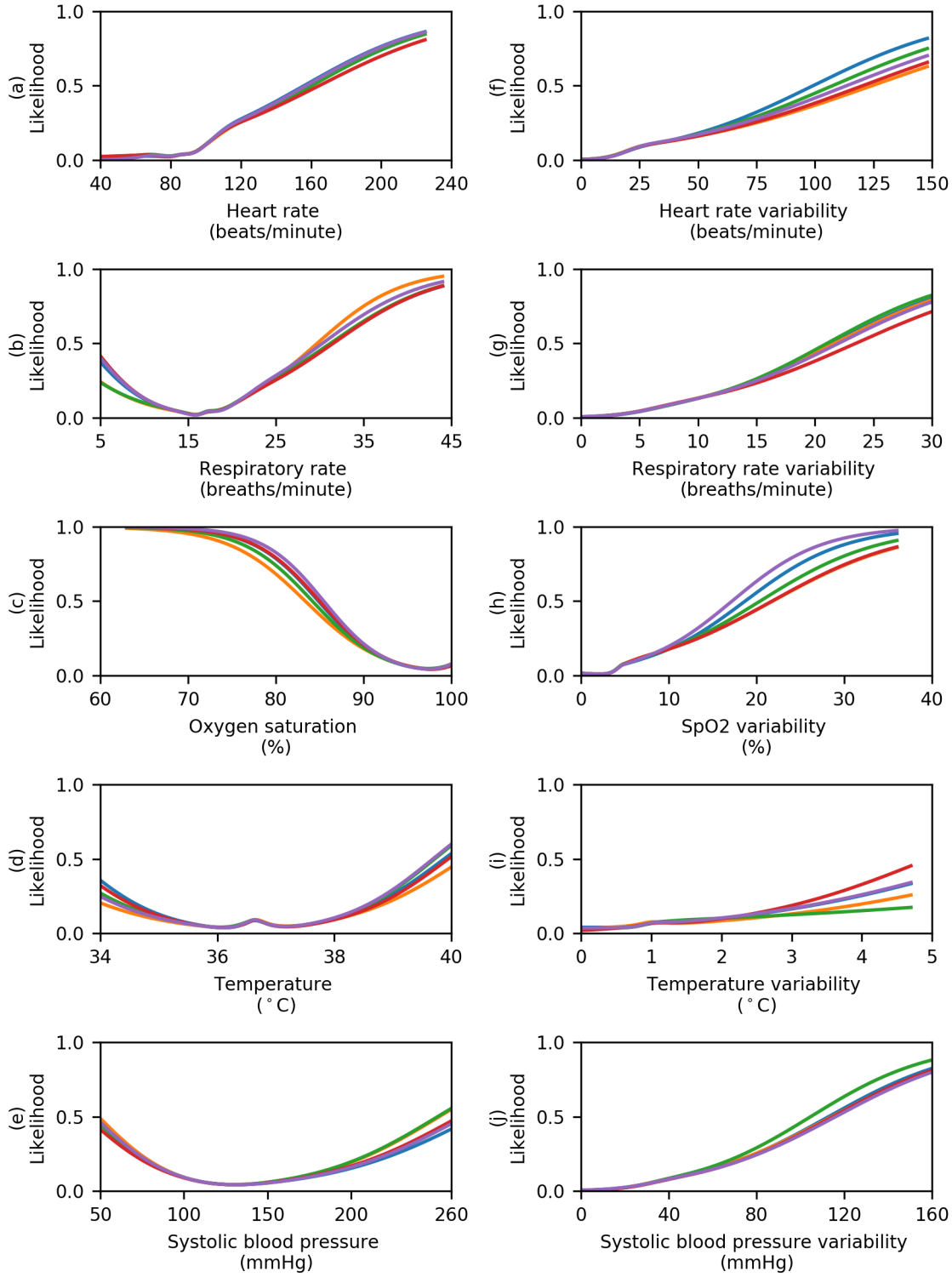


Figure 3.4: Risk profiles showing the predicted event likelihood against the vital-sign or vital-sign variability feature value for the models fitted on the training set of the five folds. Plots are shown for heart rate (a), respiratory rate (b), SpO_2 (c), temperature (d), systolic blood pressure (e), heart rate variability (f), respiratory rate variability (g), SpO_2 variability (h), temperature variability (i), and systolic blood pressure variability (j).

3.2.5 Discussion

Features of diurnal vital-sign variability were highly predictive of adverse clinical events in the CALMS-3 dataset. For example, respiratory rate variability had a mean (standard deviation) AUROC over the five folds of 0.764 (0.025) and SpO_2 variability had a mean (standard deviation) AUROC of 0.749 (0.027). Systolic blood pressure variability and heart rate variability were also highly predictive of physiological deterioration with mean AUROC values over 0.7, although heart rate variability had a high standard deviation and therefore had a slightly lower conservative mean value. Temperature variability was the least predictive feature. All features had AUROCs above 0.5 and therefore performed better than random guessing.

Interestingly, features of vital-sign variability were more predictive of adverse clinical events than vital-sign features. Respiratory rate variability and SpO_2 variability were the highest ranking features in terms of conservative mean AUROC values. For SpO_2 and systolic blood pressure, the variability features were significantly more predictive than the vital-sign features, with more than a 10% increase in conservative mean AUROC. The results suggest that the diurnal fluctuations in vital-sign values are important indicators of patient health, and that deviations from the normal diurnal patterns are related to physiological deterioration.

There may be other factors that partly explain the predictivity of the vital-sign variability features in our experiments. Baseline normal vital-sign ranges differ between patients depending on demographic factors such as age. Our variability estimation method (maximum - minimum vital-sign value over the previous 24 hours) would encode changes from baseline values and may therefore be a proxy method of normalising vital-sign values to a patient-specific baseline level. While this was not the intended purpose of our methodology, patient-specific features may be desirable in a clinical decision system so this aspect of the methodology is appropriate for an EWS. The variability estimation method may also encode trends in the trajectory of the vital sign, thus partly predicting future changes in the vital-sign value. This is also a desirable feature property for an EWS.

The risk profiles plotted in figure 3.4 show that the risk of an adverse event increases monotonically with vital-sign variability for values above a certain threshold. For example, the predicted likelihood of an event remains flat for SpO_2 variability values below 4.0, but increases significantly for values above 4.0. The risk profiles for the vital signs show higher risks in the expected vital-sign ranges, except for heart rate, which does not have an increased risk for heart rate values corresponding to

bradycardia. Most EWSs have high scores for low heart rate values, so this effect is not consistent with the clinical intuition used when developing these scores. It may indeed be related to the low prevalence of bradycardia in the dataset. Less than 0.1% of heart rate values were below 50 beats/minute, so values corresponding to bradycardia will not have had a significant contribution to the ability of the model to predict adverse events. Another unexpected finding in the risk profiles is the slight rise in the risk of an adverse event for temperature values between 36 and 37 degrees Celsius. The reason for this effect is unclear, although it is unlikely to have affected the model performance since temperature was the least predictive feature.

Different numbers of knots were required for different features. This was related to the linearity of the risk profiles for the features shown in figure 3.4. For example, respiratory rate variability is almost linear in its relation to the predicted likelihood of an event, and thus only required 3 knots to model the changes in the likelihood for different values. In contrast, respiratory rate was highly non-linear in relation to the predicted likelihood of an event, and thus performed best in terms of conservative mean AUROC by using 7 knots in the cubic spline interpolation.

A limitation of this analysis was the univariate model design, as some of the features will have encoded the same physiological phenomena. Thus, some feature combinations within a multivariate EWS model may not be significantly more predictive than one feature if they all detect similar patterns in physiology. However, the univariate methodology did allow the assessment of features independently, and was thus appropriate for measuring the predictive ability of each feature. A limitation of the univariate assessment of SpO_2 can be seen in figure 3.4c, which shows that the predicted likelihood of an adverse event does not continue to decrease for very high SpO_2 values. High SpO_2 values in post-operative care are often caused by the use of supplemental oxygen, which is given to patients with reduced respiratory function. These patients therefore have abnormal physiology, despite having high values of SpO_2. The increased likelihood of an event at high SpO_2 values is likely to be due to this effect, as patients with high SpO_2 are more likely to have supplemental oxygen support and therefore an increased risk of an adverse event. Many EWSs encode this effect by including a score for the use of supplemental oxygen.

Our results suggest that features of vital-sign variability have significant value for the prediction of adverse clinical events in post-operative care. The inclusion of vital-sign variability features in an EWS model may therefore improve model performance. Further experiments are needed to test on an independent test set the performance of an EWS model that includes vital-sign variability features.

3.3 Building an improved machine learning EWS

3.3.1 Overview

In this section, we develop a machine learning based EWS model that includes features of vital-sign variability. We use the logistic regression model with restricted cubic spline interpolation described in section 3.2, and compare the performance to several EWS models that have been described in the EWS literature. We also compare the performance of the model with and without features of vital-sign variability. We use the CALMS-3 dataset for training and the CALMS-2 dataset for testing.

3.3.2 Model design

We trained and tested a model with multiple vital-sign and vital-sign variability features, as well as a model using only the vital signs. Both models also included a categorical feature for neurological status (using the AVPU scale described in section 1.1.2) and a binary feature indicating whether the patient was receiving supplemental oxygen support. The steps for building the models are described below.

Pre-processing

As in section 3.2, we began by ensuring that all vital-sign values were physiologically plausible. We removed all vital-sign observations from the CALMS-3 and CALMS-2 datasets that were outside the physiological limits shown in table 3.1. For missing values of the supplemental oxygen feature, we assumed that the patient was not receiving oxygen support (binary feature value of 0). We encoded the neurological status feature as four separate binary "dummy variables". For each observation the binary feature for the AVPU score was set to 1 (e.g. an "A" score would correspond to an "A" feature value of 1), while the other binary features were set to 0. If the AVPU score was missing, all the binary "dummy variables" were set to 0.

Post-processing

The vital-sign variability features were calculated using the same methodology as described in section 3.2.2. Mean imputation was again performed by using the mean of the features in the training set to impute the missing values in both the training and test sets. When training and testing the model, this was performed by computing the mean from the CALMS-3 dataset and applying this value to impute the missing

values for both the CALMS-3 dataset and the CALMS-2 dataset. For feature selection, as described below, we again computed the mean from the training set in each fold, and used this for mean imputation in the training and test sets for that fold.

Feature normalisation was applied to standardise the range of the continuous features. Feature normalisation is a common step in machine learning, as it allows for more efficient convergence of gradient-based multivariate optimisation methods. For each feature, we subtracted the mean and divided by the standard deviation, resulting in features with zero-mean and unit-variance. Just as with mean-imputation, we calculated the mean and the standard deviation from the training set and used these values for scaling the features in both the training set and the test set.

Mean-imputation and feature normalisation were only applied for the continuous features (vital signs and vital-sign variability features). There were no missing values of the oxygen support and neurological status features once the pre-processing steps were completed. These features are binary and therefore have the same order of magnitude as the continuous variables (after normalisation), so did not require normalisation.

Multivariate logistic regression with cubic splines interpolation

We adapted the logistic regression model with cubic spline interpolation described in section 3.2 to allow multiple features. In the univariate setting, the feature was first expanded to several polynomial variables (equation (3.2)), which were combined in a weighted sum to form the interpolation function (equation (3.1)). The sigmoid function was applied to compute the predicted likelihood of an adverse event (equation (3.3)). We repeat the calculation of the polynomial variables for convenience:

$$X_{j+1} = \begin{cases} x & j = 0 \\ (x - t_j)_+^3 - (x - t_{k-1})_+^3 (t_k - t_j)/(t_k - t_{k-1}) + & \\ \qquad (x - t_k)_+^3 (t_{k-1} - t_j)/(t_k - t_{k-1}) & j = 1, ..., k-2 \end{cases} \tag{3.6}$$

where x is the vital-sign value, and X_j are the polynomial terms in the spline function. t_k are the positions of the k knots. The plus subscript denotes that the value is 0 when the term is negative (equivalent to multiplication by the Heaviside function).

In the multivariate setting, we compute the polynomial variables for each continuous feature. We combine the polynomial variables and the non-continuous variables (such as the binary supplemental oxygen support feature) in a weighted sum. The sigmoid function is then applied to the weighted sum of all the polynomial variables and

non-continuous variables. This is calculated as follows.

$$f(x) = \beta_0 + \sum_{w=1}^{W} \sum_{i=1}^{k_w - 1} \beta_{w,i} X_{w,i} + \sum_{u=1}^{U} \beta_u x_u \tag{3.7}$$

$$P(x) = \frac{1}{1 + e^{-f(x)}} \tag{3.8}$$

where $P(x)$ is the predicted likelihood of an event. $X_{w,i}$ are the polynomial terms in the ith spline function for the wth continuous feature (of which there are W). k_w is the number of knots for the wth feature. x_u are the non-continuous features, of which there are U. β are the model parameters.

Optimisation

The number of knots for each vital-sign and vital-sign variability feature were the optimal number of knots computed in the univariate analysis of feature importance (see table 3.3). We optimised the parameter values β of the logistic regression model using the Scikit-learn (Pedregosa et al. [2011]) package in Python, as described in section 3.2. We also optimised the model that combined both vital-sign and vital-sign variability features by choosing the best combination of features for inclusion in the model. To do so, we used a feature selection scheme based on the importance of the vital-sign and vital-sign variability features given in table 3.3.

In our feature selection methodology, we designed models of increasing complexity and evaluated their performance using 5-fold data partitioning on the CALMS-3 dataset. The simplest model had features for respiratory rate variability, which was the most important feature in terms of predicting adverse clinical events (see table 3.3), as well as features for neurological status and supplemental oxygen support. Subsequent models were then designed by adding features in order of decreasing importance. For example, the second model added SpO_2 variability to the first model and the final model included all vital-sign and vital-sign variability features.

We performed 5-fold cross-validation on the CALMS-3 dataset for each model, using the scheme described in section 3.2.3. The model with the highest conservative mean AUROC value was selected. This model was re-trained on the entire CALMS-3 dataset and tested on the CALMS-2 dataset. We also trained and tested a model with vital-sign features and features for supplemental oxygen support and neurological status (but no variability features). We call the full model that includes variability features "M1" and the model without variability features "M2".

Table 3.4: Oxford Early Warning Score (Oxford-EWS)

	Score						
Vital sign	3	2	1	0	1	2	3
Heart rate (beats/minute)	≤ 42	43 - 49	50 - 53	54 - 104	105 - 112	113 - 127	≥ 128
Respiratory rate (breaths/minute)	≤ 7	8 - 10	11 - 13	14 - 25	26 - 28	29 - 33	≥ 34
SpO_2 (%)	≤ 84	85 - 90	91 - 93	≥ 94			
Temperature (°C)	≤ 35.4		35.5 - 35.9	36.0 - 37.3	37.4 - 38.3		≥ 38.4
Systolic blood pressure (mmHg)	≤ 85	86 - 96	97 - 101	102 - 154	155 - 164	165 - 184	≥ 185
Neurological status (AVPU)				A	V		P, U

Table 3.5: National Early Warning Score (NEWS) (McGinley and Pearse [2012])

	Score						
Vital sign	3	2	1	0	1	2	3
Heart rate (beats/minute)	≤ 40		41 - 50	51 - 90	91 - 110	111 - 130	≥ 131
Respiratory rate (breaths/minute)	≤ 8		9 - 11	12 - 20		21 - 24	≥ 25
SpO_2 (%)	≤ 91	92 - 93	94 - 95	≥ 96			
Temperature (°C)	≤ 35.0		35.1 - 36.0	36.1 - 38.0	38.1 - 39.0	≥ 39.1	
Systolic blood pressure (mmHg)	≤ 90	91 - 100	101 - 110	111 - 219			≥ 220
Neurological status (AVPU)				A			V, P, U
Received supplemental oxygen		Yes		No			

3.3.3 Model evaluation

Benchmark EWS systems

We compared the performance of models "M1" and "M2" to several EWS systems that have been described in the literature. These EWSs are multi-parameter scoring systems that apply weights to the vital signs based on discrete vital-sign ranges, and summing the weights to form an overall score. Multi-parameter EWS systems designed in this way are common, as described in section 1.3.1.

The benchmark EWS systems used in this analysis are the National Early Warning Score (NEWS) (McGinley and Pearse [2012]), the VitalPac Early Warning Score (ViEWS) (Prytherch et al. [2010]), the Decision-tree Early Warning Score (DTEWS) (Badriyah et al. [2014]) and the EWS used in the Oxford University NHS Hospitals Trust, which we denote as the Oxford-EWS. These scores use the same vital signs, but apply different weights in different ranges. For example, tables 3.4 and 3.5 show the EWS weights for the Oxford-EWS and NEWS scores, respectively.

We computed the EWS scores described above for the CALMS-2 dataset. Missing values for all vital signs were assigned scores of 0. We note that some EWSs have multiple thresholds that correspond to different levels of care or escalation, which can be triggered by individual vital signs or by the total score. For example, NEWS has low, moderate and high escalation criteria, with the moderate criterion triggered by a single vital-sign score of 3 or a total EWS score of 5. These criteria do not easily allow analyses based on a single threshold. For analyses based on varying threshold levels (such as AUROC) we applied the thresholds to the overall score. For analyses based on the assigned EWS decision, we used the reported decision thresholds for the moderate level of escalation. The performance metrics are described below.

Performance metrics

We measured the performance of the machine learning and EWS models on the CALMS-2 dataset, using a number of metrics that are commonly used to measure the performance of EWSs, and binary classifiers in general. We computed the AUROC, which was described in section 3.2.3. A limitation of the AUROC as an EWS performance metric is that it assigns weight to parts of the ROC curve that would be inappropriate in a clinical scenario, for example very high FPR, which would be too burdensome on clinical workload. Instead, the partial AUROC (pAUROC) can be computed (McClish [1989]), which is the area under the part of the ROC curve that

relates to clinically relevant values of TPR and FPR. The pAUROC is defined over the range of FPRs $[0, e]$ to be the definite integral of the ROC function over the given range: $A_e = \int_0^e ROC(f)df$. When $e = 1$ the pAUROC is equivalent to the AUROC. To standardize the range of pAUROC values to be between 0 and 1, we scale the pAUROC using the methodology of Ma et al. [2013], as follows:

$$\hat{A}_e = \frac{1}{2}\left(1 + \frac{A_e - e^2/2}{e - e^2/2}\right) \tag{3.9}$$

where $\hat{A}_e$ is the scaled pAUROC for FPRs in the range $[0, e]$, based on the unscaled definite integral A_e.

The scaled pAUROC calculated in this way (henceforth denoted as pAUROC – without reference to scaling) has a range of 0 to 1, with 0.5 equivalent to random guessing. We use $e = 0.4$, thus evaluating the pAUROC for FPRs between 0 and 0.4, which is equivalent to specificity of 0.6 to 1.0 (as FPR is 1 - specificity). We chose this value because a specificity below 0.6 would be too burdensome on workload.

While most published studies have used ROC analysis to evaluate EWS performance (Smith et al. [2008a,b]), there is evidence that this can lead to overestimating the benefit of an EWS, or underestimating the cost in terms of clinical resources (Romero-Brufau et al. [2015]). The AUROC can describe how models rank order positive and negative outcomes, but is not a function of the actual predicted probabilities (Cook [2007]). For diagnostic tests with low positive outcome prevalence such as an EWS (Escobar et al. [2012]), the positive predictive value (PPV) may provide more information than specificity for determining the effect of false positives (Romero-Brufau et al. [2015]). PPV is the proportion of true positives that lead to an adverse event, as follows.

$$PPV = \frac{TP}{TP + FP} \tag{3.10}$$

where TP and FP are the number of true and false positive predictions.

Similarly to ROC analysis, it is common to assess the PPV for a number of thresholds and to plot the values against the sensitivity at those thresholds. This is a precision-recall plot; precision and recall are alternative terms for PPV and sensitivity, respectively. We calculated the area under the precision recall curve (AUCPR), which is a measure of a classifier's precision-recall performance for all decision thresholds.

Metrics such as AUROC and AUCPR are useful for assessing model performance independently from the chosen decision threshold. However, these metrics do not answer clinically important questions such as how many patients will be correctly

detected or how many false alarms will have to be tolerated. These questions require the model decisions to be known, and therefore require an explicit decision threshold. Choosing a decision threshold allows the evaluation of metrics such as sensitivity, specificity and PPV for the predicted decisions from the model, from which comparisons can be made with the decisions of the benchmark EWSs. Ideally, the decision threshold will be based on clinical requirements, with the aim of achieving the highest possible sensitivity while maintaining a reasonable specificity such that clinical staff are not overly burdened by false alarms. However, there is no consensus on recommended minimum sensitivities and specificities. Thus, we choose an "optimal" decision threshold by minimizing the loss function g as follows.

$$g(o) = v(1 - Sens(o)) + (1 - v)Spec(o) \tag{3.11}$$

where $Sens$ is the sensitivity and $Spec$ is the specificity, at the threshold value o. v corresponds to the cost of a false negative (Pimentel [2015]).

If the relative importance of false positives and false negatives is equal, then $v = 0.5$. Choosing the threshold at $v = 0.5$ is equivalent to maximising *Youden's J statistic* ($J = sensitivity + specificity - 1$), which is a common heuristic for choosing the operating point of a diagnostic test (Youden [1950]). Geometrically, this is the closest point on the ROC curve to the point $[0, 1]$. In this chapter we use the threshold that minimises equation (3.11) with $v = 0.5$, assuming equal importance of false positives and false negatives. We use this threshold to compute predicted outcomes, from which we calculate sensitivity, specificity and PPV. We also computed the sensitivity of "M1" and "M2" at the threshold that achieved the same specificity as NEWS and likewise computed the specificity at the threshold that achieved the same sensitivity as NEWS.

Finally, we conducted a time to event (TTE) analysis to discover the amount of early warning time given by the models before an adverse clinical event. For each adverse clinical event, the time to event (TTE) was the time of the first positive alert before the event that occurred within 24 hours of the event. If no positive alert occurred within the 24-hour window, the TTE was zero. The overall TTE was the median TTE of all the adverse events. We computed the TTE for each model, using model decisions based on the optimal threshold (see equation (3.11)) for "M1" and "M2", and using the clinical escalation thresholds for the benchmark EWSs. For "M1" and "M2" we also computed the TTE using thresholds that achieved the same sensitivity as NEWS, and the same specificity as NEWS. We created TTE curves for each model by plotting the TTE against the FPR using different decision thresholds.

In summary, we used several metrics to measure the performance of models "M1",

"M2" and the benchmark EWSs on the CALMS-2 dataset. We plotted ROC and precision-recall curves, and computed the AUROC, the pAUROC and the AUCPR. We computed the sensitivity, specificity and PPV to compare the model decisions, using derived decision thresholds for "M1" and "M2", and clinical escalation thresholds for the benchmark EWSs. We also conducted a TTE analysis.

Statistical analysis

We report upper and lower quartiles around the median for the TTE estimates. We computed 95% confidence intervals for the AUROC, pAUROC, AUCPR, sensitivity, specificity and PPV using the logit interval (Boyd et al. [2013]). The logit interval is based on the binomial interval method, which interprets the test metric θ as a probability associated with some binomial(1,θ) variable, thus allowing the standard normal approximation of confidence intervals. This is appropriate as the listed test metrics are proportions. The binomial confidence interval can be calculated as follows:

$$CI_{binomial} = \hat{\theta} \pm 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \tag{3.12}$$

where $CI_{binomial}$ are the 95% confidence intervals calculated by the binomial interval method, for test metric estimate $\hat{\theta}$. n is the number of samples. 1.96 is the value of the 97.5 percentile of the Student's t-distribution (valid for large n) (Boyd et al. [2013]).

A disadvantage of the binomial method is that it may produce confidence intervals outside of [0, 1]. The logit interval corrects for this disadvantage by applying a logit transformation to the test metric, thus guaranteeing confidence intervals in [0, 1]. The logit interval is calculated as follows.

$$CI_{logit} = \left[\frac{e^{\hat{\eta}-1.96\hat{\tau}}}{1+e^{\hat{\eta}-1.96\hat{\tau}}}, \frac{e^{\hat{\eta}+1.96\hat{\tau}}}{1+e^{\hat{\eta}+1.96\hat{\tau}}}\right] \tag{3.13}$$

where CI_{logit} are the 95% confidence intervals calculated by the logit interval method, $\hat{\eta} = log\left(\frac{\hat{\theta}}{1-\hat{\theta}}\right)$ and $\hat{\tau} = (n\hat{\theta}(1-\hat{\theta}))^{-1/2}$.

3.3.4 Results

Feature selection

Table 3.6 shows the results for the cross-validation analysis of the optimal combination of features for the model "M1", which contained vital-sign and vital-sign variability

features, as well as features for supplemental oxygen support and neurological status. The best performing model included all vital-sign and vital-sign variability features, with a mean AUROC (standard deviation) of 0.865 (0.023) and a conservative mean AUROC of 0.843. This model was used as "M1" for the remainder of this chapter.

Table 3.6: AUROC results for the 5-fold cross-validation feature selection analysis on the CALMS-3 dataset. The features are listed in the order they were added to the model, with the order based on the univariate analysis in table 3.3. All models also included features for supplemental oxygen support use and neurological status. For example the respiratory rate variability results were for a model that included a respiratory rate variability feature, while the temperature variability results were for a model that included all vital-sign variability features. The model was logistic regression with restricted cubic spline interpolation. The number of knots for each feature were shown in table 3.3.

Added feature	**AUROC**	
	Mean (Standard deviation)	**Conservative mean**
Respiratory rate variability	0.780 (0.026)	0.754
SpO_2 variability	0.812 (0.025)	0.789
Respiratory rate	0.837 (0.023)	0.814
Systolic blood pressure variability	0.840 (0.020)	0.819
Heart rate	0.858 (0.027)	0.831
Heart rate variability	0.858 (0.027)	0.831
Temperature	0.859 (0.025)	0.835
SpO_2	0.859 (0.025)	0.835
Systolic blood pressure	0.866 (0.024)	0.842
Temperature variability	0.865 (0.023)	0.843

Performance analysis

Table 3.7 shows the results for the performance analysis of "M1" and "M2" (both trained on the CALMS-3 dataset), NEWS, VitalPac Early Warning Score (ViEWS), DTEWS and the Oxford-EWS, on the CALMS-2 dataset. Figure 3.5 shows the ROC curves, figure 3.6 shows the precision-recall curves, and figure 3.7 shows the TTE curves, for "M1", "M2", and the benchmark EWSs. Table 3.8 shows the comparison of the sensitivity, specificity and TTE using thresholds that achieve the same sensitivity and specificity as NEWS.

Table 3.7: Results for the performance analysis of the machine learning models and the benchmark EWSs on the CALMS-2 dataset. The decision based metrics were computed based on the optimal decision threshold calculated in equation (3.11) for models "M1" and "M2" and with clinical decision thresholds for the benchmark EWSs. Here Oxford refers to the Oxford-EWS (see table 3.4), CI refers to 95% confidence intervals and IQR refers to the inter-quartile range (see section 3.3.3).

	Curve-based metrics			**Decision-based metrics**			
Model	**AUROC (95% CI)**	**pAUROC (95% CI)**	**AUCPR (95% CI)**	**Sensitivity (95% CI)**	**Specificity (95% CI)**	**PPV (95% CI)**	**TTE (hours) (IQR)**
M1	0.886 (0.883, 0.890)	0.843 (0.839, 0.847)	0.285 (0.280, 0.290)	0.778 (0.774, 0.782)	0.839 (0.835, 0.843)	0.116 (0.113, 0.120)	19.75 (10.04, 22.92)
M2	0.871 (0.867, 0.874)	0.822 (0.818, 0.826)	0.256 (0.252, 0.261)	0.794 (0.790, 0.798)	0.779 (0.775, 0.784)	0.089 (0.086, 0.092)	18.92 (10.04, 22.75)
Oxford	0.819 (0.814, 0.823)	0.781 (0.777, 0.785)	0.188 (0.184, 0.192)	0.570 (0.564, 0.575)	0.896 (0.893, 0.899)	0.130 (0.126, 0.134)	15.75 (7.23, 21.00)
NEWS	0.848 (0.844, 0.852)	0.795 (0.791, 0.799)	0.225 (0.221, 0.230)	0.704 (0.699, 0.709)	0.823 (0.819, 0.827)	0.098 (0.094, 0.101)	18.68 (13.76, 21.33)
ViEWS	0.851 (0.847, 0.855)	0.797 (0.792, 0.801)	0.223 (0.219, 0.228)	0.681 (0.676, 0.686)	0.847 (0.843, 0.851)	0.108 (0.105, 0.111)	18.68 (13.13, 21.33)
DTEWS	0.832 (0.828, 0.836)	0.776 (0.772, 0.781)	0.195 (0.191, 0.200)	0.872 (0.868, 0.875)	0.595 (0.589, 0.600)	0.055 (0.053, 0.058)	20.83 (17.04, 23.08)

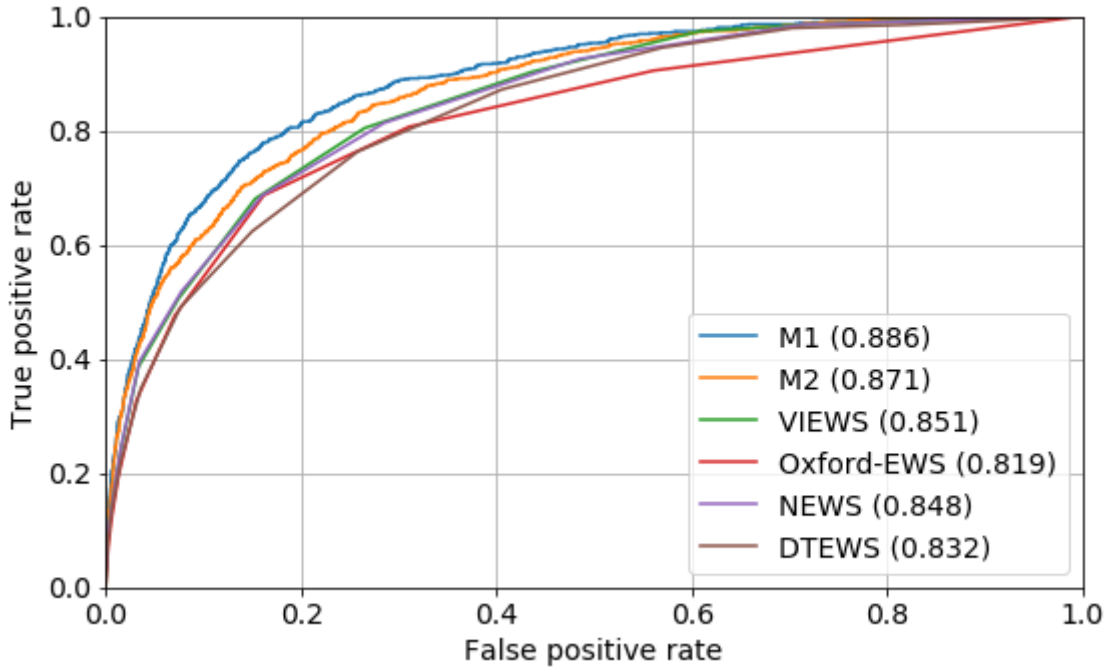


Figure 3.5: ROC curves plotting true positive rate against false positive rate for the models tested on the CALMS-2 dataset. The AUROC is shown for each model.

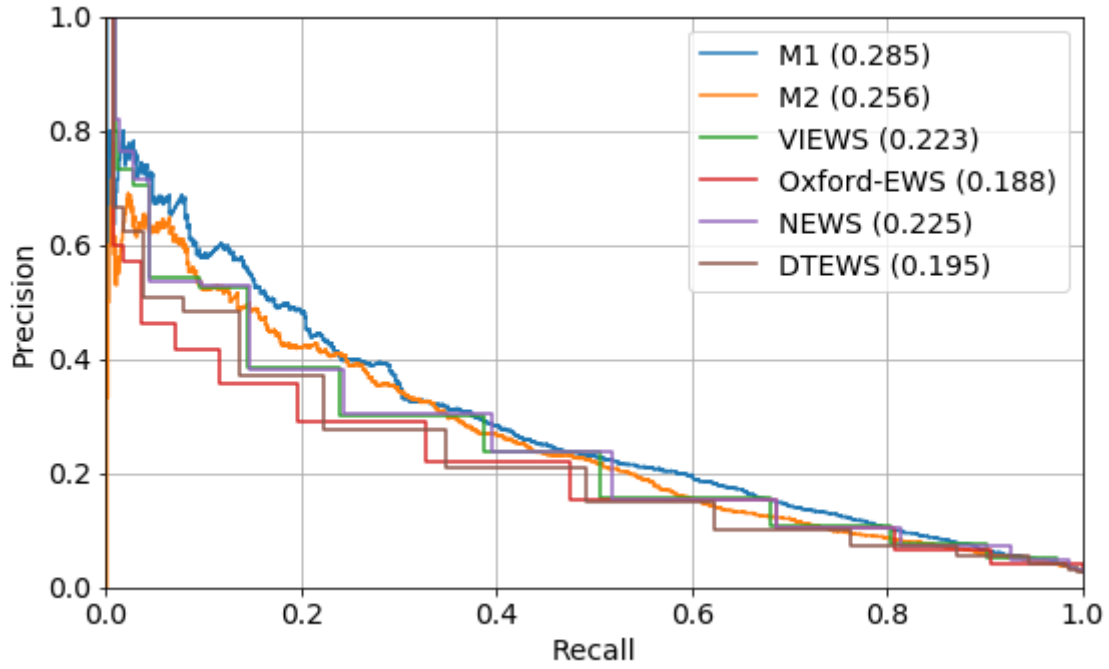


Figure 3.6: Precision-recall curves plotting precision (PPV) against recall (sensitivity) for the models tested on the CALMS-2 dataset. The AUCPR is shown for each model.

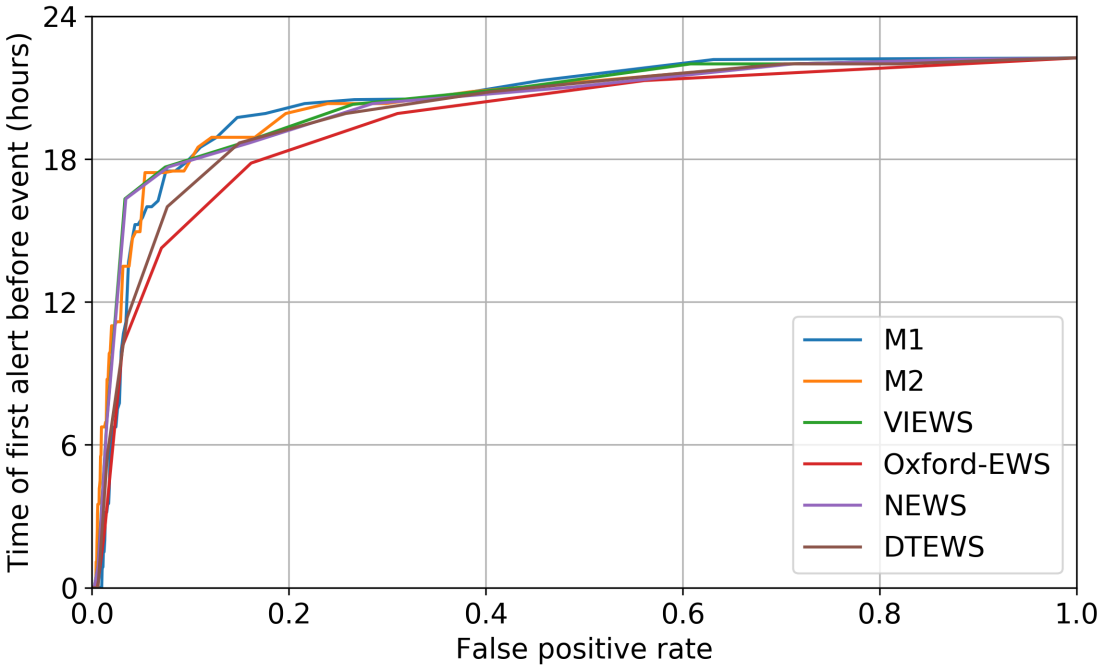


Figure 3.7: Time-to-event curves plotting the median time of the first alert (positive model prediction) within 24-hours of an adverse event against the false positive rate for the models tested on the CALMS-2 dataset.

Table 3.8: Analysis of the sensitivity, specificity and TTE of "M1", "M2" and NEWS using decision thresholds that achieve the sensitivity, and the specificity, of NEWS. CI refers to 95% confidence intervals and IQR refers to inter-quartile range.

Threshold type	Model	Sensitivity (CI)	Specificity (CI)	TTE (IQR)
Sensitivity	NEWS	0.704 (0.699, 0.709)	0.823 (0.819, 0.827)	18.66 (13.76, 21.33)
	M1	0.703 (0.698, 0.707)	0.884 (0.881, 0.888)	18.92 (7.51, 22.53)
	M2	0.703 (0.698, 0.707)	0.859 (0.855, 0.862)	18.92 (9.23, 22.71)
Specificity	NEWS	0.704 (0.699, 0.709)	0.823 (0.819, 0.827)	18.66 (13.76, 21.33)
	M1	0.789 (0.785, 0.794)	0.824 (0.820, 0.828)	19.92 (10.04, 22.92)
	M2	0.740 (0.736, 0.745)	0.823 (0.818, 0.827)	18.92 (10.04 - 22.75)

3.3.5 Discussion

The machine learning based EWS model with vital-sign variability features ("M1") performed better than all other models for predicting adverse clinical events in the CALMS-2 dataset. Table 3.7 and figures 3.5 and 3.6 show that "M1" was the best performing model in terms of AUROC, pAUROC and AUCPR. We note that an improvement in AUROC of 0.848 (NEWS) to 0.886 ("M1") is an important difference, and that "M1" outperforms other EWS models presented in the scientific literature (Pimentel [2015]) to the best of our knowledge. The machine learning model without vital-sign variability features ("M2") was the next best performing model, showing that the use of machine learning to train a model was advantageous over heuristic methods used to design the benchmark EWSs. Of the benchmark EWSs, NEWS and ViEWS had similar performance (with overlapping 95% confidence intervals) and were the highest scoring in terms of AUROC, pAUROC and AUCPR.

The AUROC and pAUROC values of "M1" were higher than those metrics for NEWS with non-overlapping confidence intervals. It is therefore reasonable to conclude with confidence that "M1" performed better than the benchmark EWSs on the CALMS-2 dataset. Similarly, "M1" outperformed "M2" in terms of AUROC and pAUROC with non-overlapping confidence intervals. Thus, the addition of vital-sign variability features to the model caused an increase in discriminative performance.

The univariate feature analysis in section 3.2 provided a ranking of the vital-sign and vital-sign variability features for predicting adverse clinical events. The feature selection process, with results given in table 3.6, showed that while some features had a poor performance in the univariate setting, the best performing version of "M1" included all possible vital-sign and vital-sign variability features. Interestingly, some of the lesser performing features in the univariate setting provided some of the largest increases in conservative mean AUROC. For example, systolic blood pressure was the second worst performing feature in the univariate analysis, but provided an increase in conservative mean AUROC of 0.007 when added to model "M1" (one of the largest increases). Likewise, heart rate was only the 5th most informative feature in the univariate setting but provided a large increase in model performance when added to "M1". Not surprisingly, the univariate analysis likely had limitations with respect to ranking features as the input of a multivariate model. All features provided independent information for predicting adverse clinical events, and were included in the final version of "M1".

The results of this analysis confirm those presented in the univariate analysis in

section 3.2, showing that features of diurnal vital-sign variability are predictive of adverse clinical events. However, the increase in AUROC of 2% between "M2" and "M1" (87% to 89%) is surprisingly modest given the high importance found for features of vital-sign variability in the univariate analysis. This may be because vital-sign variability features have a marginal benefit when the vital signs are already in the model, or it may be because there are "diminishing returns" as the model performance is already at a high level given the data available. However, vital-sign variability features are still worth including in a machine-learning model as there is negligible additional computational burden, and even a small increase in predictive performance could yield clinical benefits for some patients.

While the machine-learning models performed best in terms of curve-based metrics such as AUROC, the results in table 3.7 suggested that other models performed better in terms of sensitivity, specificity and PPV. For example, DTEWS had the highest sensitivity (0.872) and the Oxford-EWS had the highest specificity (0.896) and PPV (0.130). This means that DTEWS would have detected a higher proportion of adverse clinical events, and the Oxford-EWS would have had a lower proportion of false alarms. However, the EWSs achieved these high metrics by applying thresholds that traded-off a poor performance in specificity (for DTEWS) and sensitivity (for the Oxford-EWS). Either the models were tuned to achieve these results (and tolerate these trade-offs) for particular clinical purposes, or the decision threshold was poorly chosen. The imbalance in sensitivity and specificity in these models may also be a function of the discretised design of EWSs, with integer decision thresholds (for ease of calculation) offering few possible choices of the optimal value. Conversely, the predicted likelihoods from "M1" and "M2" allow any value in the range [0, 1] to be chosen as the decision threshold.

Table 3.8 showed that "M1" can be tuned to match the sensitivity of NEWS while achieving higher values of specificity. Likewise, "M1" can be tuned to match the specificity of NEWS while achieving a higher sensitivity. We conclude that, on balance, "M1" performed the best in terms of sensitivity, specificity and PPV, with some of the highest values for all metrics. Furthermore, "M1" can be tuned to meet the particular clinical needs of a ward, for example by achieving high sensitivity while maintaining an acceptable specificity.

Badriyah et al. [2014] claimed that an EWS with reasonable discrimination is indicated by AUROC values of 0.7 to 0.8 and good discrimination by values exceeding 0.800. On this basis, the models presented in this chapter perform well, with AUROC values of up to 0.886. However, the PPV values of the machine learning and benchmark

EWS models presented in table 3.7 are low, mostly around 0.08 to 0.12. This means that approximately 10% of alerts generated by the models lead to an adverse clinical event. This result is partly a function of the low prevalence of adverse clinical events in the dataset (prevalence affects PPV), which is exacerbated by the method of assessing the model performance by observation set, as even "abnormal" patients contribute more negative observations than positive observations. The low PPV values may also be partly caused by "abnormal" observations that led to successful treatment, and therefore did not lead to an adverse event. Such observations had negative labels as they were not within 24 hours of an adverse event, but may have been truly abnormal (and therefore required alerting) before the intervention, thus potentially affecting the value of PPV. Details of interventions were not available for the CALMS-2 and CALMS-3 datasets, so it was not possible to investigate this effect. Regardless, "M1" had a higher PPV than most models that are currently in clinical use (except the Oxford-EWS which sacrifices a high sensitivity to achieve this) and thus has the potential to reduce this false alert burden.

The TTE analysis showed that DTEWS alerted the earliest of all the models with a median TTE of 20.83 hours (see table 3.7). However, this was likely to be related to the high sensitivity and low specificity trade-off that DTEWS makes with the decision threshold (as discussed earlier). Of the other models, "M1" alerts earliest (19.75 hours) followed by "M2" (18.92). "M1" alerts over 1 hour earlier than benchmark EWSs such as NEWS and ViEWS, which both have TTEs of 18.68 hours. Figure 3.7 shows the median TTE against the FPR for different decision thresholds, and demonstrates that "M1" and "M2" can be tuned to have higher or lower FPR (to achieve a high sensitivity) and will still alert relatively early, as evidenced by the high TTE curve for these models. Early alerting is important as each hour of delayed admission to the ICU is estimated to increase the risk of death by 1.5% (Cardoso et al. [2011]). Thus, models "M1" and "M2" may have the potential to improve patient care by alerting clinical staff to patients who need intervention in a timely manner.

The strong performance of model "M2" compared to benchmark EWSs with similar features shows that machine learning partly accounts for the improvement in model performance seen in both "M1" and "M2" (with the addition of variability features accounting for the increase in performance of "M1" with respect to "M2"). Interestingly, DTEWS was designed using machine learning – Badriyah et al. [2014] used decision tree models to learn optimal decision thresholds for the EWS. The discrepancy in the performance between DTEWS and "M2" may be caused by the discretisation of risk intervals in the design of DTEWS, which has a similar design to

the Oxford-EWS and NEWS shown in tables 3.4 and 3.5, with ranges of vital-sign values with different weights. Conversely, the logistic regression model with cubic spline interpolation allows the likelihood of an adverse event to change continuously with the vital-sign value, which can be visualised with the risk profiles in figure 3.4. It is only possible to learn such a continuous scale (rather than it being designed heuristically) and the score cannot be computed manually, so there is a clear difference between machine learning EWS models and traditional EWSs designed for manual calculation.

A strength of adding diurnal vital-sign variability features to an EWS model is that it allows more complex EWSs to be computed, while retaining the benefit of clinical interpretability. Complex methods such as deep learning[1] have vastly improved the performance of machine learning models in fields such as computer vision, but they are "black-box" models that do not easily allow interpretation of the origin of a decision. This makes it difficult to translate such complex methods to the field of EWSs, as model decisions would not be explainable. Vital-sign variability features provide additional complexity to EWS models by encoding properties such as the abnormality of a patient's diurnal cycle, trends in a patient's trajectory of deterioration, and changes from patient-specific baseline levels (see section 3.2.5). These properties can be interpreted clinically and are thus appropriate for an EWS.

A limitation of the calculation of diurnal vital-sign variability from manual vital-sign observations is that physiological changes are likely to be under-sampled, as they may occur between the times of clinical visits. The performance of vital-sign variability features may be affected by this limitation, as the real peaks and troughs in the vital-sign value during the diurnal cycle may not have been sampled, thus estimating an incorrectly low level of variability. This limitation could be addressed with continuous vital-sign monitoring, as physiological changes could be tracked throughout the diurnal cycle. Indeed, continuous vital-sign monitoring may also improve the timing of alerts (TTE) by allowing persistent physiological changes to be detected early in their onset. Further experiments should assess the value of vital-sign variability features computed from continuous vital-sign measurements for improving EWS models.

[1]Deep learning is the use of neural network models with many hidden layers (parameter numbers can exceed 1 million), allowing the model to learn complex features that are highly predictive for the pattern recognition task.

3.4 Conclusions

In this chapter, we have demonstrated that features of diurnal vital-sign variability are predictive of adverse clinical events in post-operative care. We conducted a univariate analysis of vital-sign and vital-sign variability features on the CALMS-3 dataset using a logistic regression model with cubic spline interpolation. We found that in many cases, features of diurnal vital-sign variability had more predictive value than features of the corresponding vital signs. Respiratory rate variability and SpO_2 variability were the most important features.

We then trained a machine learning model (logistic regression with cubic spline interpolation) to detect adverse clinical events in the CALMS-3 dataset, which we then tested on the CALMS-2 dataset. The machine learning model with vital-sign variability features outperformed the model without vital-sign variability features, as well as several benchmark EWSs used in clinical care. We conclude that machine learning can be used to train EWS models with better discriminative performance than traditional EWSs designed for manual calculation. We also conclude that the addition of features of diurnal vital-sign variability to an EWS can improve identification of the onset of physiological deterioration.

Further improvements in predictive performance could be made to EWS systems for post-operative care. Continuous vital-sign monitoring could identify trends in physiological values early in their onset, allowing a correctly-sampled estimate of diurnal vital-sign variability, and perhaps more timely detection of physiological deterioration. Further experiments should be conducted to assess the use of EWSs with continuous vital-sign monitoring, and also the computation of diurnal vital-sign variability features from continuous vital-sign measurements.

Chapter 4

Method comparison of continuous and manual vital-sign data

The previous chapter demonstrated how EWSs can be designed to predict future adverse clinical events within the next 24 hours with high sensitivity and specificity, particularly with the inclusion of features of vital-sign variability. As discussed in section 1.4, the detection of physiological deterioration may be further improved with continuous vital-sign monitoring, allowing changes in vital-sign values to be tracked in real-time. This could improve the computation of vital-sign variability, as important vital-sign changes between clinical observations would be included in the variability calculation. The integration of EWSs with vital-sign variability features and continuous monitoring sources could lead to earlier detection of clinically important signs of physiological deterioration.

Translating EWS scores designed for computation with manually-recorded vital-sign observations to continuous monitoring sources may not be straightforward. Previous studies have reported that vital-sign observations collected manually may be affected by the vital-sign taker due to phenomena such as clinical inaccuracies (Clifton et al. [2015]), clinical bias towards normal physiology when recording the vital signs (Taenzer et al. [2014]), or arousal of the patient's physiology caused by the presence of clinical staff (Martin and McGrath [2014]). It is important to understand whether these phenomena exist in a post-operative setting before translating EWS values derived for manual vital-sign observations to continuous monitoring sources.

In this chapter, we conduct a method comparison study between manual and continuous vital-sign values from bedside monitors. We focus on two potential sources of inaccuracies that have been described in the literature. Primarily, we assess whether manual vital-sign observations are biased towards normal physiology, which Block [1991] denoted the "smoothing effect" as this phenomenon smooths the vital-sign record of extremal values. If present, this would affect the vital-sign variability signal found in the previous chapter, as vital-sign variability would have been smoothed in

the manual record. We also assess whether there is an "arousal effect" caused by the presence of the vital-sign taker (Taenzer et al. [2014]), which would affect the manual vital-sign record as values would reflect slightly different physiology than what would be measured by an automated continuous system without the input of clinical staff.

4.1 Previous work

4.1.1 Clinical errors

Several previous studies have shown manual measurements of the main vital signs to occasionally be inaccurate (Hooker et al. [1989]; Van Leuvan and Mitchell [2008]). In particular, respiratory rate is known to be difficult to measure accurately (Lovett et al. [2005]). Errors in the calculation of clinical risk scores such as EWSs have also been described (Prytherch et al. [2006]; Subbe et al. [2007]), although Clifton et al. [2015] found that the values documented by clinicians often preceded similar "true" values, thus suggesting that such errors may actually encode subjective clinical judgements of future physiological status.

The "smoothing effect" was first reported to occur during anaesthesia (Block [1991]; Cook et al. [1989]; Lerou et al. [1988]; Reich et al. [2000]; Thrush [1992]; Zollinger et al. [1977]). These studies were mostly conducted in the transition from handwritten patient records to automated anaesthetic records, with the elimination of the "smoothing effect" cited as a potential benefit of the adoption of automated record systems. More recently, vital-sign smoothing has been reported in acute ward monitoring (Sapo et al. [2009]; Taenzer et al. [2014]). Similarly, these studies cited the "smoothing effect" as a potential reason to adopt continuous measurement technology for surveillance monitoring of patients.

The influence of the measurement environment on the value of a patient's vital signs has been studied in the context of blood pressure measurement in general practice. Known as "white-coat hypertension", patients can exhibit higher blood pressure values in the clinic than they would otherwise, which may be due to the anxiety of being monitored (Martin and McGrath [2014]). Similarly, it is possible that the presence of clinical staff taking manual observations in the ward environment may have an "arousal effect" or indeed a "wake-up effect" (due to the interruption of sleep), although Taenzer et al. [2014] found no evidence of this occurring in their dataset. Limited evidence is available on the "arousal effect" in hospital wards so further studies are needed.

4.1.2 Methodological issues

There may be concerns with the methodologies that have been used to study the "smoothing effect". For example, four previous studies compared the magnitude and frequency of extreme values in manual and automated vital-sign measurement where the automated measurement had much higher measurement rates (Cook et al. [1989]; Devitt et al. [1999]; Lerou et al. [1988]; Thrush [1992]). Signals sampled frequently are more likely to capture transient extreme measurements than sparsely-sampled signals so the manual observations may not have coincided with times of large fluctuations in vital-sign values. Thus, the "smoothing effect" found in these studies may perhaps be explained by under-sampling. Reich et al. [2000] assessed the "smoothing effect" by comparing manual and automated measurements from different patient cohorts in a matched sample design study. While more extremal values were found in the automated measurement data, it is difficult to tell whether this discrepancy was due to differences in vital-sign values between cohorts or differences in measurement methods. Other studies have only presented data from automated measurements (Block [1991]), and used simulated vital-sign measurements from mannequins (Devitt et al. [1999]).

A number of studies compared manual vital-sign observations with continuous monitoring data without using any form of temporal averaging to remove potential signal artefact (Block [1991]; Cook et al. [1989]; Thrush [1992]; Zollinger et al. [1977]). Continuous vital-sign monitoring signals can be noisy and often fluctuate above and below the true physiological value. Thus, when sampled randomly, continuous monitoring data are likely to differ from the true physiological value. Furthermore, extremal vital-sign values are often associated with spurious values, as demonstrated by Sapo et al. [2009] for SpO_2 measurements. Clinicians may apply a legitimate form of data smoothing by sampling the value that best summarises the trend in the data. This is different from the "smoothing effect", which refers to clinicians systematically choosing values that are more normal than the patient's true physiology. Median filtering can be used to ensure that the continuous data samples represent the trends in the data rather than the variance (Tarassenko et al. [2005]). For example, Reich et al. [2000] and Taenzer et al. [2014] both sampled data using the median value of a 5-minute window in their analyses of the "smoothing effect".

A technical methodological error was made by Taenzer et al. [2014] in their analysis of the "smoothing effect". The authors reported a difference of 6.5% between continuous and manual measurements of SpO_2 below 90% in general medical and post-operative wards. The method used continuous data to group SpO_2 measurements

above or below 90% and then found the average difference between continuous and manual measurements within each group. By comparing the differences between two measurement sources to the values of the continuous source, a spurious correlation was introduced to the data (Bland and Altman [1995]) that confounds the results and may explain why a large difference was found between measurement methods. We further explain this spurious correlation in appendix A.1. We also note that the dataset in this paper was small, with only 16 measurements below 90%.

4.1.3 Summary

If real, the "smoothing effect" may result in lost opportunities for early recognition of physiological deterioration. The "smoothing effect" may also have implications for translating EWSs designed for manual vital-sign data to continuous monitoring sources if manual data are less likely to contain extremal values. However, to our knowledge all previous studies of data smoothing have been confounded by methodological errors that make it difficult to ascertain whether this source of bias indeed exists. It is important to understand whether the "smoothing effect" truly exists using a methodology that is robust to these confounding factors.

4.2 Data preparation

4.2.1 Overview

In this chapter we assess the "smoothing effect", agreement between measurement sources, and the "arousal effect" in vital-sign measurements of respiratory rate, heart rate and SpO_2. For this purpose, we conduct a secondary analysis of the CALMS-2 dataset. Our methodology is based on the comparison of paired manual and continuous vital-sign data. We describe the data preparation methodology below.

4.2.2 Data selection

The process for extracting the data from the CALMS-2 dataset for this analysis is shown in figure 4.1 For each vital sign, we selected the patients who had continuous measurements from the Philips bedside monitor. We did not include continuous monitoring values from wearable monitors in this analysis to ensure that the continuous vital-sign measurements were reliable and were unlikely to be affected by signal artefact. We then removed the data from patients who did not have any manual vital-sign

observations that occurred when they were receiving bedside monitoring, as they had no contemporaneous continuous and manual data that could be compared. We included all manual observations with contemporaneous periods of continuous data, allowing patients to provide multiple observations in our analysis. Paired vital-sign measurements were available for respiratory rate from 263 patients (3740 paired measurements), heart rate from 267 patients (3844 paired measurements), and SpO_2 from 271 patients (3896 paired measurements). The median (inter-quartile range (IQR)) numbers of observations for each patient were 11 (7 - 18) for respiratory rate, 11 (7 - 18) for heart rate and 11 (7 - 19) for SpO_2 measurements.

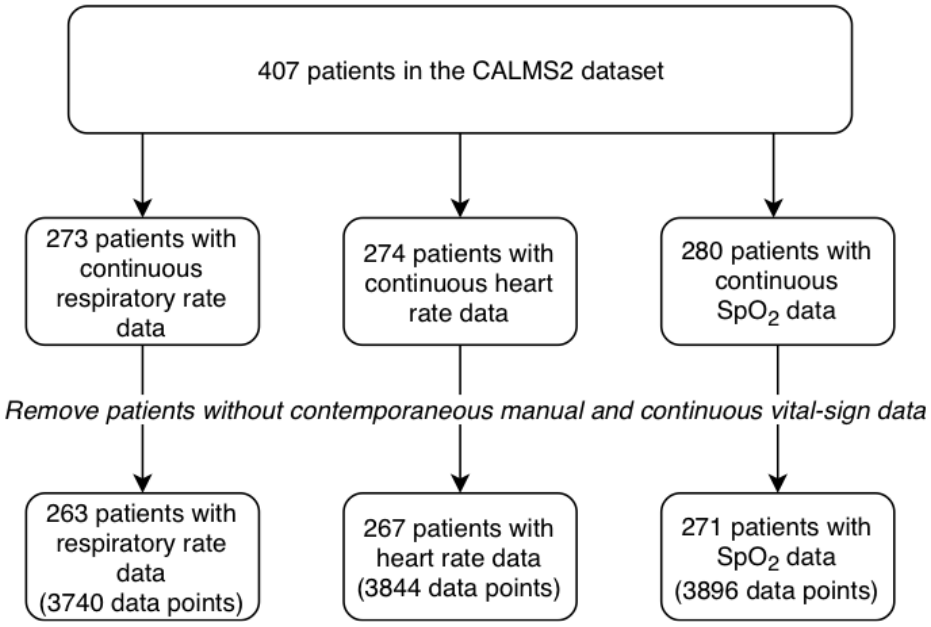


Figure 4.1: The selection process for extracting pairs of contemporaneous manual and continuous vital-sign measurements of respiratory rate, heart rate and oxygen saturation from the CALMS-2 dataset.

Manual and continuous vital-sign values were compared based on the timestamps taken from paper records and the computer-generated timestamps from the patient monitor. Clinical staff manually recorded blood pressure measurements at the time of observation, and these were also logged in the automatically-generated data record. We set the manual observation time for all vital signs to the computer-generated timestamps for blood pressure (while checking that data were correctly matched). This calibration method ensured that the data considered contemporaneous from manual and continuous measurement sources were synchronized, and thus could appropriately be used for comparison.

Data pre-processing

We sampled the continuous data at the time of each manual observation to create paired measurements of continuous and manual vital signs. This was done by extracting the median value of a 5-minute window of continuous data centred at the time of manual observation, which we call the "during" window. We used this methodology (median filtering) to summarise the continuous data without the effects of measurement noise or short-term variance, while retaining long-term trends. We selected a 5-minute window to reflect clinical practice, in line with previous work (Reich et al. [2000]; Taenzer et al. [2014]). We undertook sensitivity analyses by recomputing the "smoothing effect" analysis with "during" window sizes of 1 to 10 minutes (reported in section 4.3.2).

We also sampled the continuous data before the time of each manual observation, which we used for assessing the "arousal effect". We computed the median value of a 15-minute window ending 5 minutes before the time of observation, which we call the "before" window. This replicates the method of Taenzer et al. [2014] to allow comparison. The selection of the continuous data is shown schematically for heart rate in figure 4.2. Since the primary analysis involves the comparison of the continuous data from the "during" window to the manual observation, we refer to the value from the "during" window as the continuous vital-sign value unless stated otherwise.

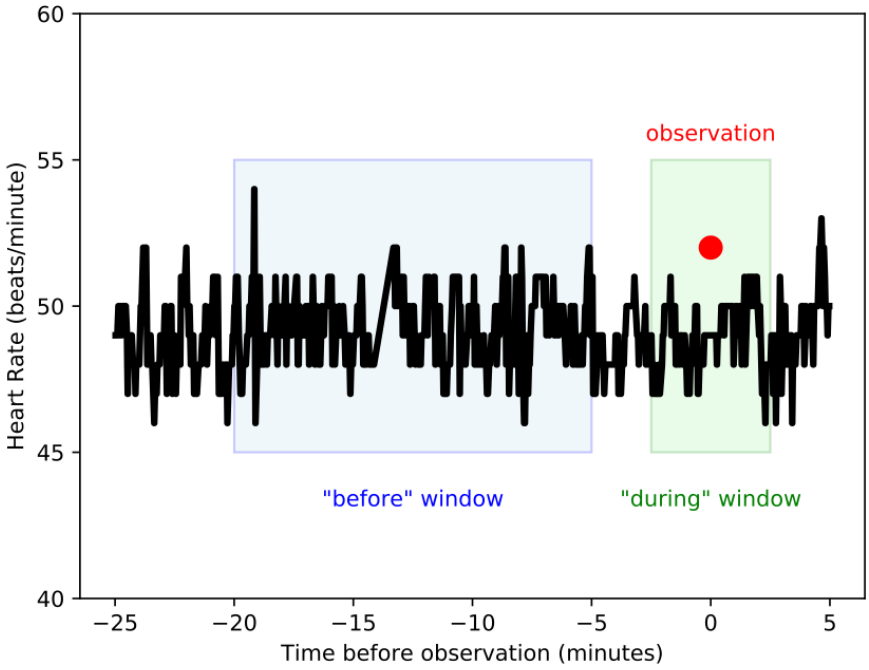


Figure 4.2: An example heart rate observation, for which the "before" window is shown in blue, the "during" window is shown in green, and the manual observation (synchronised to the time of the blood pressure measurement) is shown in red.

4.3 Assessing the "smoothing effect"

4.3.1 Methods

Our method of analysing the "smoothing effect" in our dataset is based on the canonical method of Bland and Altman [1986] for assessing agreement between two measurements sources, which is known as a Bland-Altman plot. In a Bland-Altman plot, the difference between measurement values recorded by the two sources are plotted against the average of the measurement values recorded by the two sources. The reason for using the average of the values from both measurement sources rather than values from one of the measurement sources is explained in appendix A.1. Bland-Altman plots show the distribution of the differences between the measurement sources, as well as the relationship between the differences and the measurement value. Bland-Altman plots typically show lines for the bias of the data, which is the mean of the differences, and the limits of agreement (LoA) of the data, which are limits bounding 95% of the differences. These are standard measures of agreement. We chose the Bland-Altman method because if vital-sign data were smoothed, manual measurements would be recorded above abnormally low continuous measurements and below abnormally high continuous measurements, with normal values unaffected.

We begin our analysis by creating a statistical model that relates the continuous-manual difference to the continuous-manual average. Our reason for developing this model is two-fold: First, the model allows us to quantify the relationship between the differences and the vital-sign values, which allows us to test for the "smoothing effect". Second, calculations for the limits of agreement (LoA) (in section 4.4) differ depending on whether the differences are constant, or whether they are related to the vital-sign value. Our model can therefore inform the correct LoA calculation method to choose.

To obtain the differences between the two measurement sources, we subtracted the manual vital-sign from the continuous vital-sign in each measurement pair. We also calculated the average of the continuous and manual vital-signs for each measurement pair. We used a linear mixed effects model with the continuous-manual differences as the dependent variable, the continuous-manual averages as a fixed effect and the patient as a random effect to adjust for repeated measurements. We fitted the model with the *lme4* package (Bates et al. [2015]) in R (R Core Team [2013]). This model is summarised in equation (4.1).

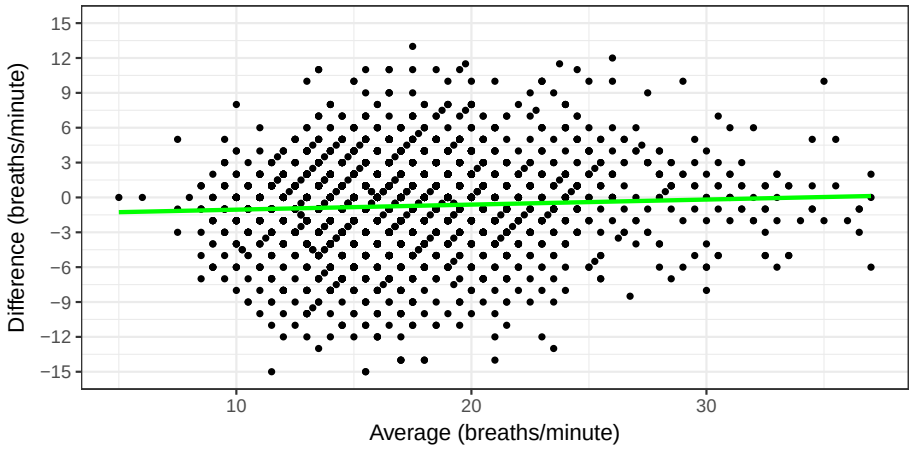
$$d_{ri} = \beta_0 + \beta_1 a_{ri} + u_{0i} + u_{1i} a_{ri} + e_{0ri} \tag{4.1}$$
$$e_{ri} \sim \mathcal{N}(0, \sigma^2_{e0}), \quad \begin{bmatrix} u_{0i} \\ u_{1i} \end{bmatrix} \sim \mathcal{N}(0, \Omega_u), \quad \Omega_u = \begin{bmatrix} \sigma^2_{u0} & \sigma_{u01} \\ \sigma_{u01} & \sigma^2_{u1} \end{bmatrix}$$

where d_{ri} is the continuous-manual difference for replicate measurement r from patient i, and a_{ri} is the continuous-manual average for replicate measurement r from patient i. β_0 and β_1 are the intercept and slope for the fixed effect. u_{0i} and u_{1i} are the intercept and slope for the random effect for patient i. e_{0ri} is the model residual with variance σ^2_{e0}. Ω_u is the covariance structure of the random effect with intercept and slope variance σ^2_{u0} and σ^2_{u1}, and covariance between the intercept and the slope residual for the random effect of σ_{u01}.

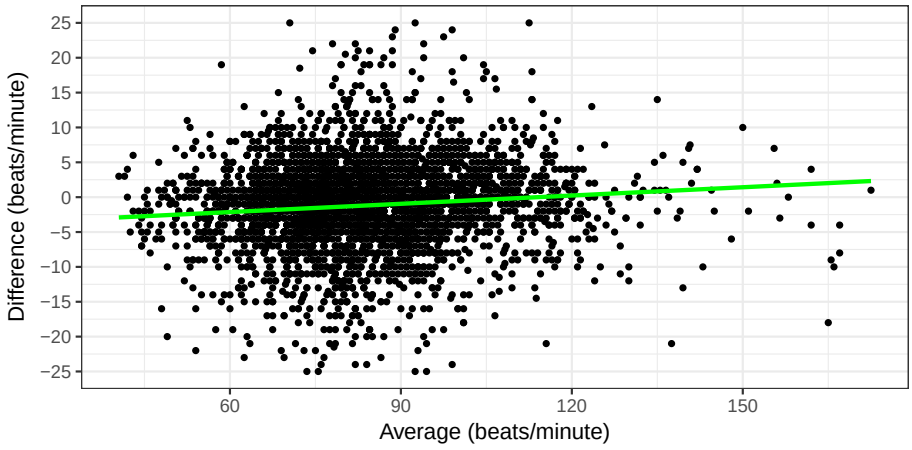
Statistical significance of the regression coefficient for the fixed effect (β_1) was calculated with an F-test (type III with the degrees of freedom approximation of Kenward and Roger [1997]), using a significance of 0.05 to reject the null hypothesis that the regression coefficient was 0 (no relationship). To assess the clinical significance of the linear relationship we simulated the differences predicted by our model at vital-sign values corresponding to important clinical thresholds. We did so by using the *ggeffects* package in R (Daniel Ldecke [2018]) to find the marginal mean continuous-average difference at a given continuous-manual average, with the patient random effect held constant. We used respiratory rate values of 8 and 24 breaths/minute, heart rate values of 40 and 130 beats/minute and an SpO_2 value of 91%, which are the maximum NEWS limits (McGinley and Pearse [2012]) for each vital sign (see section 1.3.1). As mentioned in section 4.2.2, we conducted sensitivity analysis by recomputing our results using "during" window sizes of 1 to 10 minutes.

4.3.2 Results

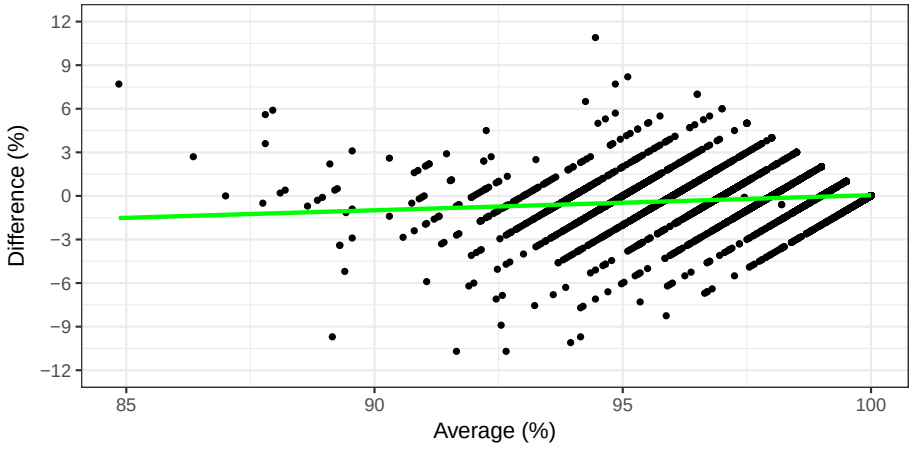
Figures 4.3a to 4.3c show scatter plots of the continuous-manual differences against the continuous-manual averages for respiratory rate, heart rate and SpO_2 (for a "during" window size of 5 minutes as described in section 4.2.2). The results for the analysis of the regression coefficient for the fixed effect in equation (4.1) are shown in table 4.1, as well as the simulated differences at vital-sign values corresponding to high NEWS values. We report results for "during" window sizes of 1 to 10 minutes.



(a) Respiratory rate



(b) Heart rate



(c) SpO_2

Figure 4.3: Scatter plots of the continuous-manual difference against the continuous-manual average of measurements of respiratory rate (a), heart rate (b) and SpO_2 (c). The data points for measurements from the CALMS-2 dataset are shown as black dots, and the predicted difference at continuous-manual average values simulated from the linear mixed effects model (see equation (4.1)) is shown in green. Difference is continuous - manual.

Table 4.1: Results for the "smoothing effect" analysis between continuous and manual measurements of respiratory rate, heart rate and SpO_2 for "during" window sizes of 1 to 10 minutes (with results for the chosen window size of 5 minutes in bold). The regression coefficient is the slope of the fixed effect in equation (4.1). C.I. is 95% confidence interval. The differences predicted at (continuous-manual average) vital-sign values corresponding to low (left column) and high (right column) NEWS thresholds are shown (8 and 24 breaths/minute for respiratory rate, 40 and 130 beats/minute for heart rate, and 91% for SpO_2).

			Predicted difference (C.I.)	
Window size (minutes)	**Regression coefficient (C.I.)**	**P-value**	**Lower limit**	**Upper limit**
Respiratory rate			**(breaths/minute)**	
1	0.17 (0.12, 0.21)	<0.001	-1.9 (-2.3, -1.6)	0.7 (0.3, 1.1)
2	0.11 (0.07, 0.16)	<0.001	-1.5 (-1.9, -1.2)	0.2 (-0.2, 0.6)
3	0.08 (0.03, 0.12)	0.001	-1.3 (-1.7, -0.9)	-0.1 (-0.5, 0.3)
4	0.05 (0.00, 0.09)	0.05	-1.1 (-1.5, -0.7)	-0.4 (-0.8, 0.0)
5	**0.04 (-0.01, 0.10)**	**0.1**	**-1.1 (-1.6, -0.7)**	**-0.4 (-0.9, 0.0)**
6	0.04 (-0.02, 0.09)	0.18	-1.1 (-1.5, -0.7)	-0.5 (-1.0, -0.1)
7	0.02 (-0.03, 0.07)	0.42	-1.0 (-1.5, -0.6)	-0.7 (-1.1, -0.2)
8	0.01 (-0.04, 0.06)	0.72	-0.9 (-1.4, -0.5)	-0.8 (-1.3, -0.3)
9	0.00 (-0.05, 0.05)	0.95	-0.9 (-1.3, -0.5)	-0.9 (-1.4, -0.4)
10	0.00 (-0.06, 0.05)	0.75	-0.8 (-1.3, -0.4)	-1.0 (-1.5, -0.5)
Heart rate			**(beats/minute)**	
1	0.07 (0.02, 0.12)	0.003	-3.9 (-5.7, -2.2)	2.4 (-0.2, 5.0)
2	0.06 (0.01, 0.10)	0.02	-3.4 (-5.2, -1.7)	1.6 (-0.9, 4.2)
3	0.05 (0.00, 0.10)	0.04	-3.2 (-5.0, -1.4)	1.2 (-1.4, 3.8)
4	0.04 (-0.01, 0.09)	0.08	-3.0 (-4.8, -1.2)	0.8 (-1.8, 3.4)
5	**0.04 (-0.01, 0.09)**	**0.1**	**-2.9 (-4.7, -1.1)**	**0.6 (-2.0, 3.3**)
6	0.04 (-0.01, 0.08)	0.14	-2.8 (-4.7, -1.0)	0.4 (-2.3, 3.1)
7	0.03 (-0.01, 0.08)	0.16	-2.8 (-4.6, -1.0)	0.3 (-2.4, 2.9)
8	0.03 (-0.02, 0.08)	0.21	-2.7 (-4.5, -0.9)	0.1 (-2.6, 2.7)
9	0.03 (-0.02, 0.08)	0.25	-2.6 (-4.4, -0.8)	-0.1 (-2.7, 2.5)
10	0.02 (-0.02, 0.07)	0.32	-2.5 (-4.3, -0.7)	-0.3 (-2.9, 2.3)
SpO_2			**(%)**	
1	0.15 (0.11, 0.19)	<0.001	-1.3 (-1.6, -1.0)	NA
2	0.12 (0.09, 0.16)	<0.001	-1.0 (-1.3, -0.7)	NA
3	0.11 (0.07, 0.14)	<0.001	-0.9 (-1.2, -0.6)	NA
4	0.11 (0.07, 0.15)	<0.001	-0.9 (-1.2, -0.6)	NA
5	**0.10 (0.07, 0.14)**	**<0.001**	**-0.9 (-1.2, -0.6)**	**NA**
6	0.10 (0.06, 0.13)	<0.001	-0.8 (-1.1, -0.6)	NA
7	0.11 (0.07, 0.14)	<0.001	-0.9 (-1.2, -0.6)	NA
8	0.10 (0.07, 0.14)	<0.001	-0.9 (-1.2, -0.6)	NA
9	0.10 (0.07, 0.14)	<0.001	-0.9 (-1.2, -0.6)	NA
10	0.10 (0.06, 0.13)	<0.001	-0.9 (-1.1, -0.6)	NA

4.3.3 Discussion

We found no clinically relevant "smoothing effect" in the CALMS-2 dataset (see results for 5-minute "during" window size in table 4.1). If vital-sign data were smoothed, manual measurements would be recorded above abnormally low continuous measurements and below abnormally high continuous measurements (with normal values unaffected), and the "smoothing effect" would cause the continuous-manual differences to be related to vital-sign values. However, differences between manual and continuous measurements of heart rate and respiratory rate were not related to the vital-sign values, as the regression coefficient of the mixed effects model (equation (4.1)) was not significantly different to 0. Differences between manual and continuous measurements of SpO_2 were related to the measurement value. However, this relationship was clinically insignificant, with minor differences shown for low values of SpO_2. For example, for SpO_2 measurements of 91% the model predicted a continuous-manual difference of -0.9 (-1.2, -0.6)%, which is within the measurement error of most pulse oximeters (Milner and Mathews [2012]). The predicted continuous-manual difference at values corresponding to high NEWS scores was also clinically insignificant for respiratory rate and heart rate, which is expected with an insignificant regression coefficient. The simulated values of continuous-manual differences from the mixed effects model (equation (4.1)) can be seen plotted in figures 4.3a to 4.3c, showing visually that there is no obvious linear relationship between the continuous-manual average and the continuous-manual difference.

Our sensitivity analysis showed that window size affected the relationship between the continuous-manual difference and the continuous-manual average but not to a clinically meaningful extent (see table 4.1). As window size was reduced, lessening the effect of median filtering, the differences predicted by the model increased, although only by 1 to 2 heart-beats or breaths. Increasing the window size did not affect our findings, suggesting that the choice of a 5-minute window (also chosen by Reich et al. [2000] and Taenzer et al. [2014]) appropriately removes error and artefact without increasing the sample size beyond what is plausible for bedside measurement. As described in section 4.1.2, four previous studies found a "smoothing effect" by using automated monitoring signals without temporal averaging, relying on manufacturer settings for artefact-removal (Block [1991]; Cook et al. [1989]; Thrush [1992]; Zollinger et al. [1977]). Sapo et al. [2009] demonstrated that SpO_2 values below 90% are associated with poor signal quality, suggesting the effects in these studies may be due to clinicians correctly removing spurious values.

A strength of this analysis is that the CALMS-2 dataset is larger than those used in most prior analyses of the "smoothing effect" (Cook et al. [1989]; Devitt et al. [1999]; Lerou et al. [1988]; Reich et al. [2000]; Taenzer et al. [2014]; Thrush [1992]; Zollinger et al. [1977]). Another advantage is the use of methodologies that are robust to the limitations of previous studies described in section 4.1.2. For example, we used contemporaneous pairs of continuous and manual vital-sign measurements, so we compare methods measuring the same physiology (Cook et al. [1989]; Devitt et al. [1999]; Lerou et al. [1988]; Thrush [1992]). We also used median filtering to remove effects of artefacts (e.g. caused by movement) in the continuous vital-sign measurements. We believe this is a better representation of the vital-sign trends than a method of randomly sampling the continuous values (that may be transient spikes) in a document (patient notes) where observations commonly occur at most hourly and are therefore intended to represent patient status for that hour. Finally, we based our analysis on the canonical Bland-Altman method (Bland and Altman [1986]) for assessing agreement between two measurement sources. We compared the continuous-manual difference to the continuous-manual average, so did not make the methodological error of Taenzer et al. [2014]. We have replicated figures 4.3a to 4.3c in appendix A.2 to demonstrate the spurious correlation that this methodological error would incur.

The finding that the "smoothing effect" does not exist in the CALMS-2 dataset has implications for the development of EWSs for continuous vital-sign monitoring. As demonstrated in chapter 3, vital-sign variability features can improve EWS performance. If the "smoothing effect" had existed, it would have suggested that the true extremal vital-sign values were not captured for vital-sign variability computation. These results suggest that the computation of 24-hour vital-sign variability features from continuous vital-sign measurements could be informative for predicting physiological deterioration.

We have provided evidence against the existence of a "smoothing effect" in post-operative care. However, this phenomenon may still exist in the context of anaesthesia. It is possible that errors in vital-sign documentation increase when clinicians record vital signs from memory after the measurement has been taken (Cook et al. [1989]; Lerou et al. [1988]). This may be more pervasive in acute episodes during surgery and may not apply to post-operative wards, where nursing staff are available to measure and document vital signs simultaneously. Our study is limited by its single-centre design, as practices may vary between hospital wards and institutions. There may have been transcription errors in the value or timing of manual vital-sign observations.

This effect was minimized by double data entry and synchronizing observation times with computer-generated timestamps of blood pressure. One bedside monitoring provider was used in this study, so we could not assess differences between monitors.

4.4 Assessing agreement

4.4.1 Methods

As described in section 4.3, we used Bland-Altman analysis to assess agreement between manual and continuous vital-sign data. We generated Bland-Altman plots for the continuous and manual vital-sign data by plotting the continuous-manual difference against the continuous-manual average. We used a "during" window size of 5 minutes as described in section 4.2.2, which was further justified in section 4.3. Horizontal lines were included on the Bland-Altman plots to show the bias and LoAs. The regression line from the linear mixed effect model described in section 4.3 was plotted to visualize the relationship between the differences and the averages.

To calculate the bias and LoA for the method comparison of continuous and manual vital-sign data, we followed the methodology of Carstensen et al. [2008]. We used the specific methodology for measurements with linked replicates across subjects (in this case patients), with a constant bias between methods. Replicates were linked (as opposed to exchangeable) because the true vital-sign value differed across replicates, and the measurements by each method (continuous and manual) were paired at each replicate (as they were measured at the same time). The assumption of constant bias between methods was validated in section 4.3 as there was no meaningful relationship between the difference between methods and the value of the vital sign. In the methodology of Carstensen et al. [2008], the measurement values from a particular measurement method are modelled with a linear mixed effects model, taking into account the possible sources of variance. The model was fitted with the *MethComp* package in R (Carstensen et al. [2016]), with a structure given in equation (4.2):

$$\begin{gathered} y_{mir} = \alpha_m + \mu_i + a_{ir} + c_{mi} + e_{mir} \\ a_{ir} \sim \mathcal{N}(0, \omega^2), \quad c_{mi} \sim \mathcal{N}(0, \tau^2), \quad e_{mir} \sim \mathcal{N}(0, \sigma_m^2) \end{gathered} \tag{4.2}$$

where y_{mir} is the replicate (r) measurement taken with method m from patient i. α_m is a fixed-effect term for method m (a constant offset) and μ_i is a fixed-effect term

for the patient i (a baseline patient measurement value). a_{ir} is a random interaction term between patient i and replicate r with variance ω^2 (variance assumed constant between items and replicates). c_{mi} is a random interaction term capturing the variation between items i for method m, with variance τ^2 (variance was assumed constant between methods as separate τ_m cannot be estimated with only two methods). e_{mir} is the model residual term with variance σ_m^2.

The bias and LoA can be estimated with the fitted parameters from the model in equation (4.2) as described by Carstensen et al. [2008]. The bias is the difference between the fixed effect (offset) terms between measurement methods (equation (4.3)). The limits of agreement are calculated from the variance of the differences (equation (4.4)). Here the hat operator $\hat{}$ refers to the measured value.

$$bias = \hat{\alpha}_1 - \hat{\alpha}_2 \tag{4.3}$$

$$LoA = \hat{\alpha}_1 - \hat{\alpha}_2 \pm 1.96 \times \sqrt{2\hat{\tau}^2 + \hat{\sigma}_1^2 + \hat{\sigma}_2^2} \tag{4.4}$$

95% confidence intervals (CIs)were calculated for the bias and the LoAs using the method recommended by Bland and Altman [1999]. The 95% CI calculation for the bias is shown in equation (4.5). The 95% CI calculation for the LoA is shown in equation (4.6).

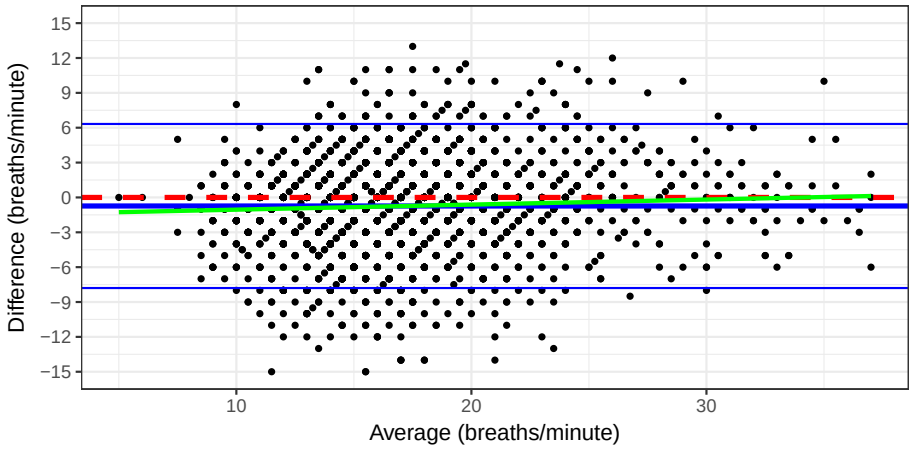
$$CI_{bias}^{95} = bias \pm 1.96 \times \frac{\sqrt{2\hat{\tau}^2 + \hat{\sigma}_1^2 + \hat{\sigma}_2^2}}{\sqrt{n}} \tag{4.5}$$

$$CI_{LoA}^{95} = LoA \pm 1.71 \times 1.96 \times \frac{\sqrt{2\hat{\tau}^2 + \hat{\sigma}_1^2 + \hat{\sigma}_2^2}}{\sqrt{n}} \tag{4.6}$$

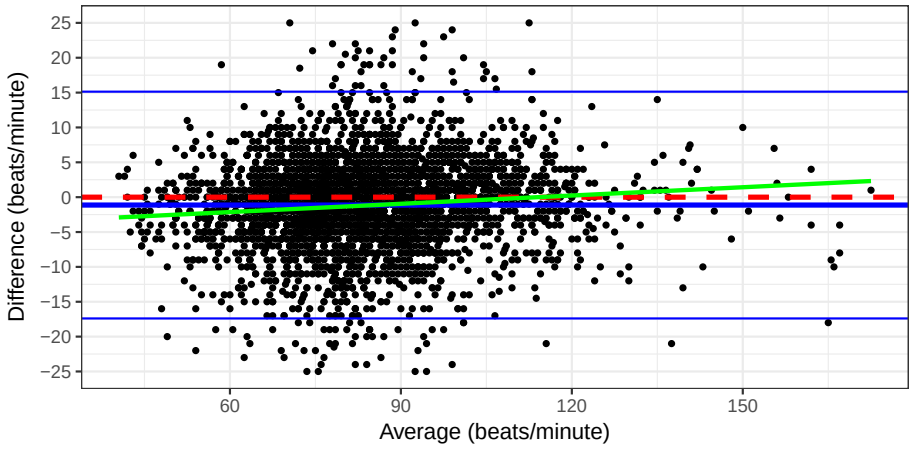
where CI_{bias}^{95} is the 95% CI for the bias and CI_{LoA}^{95} is the 95% CI for the LoA. n is the number of vital-sign measurements. 1.96 is the value of the 97.5 percentile of the Student t distribution (valid for large n) and the multiplier 1.71 was derived by Bland and Altman [1999] to allow accurate approximation of the precise CI for large n.

4.4.2 Results

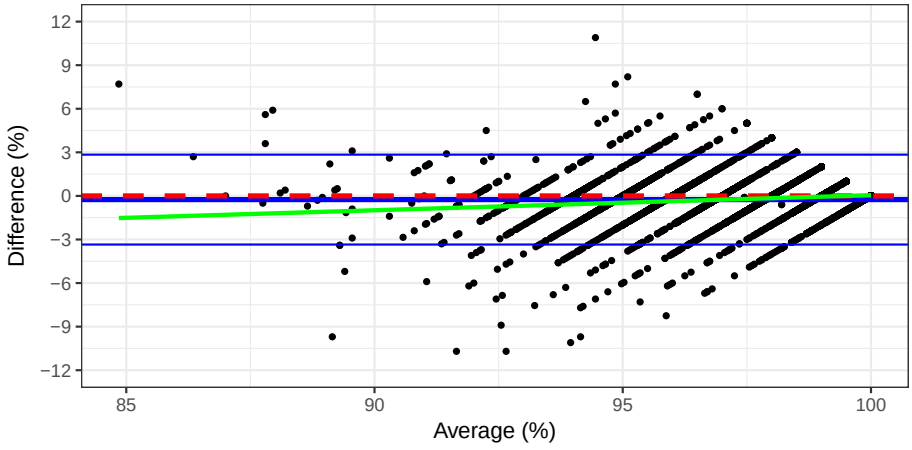
Figures 4.4a to 4.4c show Bland-Altman plots for continuous and manual measurements of respiratory rate, heart rate and SpO_2. The bias, LoAs, and the simulated predictions from the linear mixed effects model (from section 4.3) are shown on the plots. The bias and LoAs for each vital sign are given in table 4.2.



(a) Respiratory rate



(b) Heart rate



(c) SpO_2

Figure 4.4: Bland-Altman plots for the comparison of continuous and manual measurements of respiratory rate (a), heart rate (b) and SpO_2 (c). The data points for measurements from the CALMS-2 dataset are shown as black dots, the bias and LoAs are shown in blue (thick middle line is bias) and the red dashed line is the zero y-intercept. The predicted difference at continuous-manual average values simulated from the linear mixed effects model (see equation (4.1)) is shown in green. Difference is continuous - manual.

Table 4.2: Bias and LoAs for the Bland-Altman analysis comparing continuous vital-sign measurements and manual vital-sign measurements of respiratory rate, heart rate and SpO_2 (see section 4.4). C.I. is 95% confidence interval.

Vital-sign	Bias (C.I.)	Lower LoA (C.I.)	Upper LoA (C.I.)
Respiratory rate (breaths/minute)	-0.74 (-0.82, -0.66)	-7.80 (-7.94, -7.66)	6.32 (6.18, 6.46)
Heart rate (beats/minute)	-1.13 (-1.32, -0.94)	-17.4 (-17.7, -17.1)	15.1 (14.8, 15.5)
SpO_2 (%)	-0.25 (-0.29, -0.22)	-3.35 (-3.41, -3.29)	2.84 (2.78, 2.90)

4.4.3 Discussion

The bias (95% CI) values between continuously-recorded and manually-recorded vital signs were -0.74 (-0.82, -0.66) breaths/minute for respiratory rate, -1.13 (-1.32,-0.94) beats/minute for heart rate and -0.25 (-0.29, -0.22)% for SpO_2 (table 4.2). These values are low and would be considered clinically insignificant. Figures 4.4a to 4.4c show the bias to be close to zero for all three vital signs. Thus, there is no systematic bias between manual and continuous vital-sign measurements in the CALMS-2 dataset.

The LoAs (95% CI) between continuously-recorded and manually-recorded vital signs were large. Subtracting the lower LoA from the upper LoA (from table 4.2) gives values of 14.1 (13.8, 14.4) breaths/minute, 32.5 (31.9, 33.3) heart-beats/minute and 6.2 (6.1, 6.3)% SpO_2. These values are large and are clinically significant, showing that manually-recorded respiratory rate, heart rate and SpO_2 measures varied widely from the continuous measure. Respiratory rate is difficult to measure clinically using both manual and automated methods (Hooker et al. [1989]; Lovett et al. [2005]) so the high variance in the differences between continuous and manual measurements is perhaps not unexpected. However, as EWSs commonly include respiratory rate ranges between two scores of 4 breaths/minute or less (McGinley and Pearse [2012]; Tarassenko et al. [2011]), these differences would commonly impact clinical care. The LoAs for heart rate and SpO_2 were also wide and could result in different EWS values.

These results are important in the context of transitioning to the use of automated continuous monitoring sources with an integrated EWS. Most EWSs have not been designed for use with continuous data so patients would clearly alert at different threshold values if a manual monitoring method was naively exchanged for a continuous monitoring method for EWS computation. The results suggest that EWSs should be re-designed for vital-sign values recorded continuously with patient monitors. For EWS models based on machine learning, this would involve collating a database of continuous vital-sign values for training an algorithm for continuous vital-sign data.

The results in section 4.3 show that manual vital-sign measurements are not affected by a "smoothing effect", suggesting that the predictive benefits of vital-sign variability features derived in chapter 3 may apply to continuous vital-sign data. However, a continuous monitoring system with an integrated EWS algorithm using vital-sign variability features should be trained on a database of continuous vital-sign data.

4.5 Assessing the "arousal effect"

4.5.1 Methods

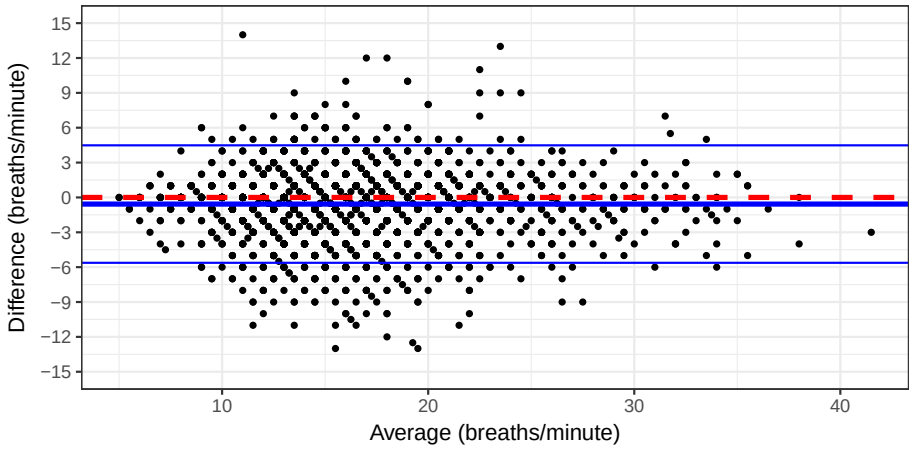
We assessed whether there is an "arousal effect" caused by the vital-sign recorder by comparing continuous data before and during the time of manual observation. To do so, we compared the continuous data from the "before" window to the continuous data from the "during" window. We used a "during" window size of 5 minutes and a "before" window size of 15 minutes, as described in section 4.2.2. To obtain the differences between the continuous vital-sign values from the two measurement windows, we subtracted the "during" window value from the "before" window value in each measurement pair. We also calculated the average of the "during" window value and the "before" window value for each measurement pair. We again used Bland-Altman analysis for this purpose, using the methodology described in section 4.4. We prepared Bland-Altman plots for the "before" window and "during" window measurements, and computed the bias and LoA.

4.5.2 Results

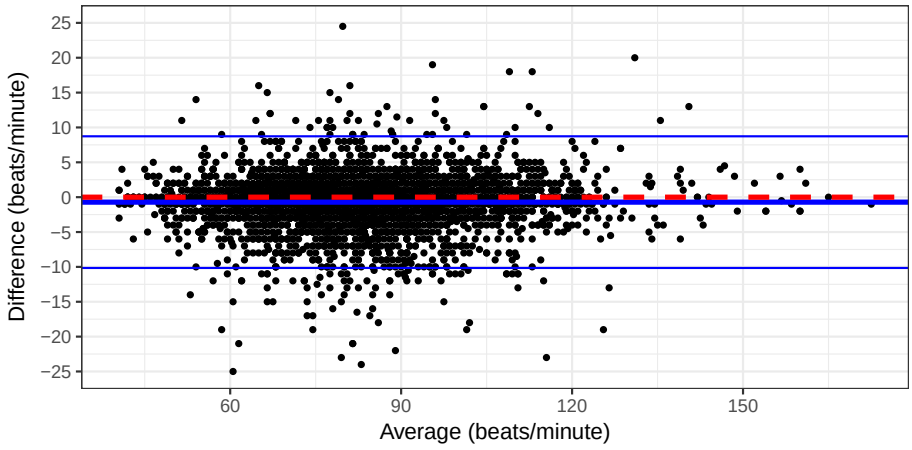
Figures 4.5a to 4.5c show Bland-Altman plots for continuous measurements of respiratory rate, heart rate and SpO_2 before and during the time of clinical observation. The bias and LoAs are shown on the plots and are given in table 4.3.

Table 4.3: Bias and limits of agreement for the Bland-Altman analysis comparing continuous measurements of respiratory rate, heart rate and SpO_2 before and during the time of clinical observation (see section 4.5). C.I. is 95% confidence interval.

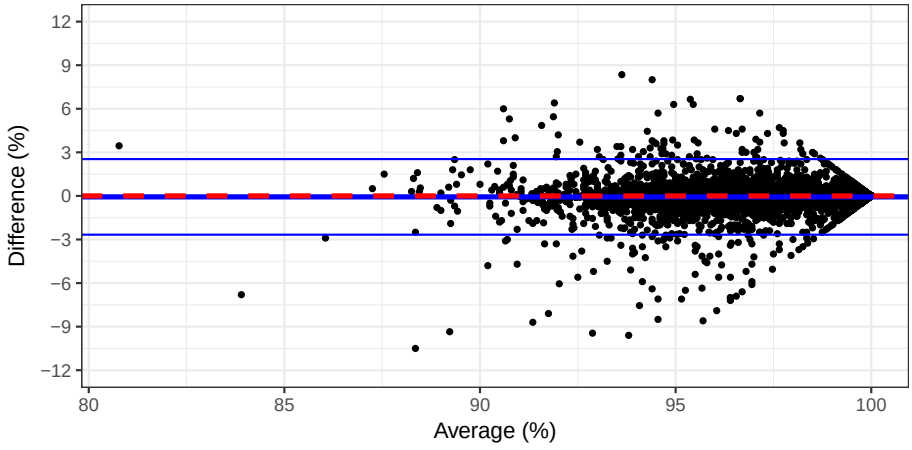
Vital-sign	Bias (C.I.)	Lower LoA (C.I.)	Upper LoA (C.I.)
Respiratory rate (breaths/minute)	-0.57 (-0.63, -0.51)	-5.63 (-5.72, -5.53)	4.48 (4.38, 4.58)
Heart rate (beats/minute)	-0.71 (-0.82, -0.60)	-10.2 (-10.3, -9.97)	8.73 (8.55, 8.92)
SpO_2 (%)	-0.07 (-0.10, -0.04)	-2.67 (-2.72, -2.62)	2.54 (2.49, 2.59)



(a) Respiratory rate



(b) Heart rate



(c) SpO_2

Figure 4.5: Bland-Altman plots for the comparison of continuous measurements of respiratory rate (a), heart rate (b) and SpO_2 (c) before and during the time of clinical observation. The data points for measurements from the CALMS-2 dataset are shown as black dots, the bias and LoAs are shown in blue (thick middle line is bias) and the red dashed line is the zero y-intercept. Difference is calculated as "before" window value - "during" window value.

4.5.3 Discussion

There was no systematic bias between the vital-sign values measured before and during clinical observation. The bias was -0.57 (-0.63, -0.51) breaths/minute for respiratory rate, -0.71 (-0.82, -0.60) beats/minute for heart rate and -0.07 (-0.10, -0.04)% for SpO_2 (table 4.3). The bias can be seen to be close to zero in Figures 4.5a to 4.5c.

If the patient's vital signs were consistently elevated during clinical observation then the vital-sign values in the "during" window would be consistently higher than those in the "before" window. However, our results show that there is no bias between continuous vital-sign values recorded in the "before" window and the "during" window. We conclude that there is no "arousal effect" from the vital-sign recorder.

Our method for assessing the "arousal effect" follows Taenzer et al. [2014], who compared SpO_2 values in the "during" window and the "before" window of continuous data, finding no evidence of an "arousal effect". Our results therefore confirm those of Taenzer et al. [2014], using a larger database. Our contribution includes the analysis of two additional vital signs (respiratory rate and heart rate).

It is important that continuous monitoring systems measure reliable values in the presence of clinical staff. While an automated continuous monitoring system would increase the frequency of vital-sign measurement and potentially reduce the burden of clinical workload, it would not replace the role of clinical staff in assessing the patient's condition. If an EWS system based on continuous vital-sign data were influenced by the presence of clinical staff, it may affect how clinical staff could interact with the system. Thus, our finding that the "arousal effect" does not exist in the CALMS-2 dataset suggests that continuous monitoring systems can be integrated with other clinical observations without having a systematic error introduced by the presence of the vital-sign recorder.

4.6 Conclusions

We found no evidence that a clinically significant "smoothing effect" exists for respiratory rate, heart rate or SpO_2 in acute post-operative care. We also found no evidence of an "arousal effect" caused by the vital-sign recorder. Differences between manually and continuously-recorded vital signs were frequently large, suggesting the methods cannot be used interchangeably. The effects of the large differences between manually-recorded and continuous vital-sign values on EWSs require investigation before measurement can be safely automated for EWS computation.

Chapter 5

Continuous vital-sign variability

In chapter 3 we demonstrated that changes in diurnal vital-sign variability are indicative of impending physiological deterioration, and that the inclusion of diurnal vital-sign variability features in a machine learning EWS model can improve the model performance for predicting adverse clinical events. However, the models developed in chapter 3, like other clinical EWS systems, used routinely collected manual vital-sign observations. Since manual observations are collected every 1-4 hours in post-operative wards, EWS systems based on manual observations are unable to detect acute physiological changes that occur between nurse observations. This restricts the timeliness of EWS systems, as physiological deterioration may remain undetected for several hours before a clinical observation occurs. Furthermore, physiological changes in the vital-sign record may be under-sampled, which affects the computation of vital-sign variability as physiology that is not measured cannot be included in the calculation of diurnal variability.

As described in section 1.4, the use of continuous vital-sign monitoring equipment may improve the detection of physiological deterioration through real-time tracking of the patient's health status throughout their stay on the ward. Clinically important changes in physiology could be quickly measured, and trends could be tracked as they develop, allowing timely delivery of clinical interventions to prevent further physiological deterioration. In addition, all changes in physiology would be measured throughout the diurnal cycle, allowing accurate computation of vital-sign variability. While delivering reliable continuous monitoring to post-operative patients has proved challenging (see section 1.4.2), the potential benefits motivate further work to improve algorithms and systems for continuous monitoring. In particular, the results from chapter 3 showing the value of diurnal vital-sign variability features suggest that the computation of diurnal variability from continuous vital-sign data, and the development of an EWS model for continfuous vital-sign data, deserve further exploration.

Computing diurnal variability and EWSs from continuous vital-sign data may seem trivial under the assumption that continuous data are the same as manual data, except

more frequently sampled. However, the method comparison study of continuous and manual vital-sign data conducted in chapter 4 showed that this is not the case. We found that continuous and manual measurements of respiratory rate, heart rate and SpO_2 had wide limits of agreement, suggesting that there can be clinically significant differences in the values measured by continuous and manual methods. The cause of these differences is unclear, as we found no evidence of any bias introduced by clinical staff through a "smoothing effect" or an "observer effect". In the context of EWS and vital-sign variability computations, these findings suggest that EWS models and diurnal vital-sign variability computation methods should be developed specifically for continuous monitoring sources since previous methods, such as those used in chapter 3, may not be optimal for continuous data.

In this chapter we investigate the computation of diurnal vital-sign variability features from continuous monitoring data. We used the continuous vital-sign data from bedside monitors and wearable ambulatory monitors collected during the CALMS-2 study. As described in chapter 2, the duration of monitoring from both bedside and wearable sources varied between patients, and many patients did not reliably comply with the use of wearable monitors. As a result, the continuous data were not of sufficient quantity or quality to be used for the development of EWS models for continuous data. We therefore did not apply the methodology used in chapter 3, in which measures of diurnal vital-sign variability were evaluated as features in EWS models. We instead assessed the value of diurnal vital-sign variability features from continuous data by exploring the differences in feature distributions and trajectories between periods of physiological deterioration and physiological stability.

We begin this chapter by describing previous work on the use of variability and trend features from continuous vital-sign data to detect physiological deterioration (section 5.1). We use case studies to explore differences between continuous and manual vital-sign time-series preceding adverse clinical events (section 5.2). We then describe the computation and evaluation of different methods of measuring vital-sign variability from continuous data (section 5.3). We provide results assessing the best variability computation methods for the CALMS-2 dataset, and present case studies of vital-sign variability preceding adverse clinical events and discharge from the ward (section 5.4). Finally, we discuss and interpret our findings in section 5.5.

5.1 Previous work

In chapter 1 we provided a review of continuous vital-sign monitoring in post-operative wards. In chapter 3 we described previous studies that used long-term variability or trend features from *routinely-collected* vital-sign data to detect physiological deterioration. Here we focus on studies related to the computation of long-term variability or trend features from *continuous* vital-sign data for detecting physiological deterioration.

The ability to continuously measure diurnal changes in the vital signs has long been recognised as a potential benefit of the introduction of continuous monitoring technology, such as wearables, to wards that do not currently provide continuous monitoring (Binkley [2003]). However, we found very few previous studies that have implemented or analysed continuous monitoring signals in this context. Most previous studies related to vital-sign variability computed from continuous monitoring data have been conducted in the ICU (Cao et al. [2008]; Li-wei et al. [2015]) or other high acuity populations such as the NICU (Moorman et al. [2011]; Saria et al. [2010b]). Vital-sign variability features have typically been computed from short term signal dynamics (as described in section 3.1) using a range of pre-specified methods (Bravi et al. [2011]) or using learned features from time-series based machine learning models (Li-wei et al. [2015]; Saria et al. [2010b]).

Of the studies that measured vital-sign changes over longer time intervals from continuous monitoring data, the most common biomarker was the diurnal variability of the interval between successive heart beats, which is used to measure autonomic nervous system[1] function and predict cardiovascular mortality (Camm et al. [1996]). This is commonly measured in cardiac patients using ambulatory Holter monitors (see section 1.4.2), for example to characterise cardiac function in patients with tachycardia (Rubenstein et al. [2010]). More relevant to our analysis, increases in diurnal inter-beat interval variability were shown to predict improved patient outcomes in rapid response team calls from general hospital wards (Salahuddin et al. [2018]). However, inter-beat interval measurement requires high-bandwidth continuous vital-sign data (such that each individual heart beat can be accurately sampled) that is not easily obtained in a long-term ambulatory post-operative setting. To our knowledge, no previous studies have analysed the diurnal variability from continuous measurements of any of the five main vital signs in a post-operative setting (either from a bedside monitor or an ambulatory monitor). We note that Hravnak et al. [2011] documented the occurrence

[1]The autonomic nervous system controls bodily functions not consciously directed, such as the heart rate, the respiratory rate, and digestion.

of repetitive cycling of vital signs above and below critical values preceding adverse clinical events in a post-operative ward, although these patterns were not encoded in their EWS computation methodology.

There is clearly an opportunity to explore the value of long-term variability features from continuous monitoring data for detecting physiological deterioration in post-operative and general wards. This has been recognised by Fritz et al. [2018], who have published a study protocol to collect continuous monitoring data from general ward patients for the development of machine learning EWS models that include temporal vital-sign patterns. Hackmann et al. [2011] also recognised this opportunity, describing a future two-tier system beginning with an EWS with temporal features (using bucketing as described in section 3.2.1) followed by the provision of ambulatory monitoring to those patients whose vital signs prompt an EWS alert. In this chapter we make use of the data collected during the CALMS-2 study to explore methodologies for computing diurnal vital-sign variability from continuous monitoring data as a first step in the development of future EWS systems for non-invasive continuous monitoring of post-operative patients.

5.2 The value of continuous vital-sign data

In this analysis we explore the additional value that continuous vital-sign monitoring may have over manual vital-sign observations. To do so, we assess the time-series of the five main vital signs from six patients in the CALMS-2 dataset, as shown in figures 5.1 to 5.6. We present three "normal" patients who did not have an adverse clinical event (figures 5.1 to 5.3), and three "abnormal" patients who suffered an adverse clinical event (figures 5.4 to 5.6). For the "normal" patients we plotted the final 48 hours of their stay on the ward before discharge (showing physiological stability), and for the "abnormal" patients we plotted the final 48 hours on the ward before the adverse clinical event (showing physiological deterioration). The six patients were chosen because they all had continuous vital-sign monitoring throughout the 48-hour period before the event/discharge. Note that the plots may have different y-axis scales due to different vital-sign ranges for each patient.

Figures 5.1 and 5.2 show time-series for patients who had stable vital-sign values, with no major changes preceding discharge. Both patients had continuous vital-sign monitoring from the ambulatory monitor, but not the bedside monitor. Continuous data were therefore collected for heart rate and SpO_2 (as the ambulatory monitor was a pulse oximeter, see chapter 2), while only manual vital-sign observations were

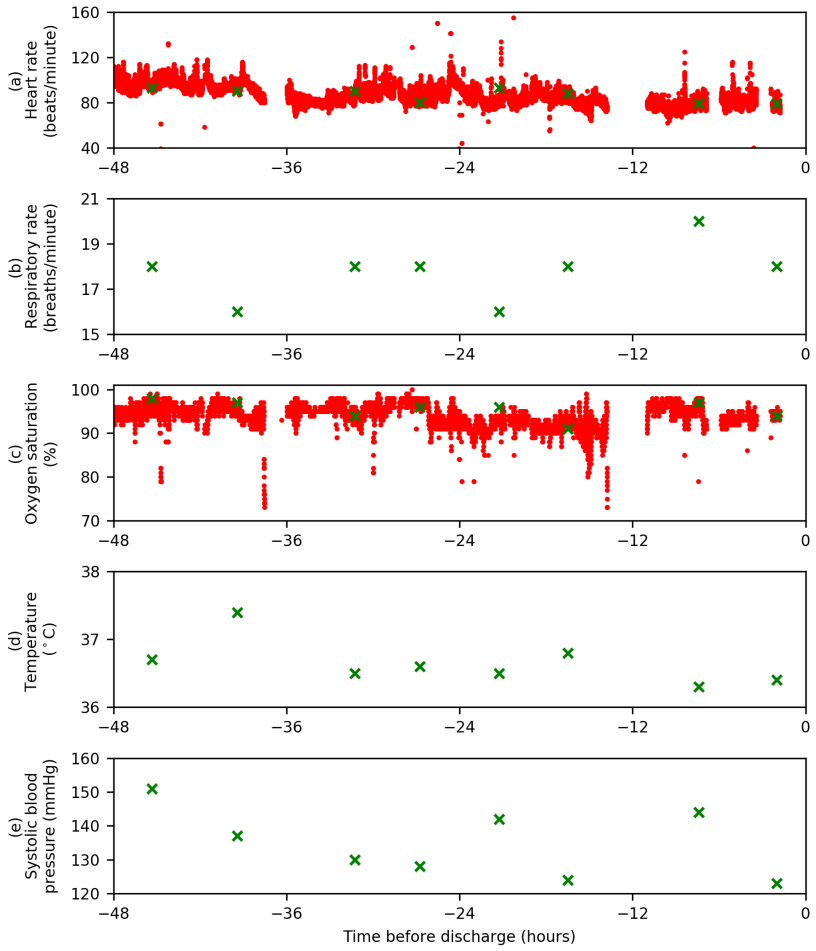


Figure 5.1: Vital-sign time-series for CALMS-2 patient 447 in the final 48 hours before discharge from the ward. Continuous vital-sign values from the ambulatory monitor are shown in red and manual observations are shown in green.

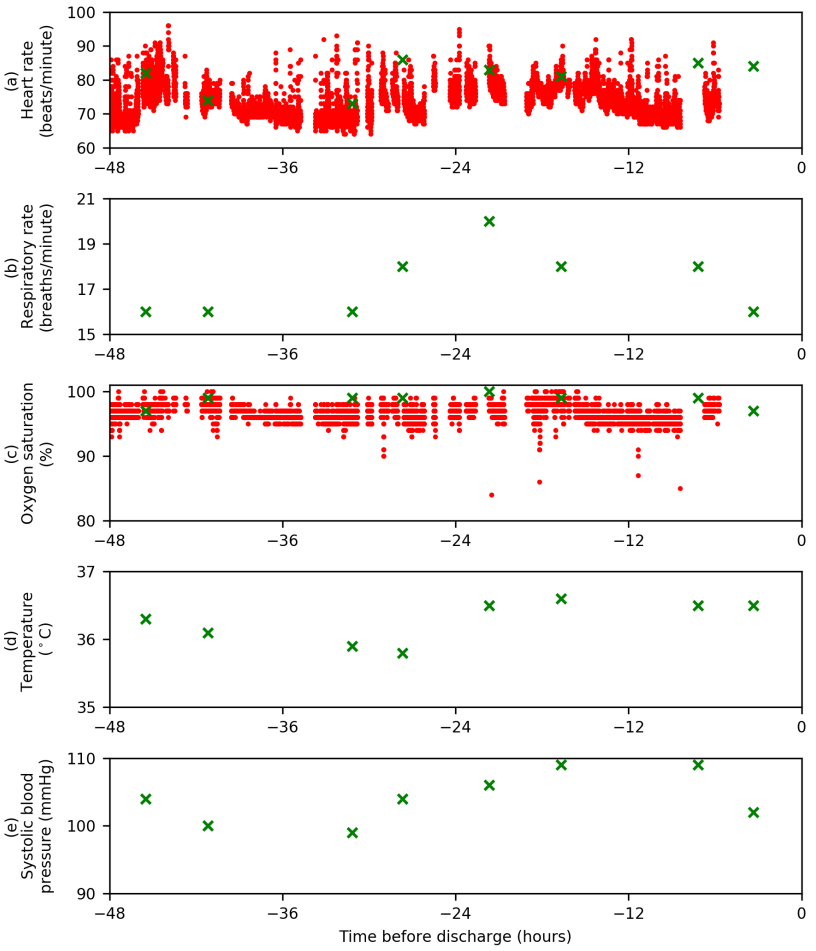


Figure 5.2: Vital-sign time-series for CALMS-2 patient 462 in the final 48 hours before discharge from the ward. Continuous vital-sign values from the ambulatory monitor are shown in red and manual observations are shown in green.

collected for respiratory rate, systolic blood pressure and temperature. The continuous data do not show any periods when the continuous monitoring signals deviated from the trend in the manual observation time-series. Thus, in this instance the continuous monitoring offered little additional value over the manual observations, which adequately sampled the patients' physiology. This is expected, as patients should have very stable physiology before they are discharged from the ward.

Figure 5.3 shows the vital-sign time-series of a "normal" patient before they were discharged from the ward. The patient had continuous bedside monitoring between 48 to 32 hours before discharge from the ward, which provided continuous data for heart rate, respiratory rate and SpO_2. The patient had continuous ambulatory monitoring of heart rate and SpO_2 for the entire 48-hour period, although there were intermittent periods of lost signal for multiple hours at a time. This demonstrates the issues with reliability and compliance of the ambulatory monitors, as discussed in chapter 2. Again, the patient's physiology was stable preceding discharge from the ward, and the continuous data followed the same trajectory as the manual data. While there were transient changes in the continuous vital-sign values (such as the variance in the respiratory signal at 36-48 hours before discharge), this is likely to be a reflection of the level of measurement noise that exists in continuous vital-sign monitoring data rather than true changes in the patient's physiology.

Figure 5.4 shows the vital-sign time-series of an "abnormal" patient in the 48 hours before an adverse event. The patient arrived on the ward 43 hours before the adverse event, so there is no data for the first 5 hours in figure 5.4. The patient only received continuous monitoring from bedside equipment, which was likely because their condition never became sufficiently stable for the patient to be moved from HDU-level care to general ward level care (these two levels were described in chapter 2). The patient's heart rate, respiratory rate and temperature increased, and SpO_2 decreased, approximately 20 hours before the adverse event. This change in physiology was captured by both the continuous and manual vital-sign data, although the change is visible approximately 4 to 5 hours earlier in the continuous data (24 hours before the event). The frequency of manual observations was greater compared with the previous three patients, particularly following admission to the ward and before the adverse event. It is common for clinical staff to increase vital-sign observation frequency when there is concern about a patient. The goal is the same as with continuous monitoring – to increase the timeliness of detecting physiological deterioration, although with continuous monitoring technology this level of surveillance can be provided at all times, and without the added workload of taking more manual measurements.

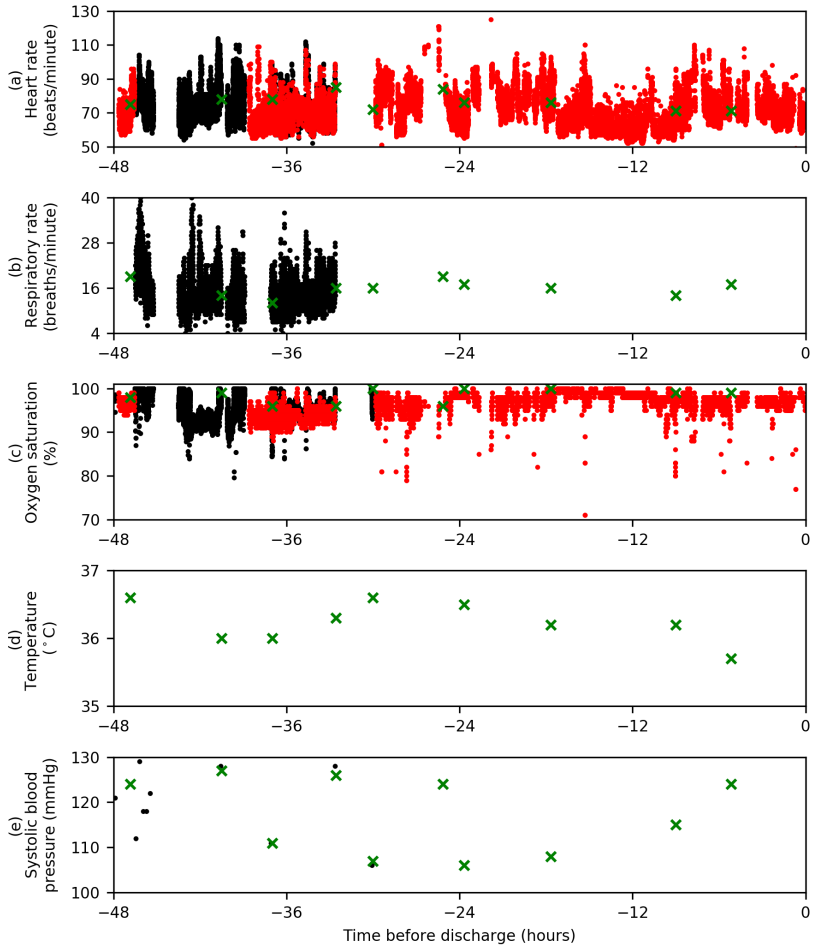


Figure 5.3: Vital-sign time-series for CALMS-2 patient 471 in the final 48 hours before discharge from the ward. Continuous vital-sign values from the ambulatory monitor are shown in red, continuous values from the bedside monitor are shown in black, and manual observations are shown in green.

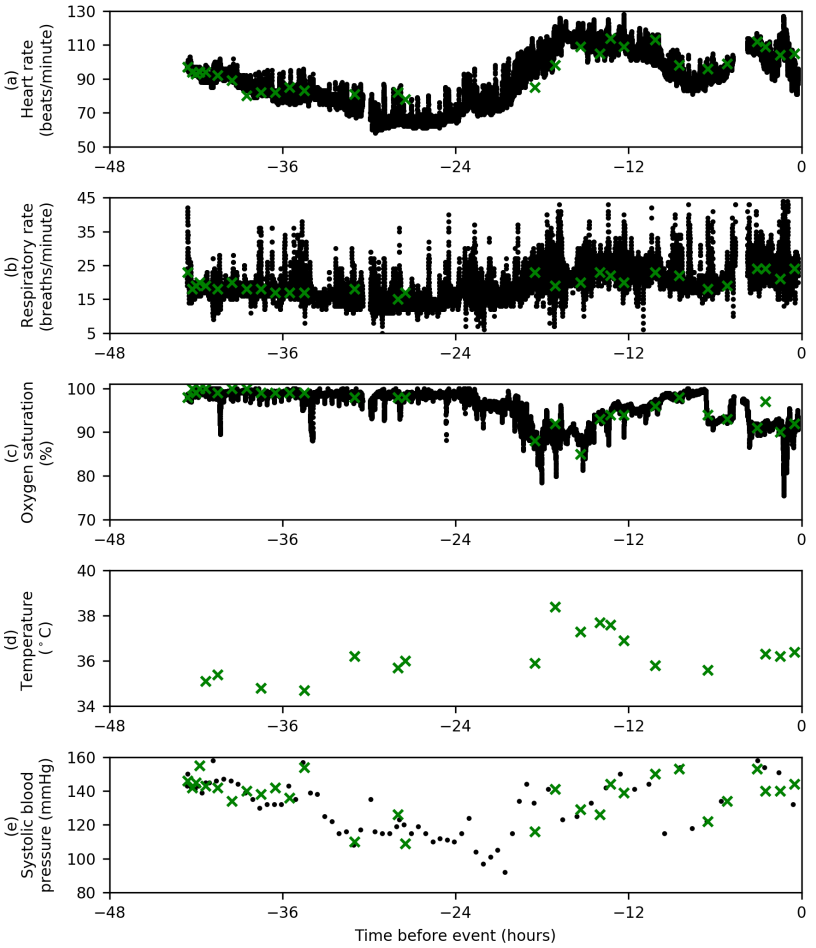


Figure 5.4: Vital-sign time-series for CALMS-2 patient 348 in the final 48 hours before an adverse clinical event. Continuous vital-sign values from the bedside monitor are shown in black and manual observations are shown in green.

Figures 5.5 and 5.6 show the vital-sign time-series of two "abnormal" patients in the 48 hours before an adverse event. Both patients received continuous monitoring from ambulatory monitors (which provided heart rate and SpO_2), before a change in physiology prompted clinical staff to provide them with continuous bedside monitoring. For the patient in figure 5.5, this probably coincided with the increase in respiratory rate and heart rate approximately 20 hours before the adverse event. This change in heart rate was tracked by the continuous ambulatory monitor slightly sooner than the manual measurements. There were also transient decreases in SpO_2 visible in the bedside monitoring data and manual observation data approximately 12 hours before the event. For the patient in figure 5.6, the prompt to provide bedside vital-sign monitoring probably coincided with the increase in heart rate, respiratory rate and systolic blood pressure, as well as the decrease in SpO_2, that occurred approximately 30 hours before the adverse event. The change in heart rate was tracked by the ambulatory monitor. Shortly after bedside monitoring was provided for the patient, their respiratory rate increased significantly, so this was tracked by both the continuous and manual data. The initial decrease in SpO_2 was not captured by continuous data, although the ambulatory monitor had been worn in the hours before and after this occurred, suggesting that there was a loss in ambulatory monitoring signal. Again, there was a significant increase in the frequency of manual observations in the 12 to 24 hours preceding the adverse event.

Overall, figures 5.1 to 5.6 demonstrate several interesting phenomena related to continuous monitoring in post-operative wards. Firstly, continuous vital-sign monitoring is not likely to improve the quality of care for patients who have stable physiology, as intermittent manual observations adequately sample their physiology during this time (figures 5.1 to 5.3). Secondly, changes in physiology may be detected earlier with continuous vital-sign monitoring (figures 5.4 to 5.6) and can provide similar levels of monitoring as increasing the frequency of manual observations, without the additional workload. Finally, continuous vital-sign monitoring must be reliable to improve the efficacy of monitoring over manual observations (see figure 5.6).

Continuous vital-sign monitoring data, when collected reliably, capture all relevant vital-sign changes, thus providing a rich data source for the computation of diurnal vital-sign variability. However, continuous signals contain transient changes in vital-sign values that may be physiological, or may be noise. Thus, the method of estimating variability with the maximum and minimum values in a 24-hour window (used in chapter 3) may not be optimal. In the following section of this chapter, we describe methods for estimating diurnal variability with continuous vital-sign data.

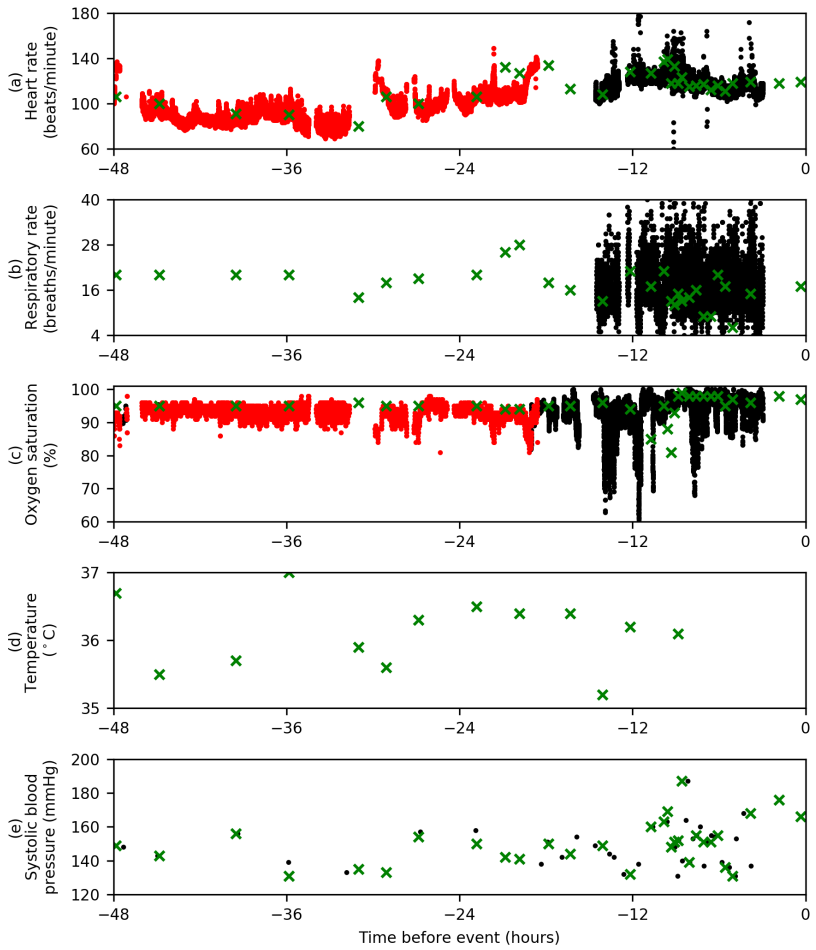


Figure 5.5: Vital-sign time-series for CALMS-2 patient 362 in the final 48 hours before an adverse clinical event. Continuous vital-sign values from the ambulatory monitor are shown in red, continuous values from the bedside monitor are shown in black, and manual observations are shown in green.

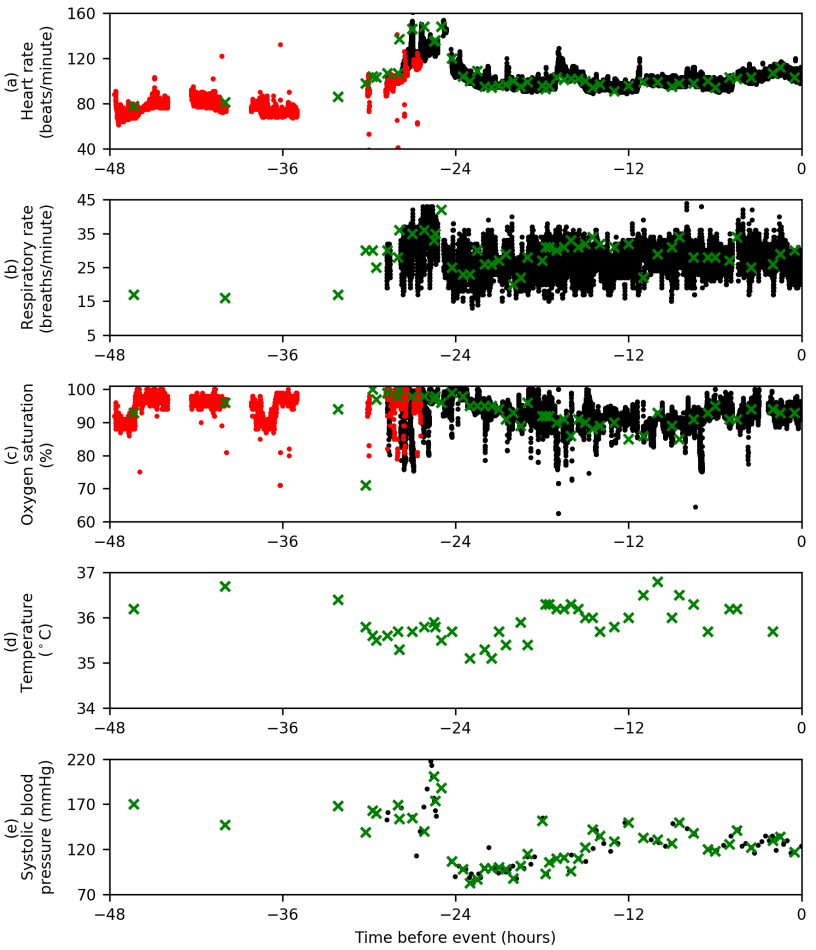


Figure 5.6: Vital-sign time-series for CALMS-2 patient 464 in the final 48 hours before an adverse clinical event. Continuous vital-sign values from the ambulatory monitor are shown in red, continuous values from the bedside monitor are shown in black, and manual observations are shown in green.

5.3 Computing variability from continuous data

5.3.1 Overview

We developed methodologies to derive diurnal variability features from continuous vital-sign data, and used several metrics to evaluate the best method for each vital sign. As continuous data were only available for heart rate, respiratory rate and SpO_2, we could only compute diurnal variability features for these vital signs. In figure 5.7 we provide a schematic overview of the vital-sign variability computation and evaluation procedure. The methodologies are described in detail in sections that follow.

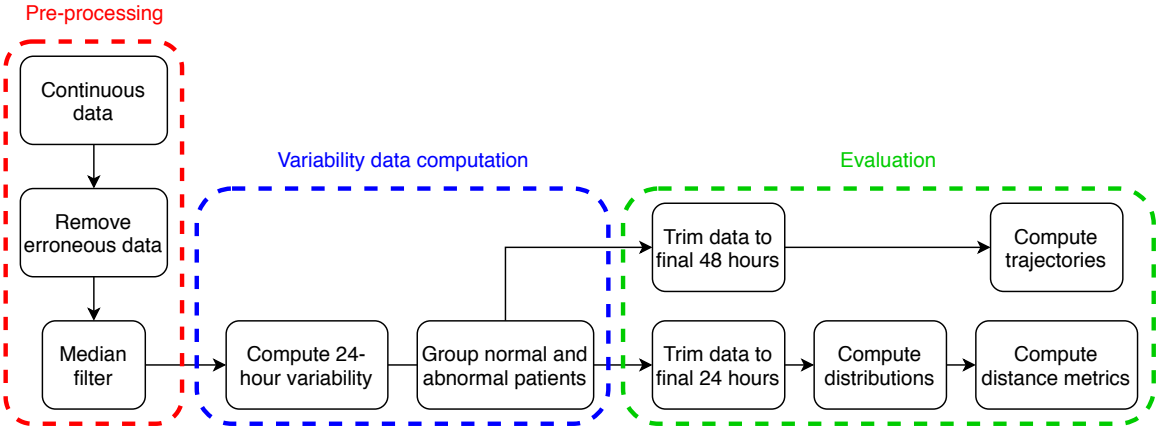


Figure 5.7: The analysis pipeline for computing and evaluating diurnal variability features from continuous vital-sign data for heart rate, respiratory rate and SpO_2. Erroneous values were removed based on physiological limits and 5-minute median filtering was applied. Variability was computed for each vital sign using several methods, and for each of these the data were separated into "normal" and "abnormal" patient groups. Probability distributions were estimated and different metrics were used to compute the distance between the distributions for each patient group in the final 24 hours before an adverse event (for "abnormal" patients) and discharge from the ward (for "normal" patients). Average variability trajectories were computed for each patient group in the final 48 hours before the event/discharge.

5.3.2 Pre-processing

We began our analysis by combining the vital-sign data from the bedside monitors with the ambulatory monitors, such that there was one continuous vital-sign time-series for each vital sign, for each patient. We then ensured that all vital-sign values were physiologically plausible by removing all vital-sign measurements from the continuous data that were outside the physiological limits shown in table 3.1. We then used

median filtering to remove transient spikes from the continuous vital-sign data, such that only persistent trends and changes were measured. Median filtering was performed with centred sliding 5-minute windows, such that the filtered data were the median of all measurements 2.5 minutes before and after the time of the samples. This was consistent with the time-window used for median filtering in chapter 4 (to compute the "during" window) and was appropriate for the computation of diurnal variability, which measures long-term changes in the vital-sign values rather than short-term fluctuations. An example 24-hour period of heart rate, respiratory rate and SpO_2 data is plotted in figure 5.8, showing the raw measurements and the filtered measurements. The filtered measurements capture persistent trends in the data, while removing transient spikes. For example, the increase in heart rate and decrease in SpO_2 in the first 12 hours were tracked with the filtered signal, without the high signal variance of the raw signal.

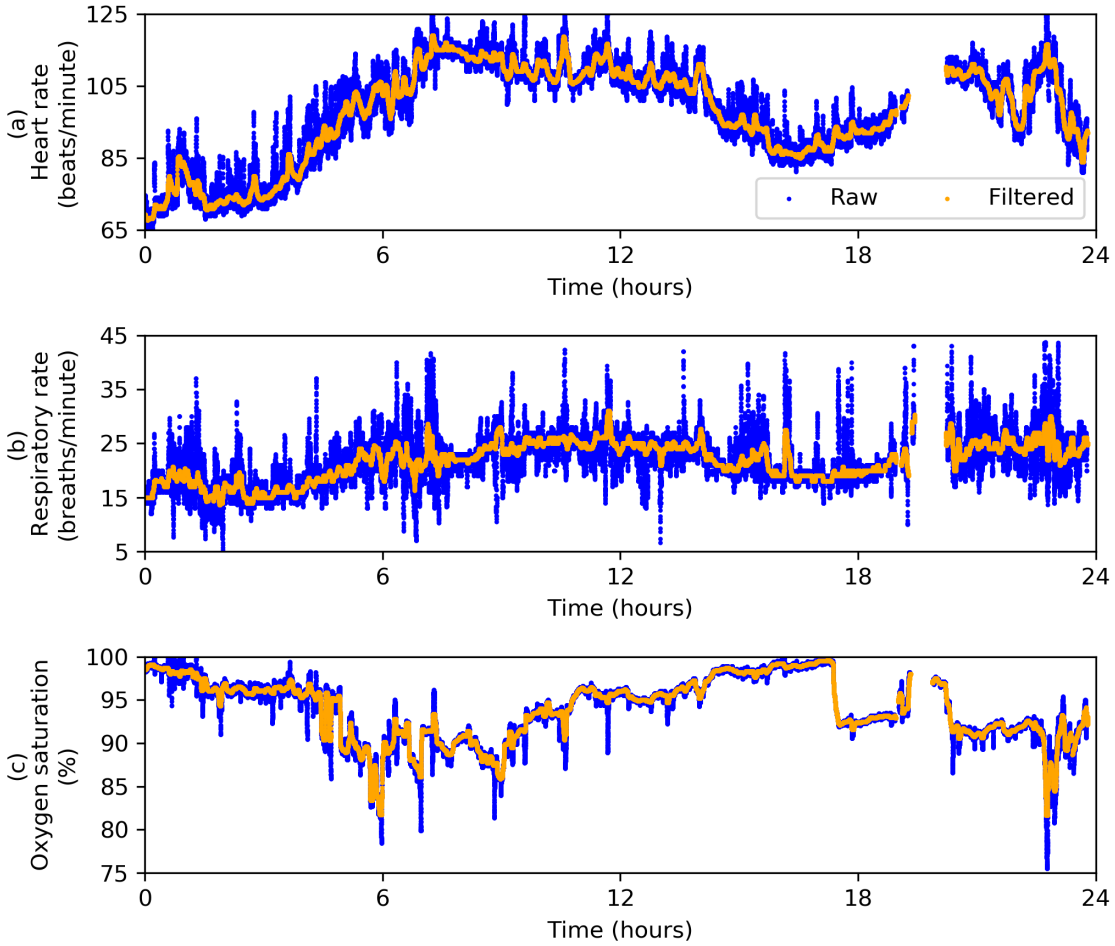


Figure 5.8: Example vital-sign time-series of heart rate (a), respiratory rate (b), and SpO_2 (c) for a 24-hour period. The raw vital-sign data are shown in blue, and the (5-minute) median-filtered data are shown in orange.

5.3.3 Variability calculation methods

To compute the diurnal vital-sign variability at a given time-point we used a similar methodology to that described in chapter 3, in which the variability was computed as the maximum - minimum value over the previous 24 hours. However, we adapted this methodology to be suitable for continuous data. Continuous vital-sign data contains changes on many time-scales. We removed the very short-term changes by median filtering, while changes over many minutes or hours remain. It is unclear how long a patient needs to stay at a particular high or low vital-sign value for it to be relevant within the diurnal cycle. Thus, we tested several different methodologies of encoding this diurnal 'range' using percentiles of the data.

To compute the variability using percentiles, we chose a lower percentile (α) of the data as the minimum, and an upper percentile ($100 - \alpha$) of the data as the maximum. Thus, the diurnal variability was computed as follows:

$$V_t = R_{100-\alpha}(X_{(t-24):t}) - R_{\alpha}(X_{(t-24):t}) \tag{5.1}$$

where V_t is the diurnal variability value computed at time t. R_α is the αth percentile rank of the data, and $X_{(t-24):t}$ is the vital-sign data over the preceding 24 hours.

We computed the diurnal variability in 5-minute intervals. To do so, we iterated over the continuous data in 5-minute steps, extracting the previous 24 hours of data at each step and computing the variability. If there were less than 12 hours of continuous data within the previous 24-hour time-window at a given time-step, no variability value was computed. We computed diurnal vital-sign variability for α values of 0 (equivalent to maximum - minimum), 1, 5 and 10. We also computed the diurnal variability as the standard deviation over the 24-hour window.

We computed the diurnal variability for heart rate, respiratory rate and SpO_2 from continuous data for all patients, using the methods described above. We also computed the diurnal variability from the manual vital-sign data, using the methods described in section 3.2.2 (the maximum - minimum value in the previous 24 hours). We then grouped the data by whether the patient was "normal" or "abnormal".

5.3.4 Evaluating variability methods

We evaluated the variability estimation methods with metrics based on probability distributions, cumulative distribution functions (CDFs) and trajectories (see below). We showed probability distributions with heart rate, CDFs with respiratory rate and

trajectories with SpO_2 (figures 5.9 to 5.11); one vital sign per metric type.

Probability distribution function based metrics

To evaluate the variability estimation methods, we first computed the probability distributions of diurnal vital-sign variability values in each group, for each method. We used data from the 24 hours before discharge from the ward (for "normal" patients) and before the adverse event (for "abnormal" patients) to show differences between data during physiological deterioration and stability. To compute the distribution for a given vital sign, variability computation method, and patient group, we used non-parametric Gaussian kernel density estimation (Bishop [2006]). The density estimate was computed as follows.

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x - X_i}{\sigma}\right)^2 \right] \tag{5.2}$$

where $f(x)$ is the density estimate at the data point x, n is the size of the dataset X (of variability values), and σ is the kernel bandwidth.

To compute the density estimates, we computed the kernel bandwidth σ with *unbiased cross-validation by least squares* (Charpentier and Flachaire [2015]). We used the constrained optimisation function *minimize_scalar* from the Scipy Python package (Jones et al. [2001], based on the algorithm of Brent [2013] using golden section search with inverse parabolic interpolation) to minimize the following objective function:

$$g(\sigma) = \int f^2(x)dx - \frac{2}{k}\sum_{i=1}^{k} f_{-i}(X_i) \tag{5.3}$$

where g is the objective function for values of the kernel bandwidth σ and f is the kernel density estimation function described in equation (7.1). The second term is a k-fold cross-validation procedure, where the data are split randomly k times to produce k datasets X_i; we used $k = 10$. In each fold, f_{-i} is the density estimation function fitted to all data not in X_i.

We constrained *minimize_scalar* to search for the minimum of g for $1 < \sigma < 10$. We found $\sigma = 10$ to be an empirical upper-bound on the global minimum. We chose the lower-bound $\sigma = 1$ to ensure that the estimate did not over-fit spikes in the sample distribution at integer values, which occurred because the vital-sign recordings were rounded to the nearest integer (by the monitoring equipment or by clinical staff).

In cases where the vital-sign variability sample had a high number of low values (mostly in the "normal" group), the density estimation assigned probability mass to

negative values (since a Gaussian kernel centred at zero extends to negative values), which are impossible due to the variability calculation method. In such a case, the probability distribution over the zero-bounded domain does not integrate to one, so is not strictly a true probability distribution. However, we do not expect that this significantly changed the shape of the distribution and therefore did not qualitatively affect our analysis. We note that log-transformation can be used to bound support to positive values (as used in chapter 7), but this asymptotically reduces the probability mass to zero as the values approach zero, so does not effectively model distributions with many low values and would not be an improvement for this application.

In figure 5.9 we show example distributions for diurnal heart rate variability computed from continuous data for "normal" and "abnormal" patient groups, using the methods described in section 5.3.3. We also show the distributions of diurnal heart rate variability computed from manual data for the two groups. The range of values, the shapes of the distributions, and the amount of overlap between the distributions of the two patient groups, differed for each variability method.

We quantified the difference between the distributions of the "normal" and "abnormal" patient groups for each variability computation method using statistical distance metrics. Our goal was to identify the variability computation method that maximised the distance between the two patient groups. If a particular method of computing variability results in very different distributions of values during physiological normality compared to physiological abnormality, then the resultant variability feature could be highly discriminatory within a machine-learning EWS.

The first distance metric was the Bhattacharyya distance (Bhattacharyya [1946]), which measures the amount of overlap between probability distributions. The Bhattacharyya distance has previously been used as a measure of class separability in machine learning classification tasks (Choi and Lee [2003]), so it is an appropriate metric for evaluating the distance between "normal" and "abnormal" data in the context of developing machine learning based EWSs. The Bhattacharyya distance D_B between two continuous probability distributions p and q is computed as follows.

$$D_B(p, q) = -log\Big(\int \sqrt{p(x)q(x)}\, dx \Big) \tag{5.4}$$

We also measured the distance between the distributions of "normal" and "abnormal" data using the Jensen-Shannon distance (Endres and Schindelin [2003]). The Jensen-Shannon distance is based on the Kullback-Leibler divergence, which compares the entropy of two distributions of the same random variable, and measures the

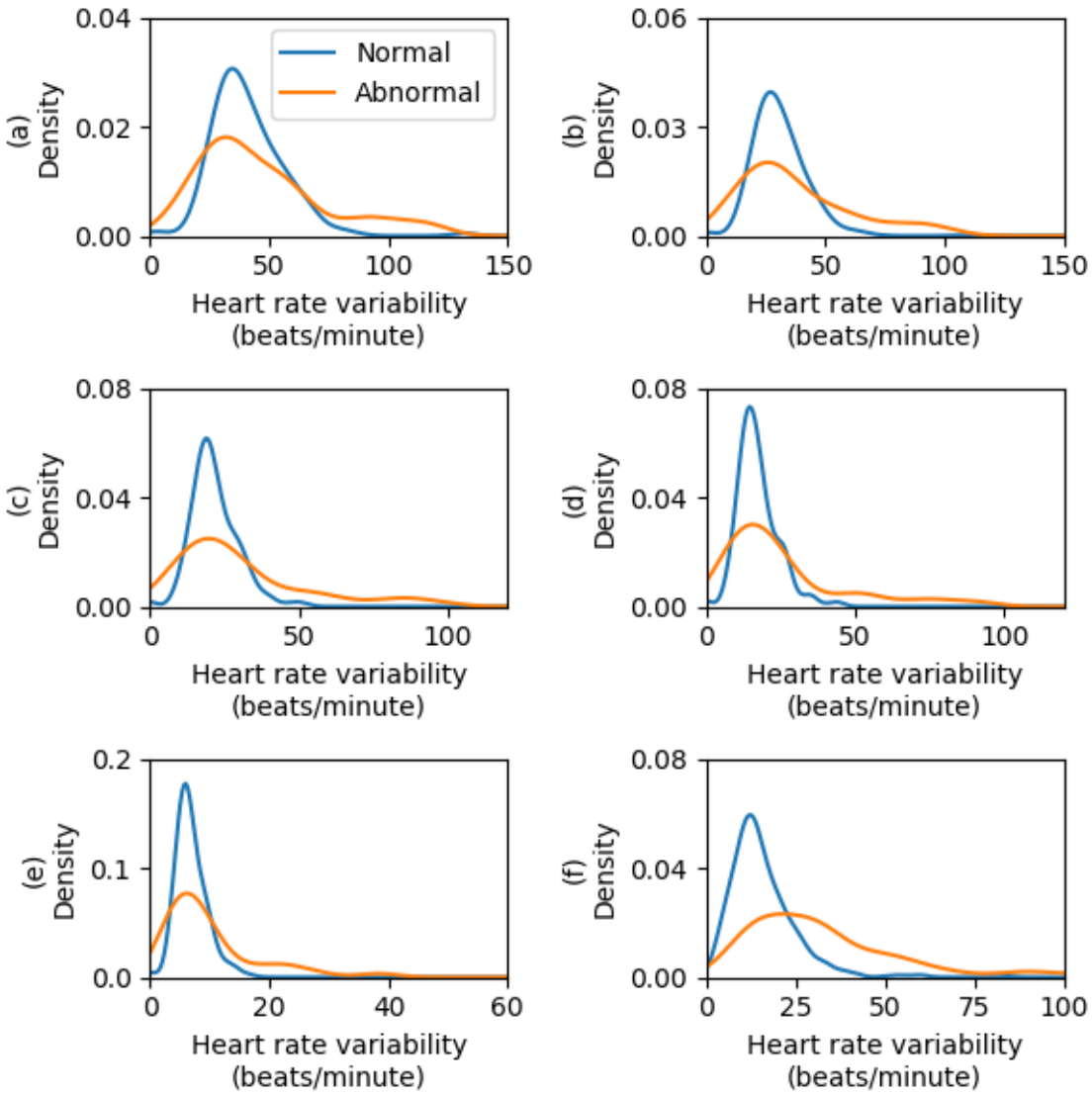


Figure 5.9: Distributions of diurnal heart rate variability for the "normal" and "abnormal" patient groups in the CALMS-2 dataset. The continuous variability was calculated from percentiles as described in section 5.3.3, using α values of 0 (a), 1 (b), 5 (c) and 10 (d), as well as using the standard deviation (e). The distributions of heart rate variability computed from manual data are also shown (f).

number of additional bits required when encoding a random variable x with a distribution $p(x)$ using an alternative distribution $q(x)$ (Kullback and Leibler [1951]). The Jensen-Shannon distance is a symmetrical and smooth version of the Kullback-Leibler divergence that can be used to measure distance, and is computed as follows.

$$D_{KL}(p||q) = \int p(x) log\left(\frac{p(x)}{q(x)}\right) dx \tag{5.5}$$

$$D_{JS}(p,q) = \sqrt{\frac{1}{2}D_{KL}(p||m) + \frac{1}{2}D_{KL}(q||m)} \tag{5.6}$$

where $D_{KL}(p||q)$ is the Kullback-Leibler divergence from $q(x)$ to $p(x)$, $D_{JS}(p,q)$ is the Jensen-Shannon distance between $q(x)$ and $p(x)$, and $m = \frac{1}{2}(p+q)$.

Cumulative distribution function based metrics

We computed statistical distance metrics based on the CDFs of the "normal" and "abnormal" patient data for each variability computation method. We computed the CDF empirically from percentiles of the data. In figure 5.10 we show example CDFs for diurnal respiratory rate variability computed from continuous data for "normal" and "abnormal" patient groups using the methods described in section 5.3.3, and respiratory rate variability computed from manual data.

To quantify the difference between the CDFs of "normal" and "abnormal" variability data we used two additional statistical distance metrics. We computed the Kolmogorov-Smirnov distance, which measures the distance between the empirical distributions of two samples, as follows (Chakravarty et al. [1967]).

$$D_{KS}(p,q) = \sup\left(|P(x) - Q(x)|\right) \tag{5.7}$$

where $D_{KS}(p,q)$ is the Kolmogorov-Smirnov distance between two probability distributions $p(x)$ and $q(x)$ with CDFs $P(x)$ and $Q(x)$, respectively, and $\sup(d)$ is the supremum of the set of distances (d).

Finally, we computed the Wasserstein distance, which is similar to the Kolmogorov-Smirnov distance but rather than measuring the supremum of the set of distances between two CDFs, it measures the integral of the distances between two CDFs (Ramdas et al. [2017]). It can be seen as the total amount of 'work' required to transform $p(x)$ into $q(x)$. The Wasserstein distance is computed as follows.

$$D_W(p,q) = \int |P(x) - Q(x)|dx \tag{5.8}$$

When evaluating the different methods of computing diurnal variability from continuous vital-sign data, we considered the distance metrics defined in equations (5.4) and (5.6) to (5.8). We also qualitatively assessed the trajectories of the average "normal" and "abnormal" vital-sign variability time-series in the 48 hours before discharge from the ward (for "normal" patients) or the adverse clinical event (for "abnormal" patients), as described below. We chose the method for each vital sign that had the largest

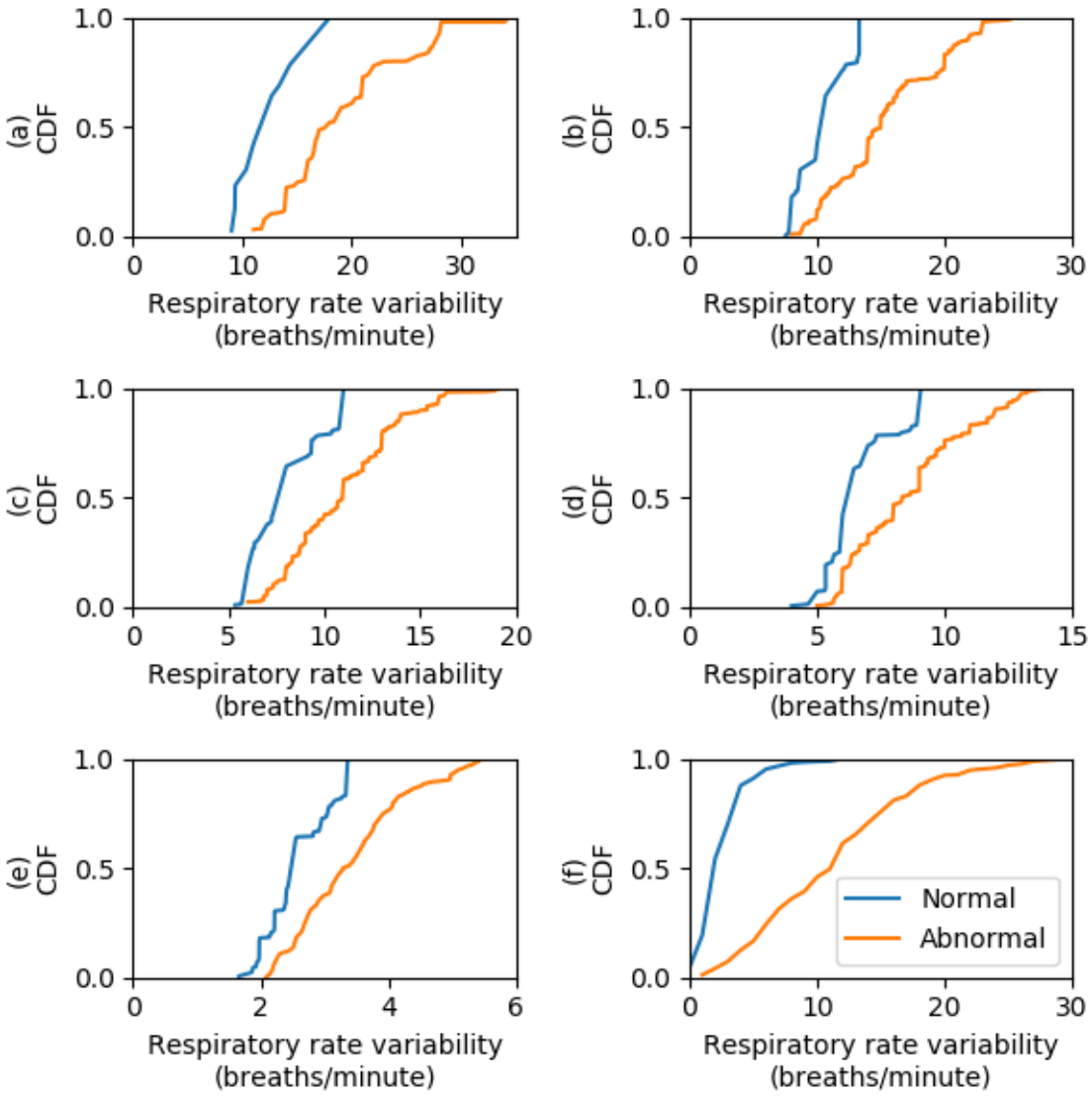


Figure 5.10: CDFs of diurnal respiratory rate variability for the "normal" and "abnormal" patient groups in the CALMS-2 dataset. The continuous variability was calculated from percentiles as described in section 5.3.3, using α values of 0 (a), 1 (b), 5 (c) and 10 (d), as well as using the standard deviation (e). The CDFs of respiratory rate variability computed from manual data are also shown (f).

values over the four distance metrics, and had the most divergence between vital-sign variability trajectories between periods of physiological stability and deterioration.

Trajectories of diurnal vital-sign variability

We computed the trajectories of diurnal vital-sign variability values for the "normal" and "abnormal" patient groups for each variability computation method. For each

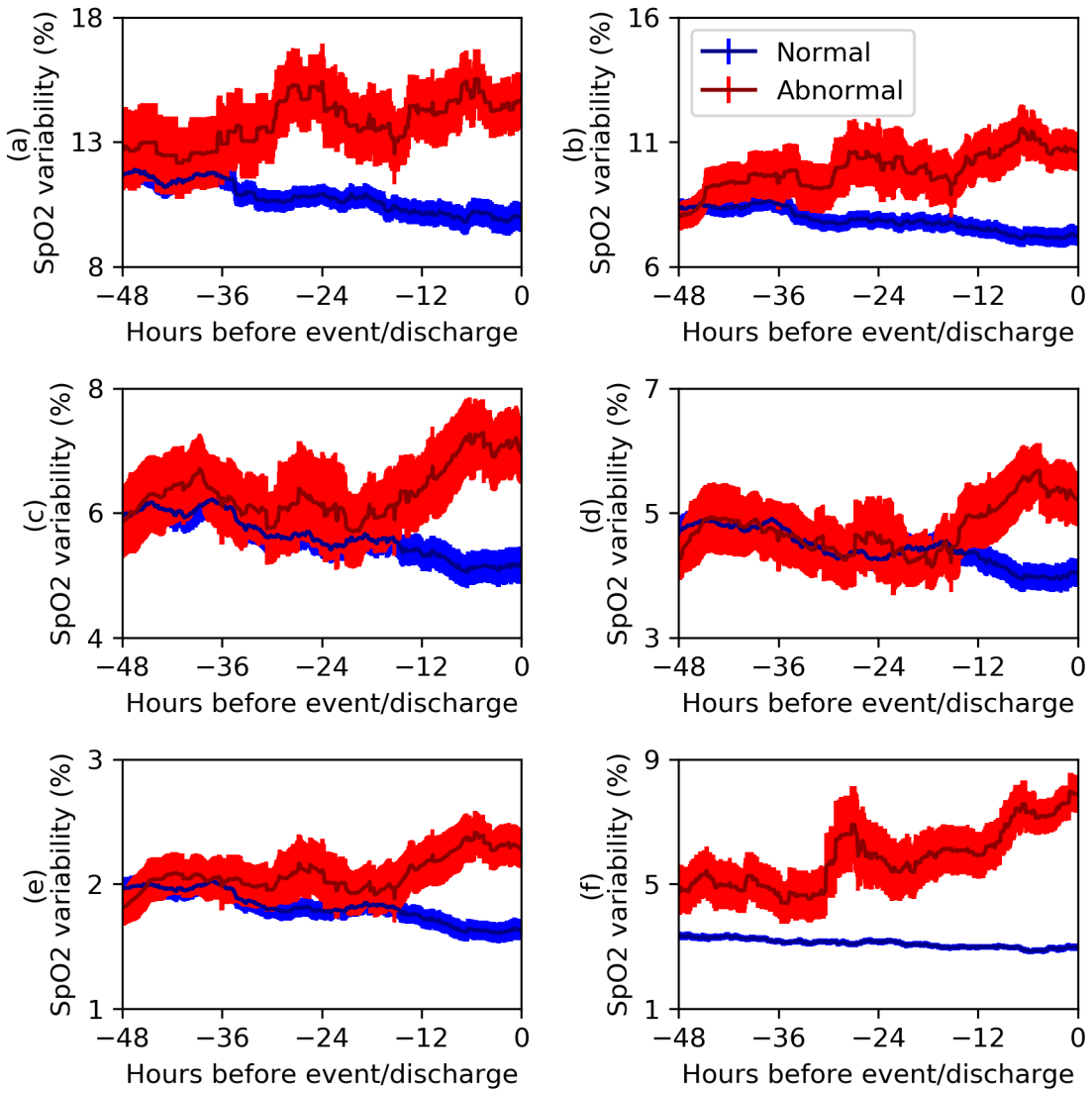


Figure 5.11: Average trajectories of diurnal SpO_2 variability in the 48 hours before discharge from the ward for the "normal" patient group and before an adverse clinical event for the "abnormal" patient group in the CALMS-2 dataset. The continuous variability was calculated from percentiles as described in section 5.3.3, using α values of 0 (a), 1 (b), 5 (c) and 10 (d), as well as using the standard deviation (e). The average trajectories of SpO_2 variability computed from manual data are also shown (f). Error bars are given for standard errors of the mean variability trajectories.

group, we created a time index of 5-minute intervals from 48 hours to 0 hours before discharge from the ward or the adverse event. We then computed the variability value at each time in the time-index for each patient and found the mean vital-sign variability at each time-point for the patient group. For the manual data, we used zero-order hold interpolation to forward-fill the time-points up to a maximum of

4 hours after the immediate previous value. If a patient did not have a vital-sign variability value at a particular time-point due to missing data, this was not included in the computation of the mean. The standard error of the mean was also computed at each time-point. For each method, the average trajectories for the "normal" and "abnormal" patient groups were then plotted with standard error bars.

In figure 5.11 we show example trajectories of diurnal SpO_2 variability computed from continuous data for "normal" and "abnormal" patient groups using the methods described in section 5.3.3. We also show the average trajectories of diurnal SpO_2 variability computed from manual data for the two groups. The shape of the trajectories of "normal" and "abnormal" patient data depended on the variability computation method, as did the gap between the trajectories and the time at which they began to diverge. For example, the "abnormal" trajectories for continuous variability in figure 5.11a ($\alpha = 0$) diverged from the "normal" trajectories 36 hours before the event, while the divergence occurred 12 hours before the event in figure 5.11d ($\alpha = 10$).

5.4 Results

Variability method selection

Table 5.1 shows the results for the analysis of statistical distances between distributions of diurnal heart rate, respiratory rate and SpO_2 variability before adverse clinical events or discharge from the ward, for different methods of computing variability. For heart rate, the method of computing diurnal variability from continuous data that best separated the distributions during physiological normality and stability was from percentiles with $\alpha = 5$ (95th - 5th percentile). For respiratory rate and SpO_2, computing variability from percentiles with $\alpha = 0$ (maximum - minimum values) provided the best separation between distributions of "normal" and "abnormal" data. The distributions of vital-sign variability computed with manual data were separated by a larger statistical distance than the distributions of variability computed with continuous data for all variability computation methods, over all distance metrics.

We assessed the average trajectories of diurnal vital-sign variability in the 48 hours before an adverse event or discharge from the ward for the "normal" and "abnormal" patients in the CALMS-2 dataset. Example trajectories for diurnal SpO_2 variability were shown in figure 5.11. We found (qualitatively) that the methods that resulted in the largest and earliest divergence in trajectories between physiological abnormality and stability were consistent with those chosen by computing statistical distance.

Table 5.1: Statistical distance metrics for different methods of computing diurnal variability for heart rate, respiratory rate and SpO_2. Diurnal variability was computed from percentiles of continuous data, with different values of α as described in section 5.3, as well as the standard deviation. We also gave values for variability computed from manual data for comparison. The distance metrics used were the Battacharyya distance (D_B), the Jensen-Shannon distance (D_B), the Kolmogorov-Smirnov distance (D_{KS}) and the Wasserstein distance (D_W), as described in section 5.3.3. The largest value for each distance metric (for continuous data) is shown in bold, for each vital sign.

		Distance metric			
Vital sign	**Method**	D_B	D_{JS}	D_{KS}	D_W
Heart rate	$\alpha = 0$	0.086	0.256	0.149	9.09
	$\alpha = 1$	0.121	0.295	0.218	9.48
	$\alpha = 5$	**0.166**	**0.333**	**0.221**	**10.3**
	$\alpha = 10$	0.164	0.325	0.214	9.49
	Standard deviation	0.157	0.319	0.201	3.48
	Manual	0.194	0.388	0.451	16.7
Respiratory rate	$\alpha = 0$	**0.293**	**0.451**	0.618	**6.94**
	$\alpha = 1$	0.268	0.439	**0.680**	4.98
	$\alpha = 5$	0.127	0.326	0.492	3.01
	$\alpha = 10$	0.133	0.326	0.445	2.03
	Standard deviation	0.066	0.234	0.492	0.827
	Manual	0.581	0.577	0.753	8.66
SpO_2	$\alpha = 0$	**0.078**	0.256	**0.311**	**3.82**
	$\alpha = 1$	0.077	**0.259**	0.310	2.90
	$\alpha = 5$	0.035	0.169	0.215	1.20
	$\alpha = 10$	0.021	0.111	0.178	0.708
	Standard deviation	0.075	0.111	0.250	0.433
	Manual	0.219	0.381	0.466	3.93

Patient case studies

We used the chosen methods for each vital sign to compute diurnal variability values from continuous data. We plotted the diurnal variability time-series from continuous and manual data using the same six example patients as in section 5.2. Figures 5.12 to 5.14 show diurnal variability time-series from continuous data for heart rate, respiratory rate and SpO_2, as well as from manual data for the 5 main vital signs, for three "normal" patients 48 hours before discharge from the ward. Similarly, diurnal vital-sign variability time-series are shown in figures 5.12 to 5.14 for three "abnormal" patients 48 hours before an adverse event. Note that the plots may have different y-axis scales due to different vital-sign variability ranges for each patient.

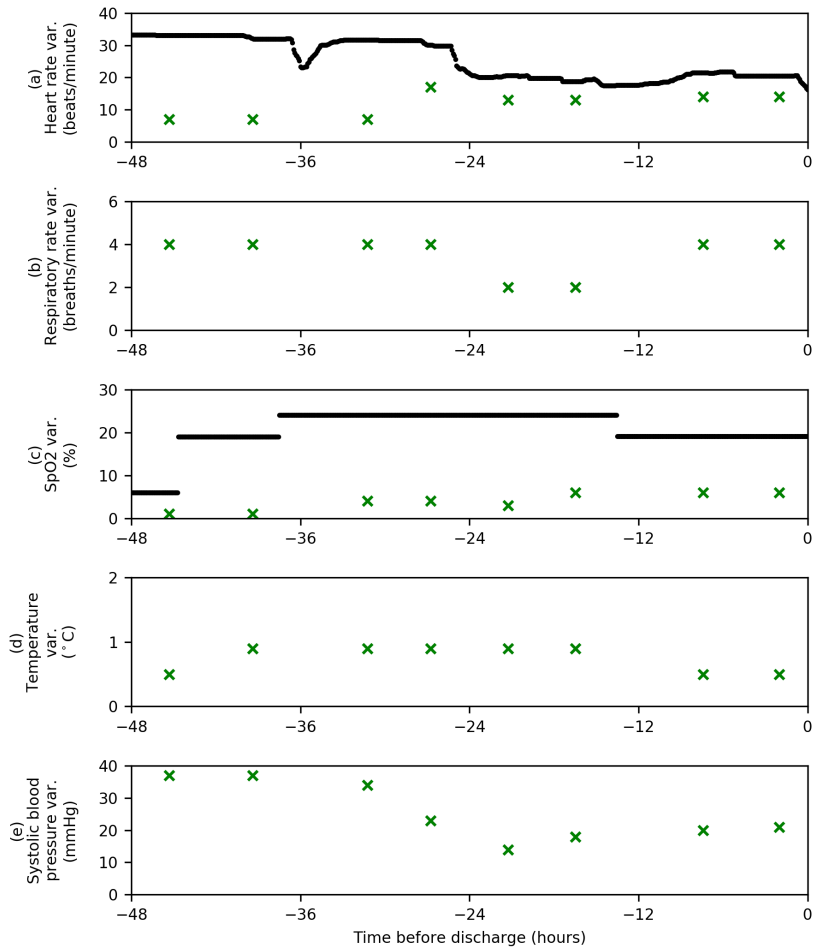


Figure 5.12: Diurnal vital-sign variability time-series for CALMS-2 patient 447 in the final 48 hours before discharge from the ward. Variability values were computed from continuous data (in black), and manual observations (in green). Here 'var.' refers to variability.

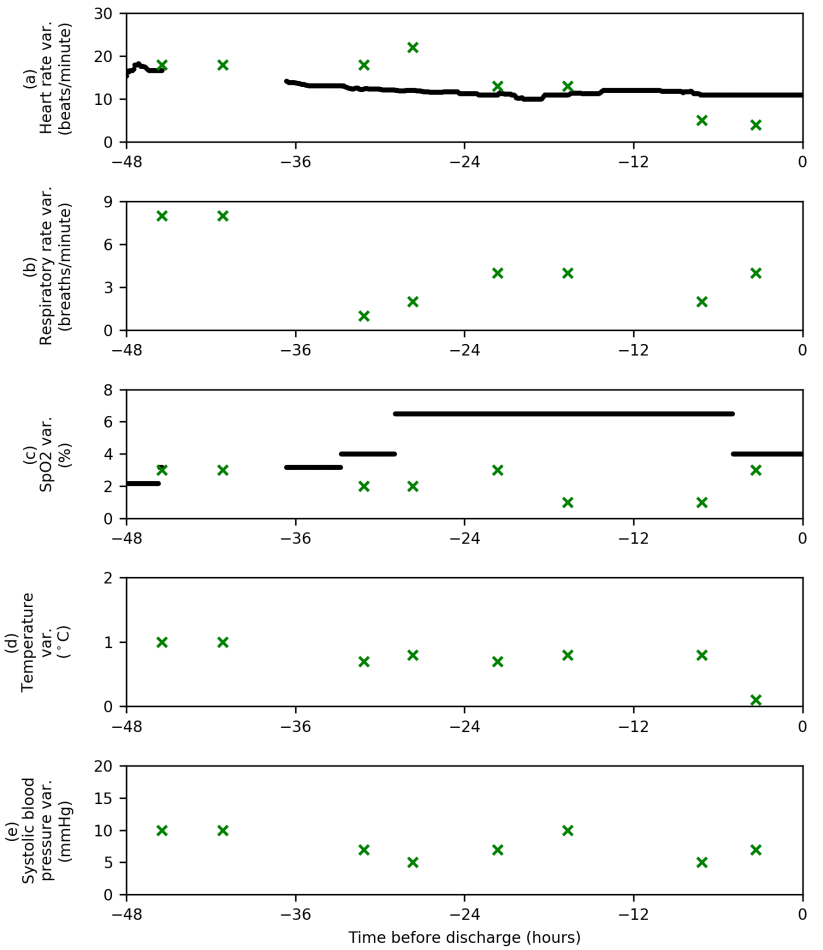


Figure 5.13: Diurnal vital-sign variability time-series for CALMS-2 patient 462 in the final 48 hours before discharge from the ward. Variability values were computed from continuous data (in black), and manual observations (in green). Here 'var.' refers to variability.

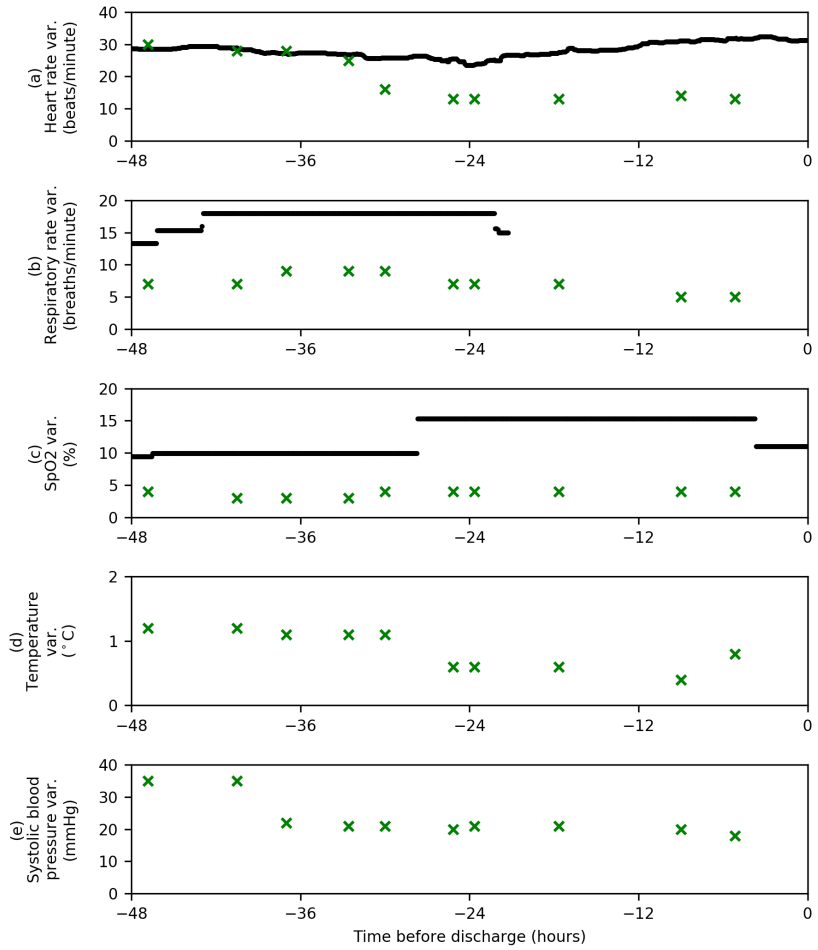


Figure 5.14: Diurnal vital-sign variability time-series for CALMS-2 patient 471 in the final 48 hours before discharge from the ward. Variability values were computed from continuous data (in black), and manual observations (in green). Here 'var.' refers to variability.

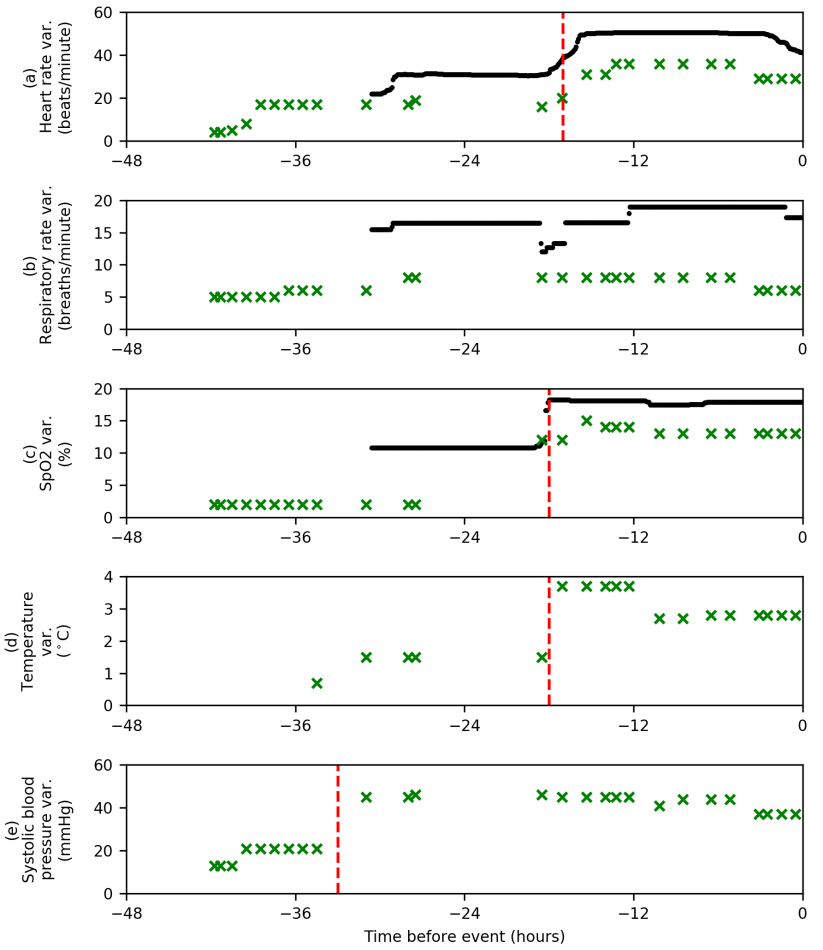


Figure 5.15: Diurnal vital-sign variability time-series for CALMS-2 patient 348 in the final 48 hours before an adverse event. Variability values were computed from continuous data (in black), and manual observations (in green). Red dashed lines show important changes in variability. Here 'var.' refers to variability.

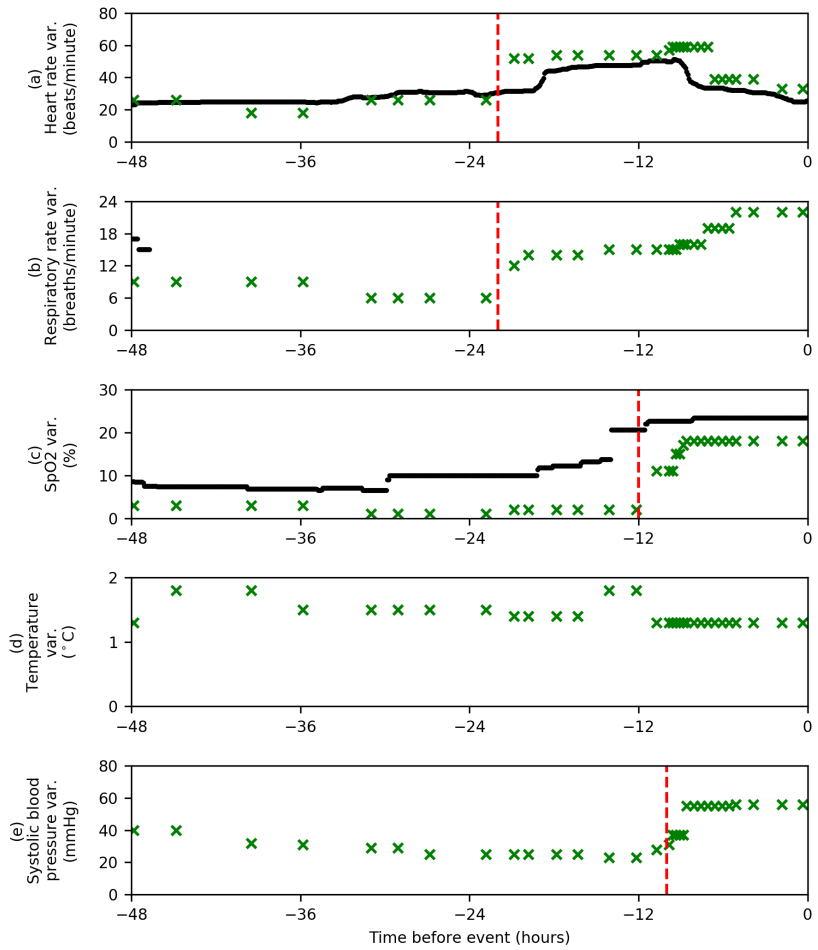


Figure 5.16: Diurnal vital-sign variability time-series for CALMS-2 patient 362 in the final 48 hours before an adverse event. Variability values were computed from continuous data (in black), and manual observations (in green). Red dashed lines show important changes in variability. Here 'var.' refers to variability.

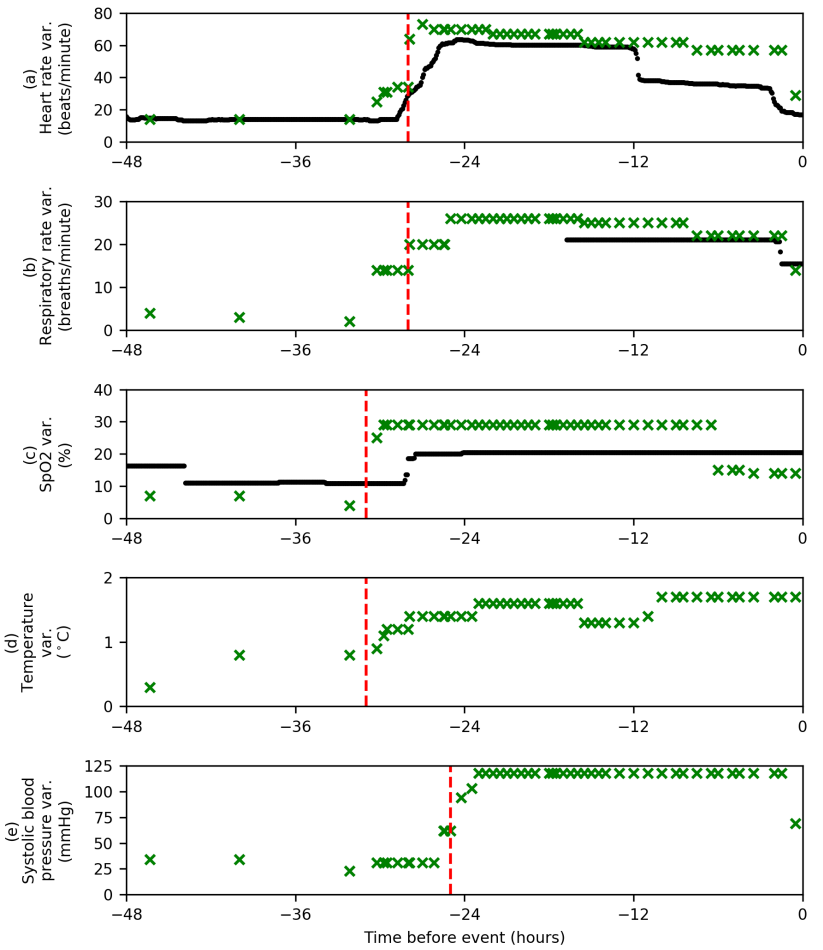


Figure 5.17: Diurnal vital-sign variability time-series for CALMS-2 patient 464 in the final 48 hours before an adverse event. Variability values were computed from continuous data (in black), and manual observations (in green). Red dashed lines show important changes in variability. Here 'var.' refers to variability.

5.5 Discussion

The continuous variability method selection analysis, with results given in table 5.1, showed that continuous variability computed from percentiles of the data could separate distributions of "normal" and "abnormal" patient data. Furthermore, the percentile levels used (quantified by the value of α) could determine the statistical distance between distributions of "normal" and "abnormal" patient data, which we used as a proxy measure of the discriminative power of the variability method for each vital sign. For example, the statistical distance for SpO_2 variability computed with $\alpha = 0$ was approximately double the distance for SpO_2 variability computed with $\alpha = 10$. The method that resulted in the largest statistical distance for each vital sign was relatively consistent for each metric, with $\alpha = 5$ having the largest distance for all metrics for heart rate, and $\alpha = 0$ having the largest distance for 3 of the 4 metrics for respiratory rate and SpO_2.

While there may be other methods of computing diurnal variability from continuous vital-sign data that have greater discriminative power than the method of using percentiles of the data, we note that using the standard deviation over the previous 24 hours produced low values of the statistical distance between "normal" and "abnormal" data, and thus did not improve discriminatory performance. We also note that the best value of α for computing variability will depend on the amount of measurement noise in the continuous data, so will be affected by the measurement equipment used, and practices such as median filtering. Thus, this analysis should be interpreted as a demonstration of the effect that the variability computation method can have on the ability to discriminate between physiological abnormality and stability, rather than a definitive selection of variability computation methods for each vital sign.

For all three vital signs, the distributions of "abnormal" and "normal" patient data were separated by greater statistical distance for variability computed from manual data than from continuous data. This was the case regardless of the measurement method, and was true for all four statistical distance metrics. Additionally, the trajectories of "abnormal" and "normal" patient data were more separated for variability computed from manual data than from continuous data. It is likely that this result was related to the high amounts of missing continuous vital-sign data in the CALMS-2 dataset, as many periods of physiological deterioration (which were more rare in the CALMS-2 dataset than physiological stability) would not have been captured. This can be seen in figure 5.16, which only has approximately 1 hour of continuous respiratory rate data. The wearable devices in the CALMS-2 study were often unreliable, as described

in chapter 2, so the data were not sufficient to allow a fair comparison between diurnal variability computed from continuous and manual data.

There may also be an effect whereby the manual vital-sign values are influenced by the nurses' perceptions of the patient's physiology. We showed in chapter 4 that there is no "smoothing effect" in the dataset, so we can rule out any clinical bias towards normality affecting the manual observations. However, we also showed that there is poor agreement between manual and continuous data, so a different clinical effect may exist. For example, Clifton et al. [2015] found that clinical observations predicted future changes in the vital-sign values in the CALMS-2 dataset (phase I), concluding that this showed that clinical staff encode their subjective assessment of the patient in their vital-sign measurements. Hence, the manual vital-sign values collected before adverse clinical events may have been influenced by clinical perceptions, thus affecting the difference in distributions of "normal" and "abnormal" patient data.

Figures 5.12 to 5.14 show vital-sign variability time-series for "normal" patients before discharge from the ward. The diurnal variability values are mostly low for both continuous and manual data, which is expected as the patients had stable physiology immediately before discharge from the ward. The SpO_2 variability values computed from continuous data in figure 5.12 were higher than the values computed from manual data. This was likely due to transient low SpO_2 values, for example shown approximately 37 hours before ward discharge in figure 5.1. While 5-minute median filtering can be used to remove transient values, it is possible that a misaligned pulse oximeter finger probe or other source of measurement artefact could cause erroneous values to persist in the time-series. This is a limitation of computing variability from the maximum and minimum values in the previous 24 hours, as a single persistent erroneous value could influence the diurnal variability values for the next 24 hours. More robust signal processing methods of removing artefact are needed before such an approach can be used to reliably compute diurnal variability from continuous data. As described in section 5.2, there is limited additional benefit of continuous monitoring during physiological stability as there are no significant changes in physiology that can be better captured with continuous data.

Figures 5.12 to 5.14 show vital-sign variability time-series for "abnormal" patients before an adverse clinical event. For all three patients there were clear increases in the values of diurnal variability, often for all five vital signs, that preceded the adverse events – usually by about 24 hours (see red dashed lines). The changes in variability values computed from continuous data mostly tracked the variability values computed from manual data, for example the heart rate variability values from continuous and

manual data in figures 5.13 and 5.16 were similar. In figure 5.12, the change in SpO_2 variability preceding the adverse event is captured several hours earlier with the continuous data, suggesting that this could help detect physiological deterioration earlier. There were also periods when the continuous variability time-series lagged the changes in the manual variability time-series, for example for SpO_2 variability in figure 5.14. There were missing continuous respiratory rate data, and thus major changes in respiratory rate variability in figures 5.16 and 5.17 were not captured. With more reliable continuous monitoring methods, important changes in vital-sign variability could be measured to detect physiological deterioration.

The practice of increasing manual observation frequency during periods of physiological deterioration can be seen in figures 5.13 and 5.14, with some measurements taken several times per hour. During these times, the differences between continuous and manual data are less obvious, as both monitoring methods record changes in physiology in a timely manner. This is the case for vital-sign and vital-sign variability time-series, as physiology is adequately sampled over the diurnal cycle with both methods. However, increasing the frequency of manual measurements relies on clinical staff detecting changes in physiology and subjectively assessing a need for increased monitoring, while this level of surveillance is always available with continuous monitoring (if the equipment operates correctly). Furthermore, frequent manual vital-sign observations lead to high clinical workloads, which could be partially avoided with the use of continuous monitoring.

A limitation of non-invasive continuous vital-sign monitoring is that only heart rate and SpO_2 can be reliably measured. Bedside monitoring can provide continuous respiratory rate measurements (using the ECG electrodes), but this cannot be used for long-term measurement in post-operative patients. While algorithms have been developed to extract respiratory rate from PPG waveforms (Charlton et al. [2017]; Meredith et al. [2012]), further studies are needed to validate this approach for long-term monitoring of ward patients. To the best of our knowledge, there are no reliable long-term ambulatory methods for continuous monitoring of temperature and blood pressure. It is clear from figures 5.15 to 5.17 that changes in respiratory rate and systolic blood pressure can precede adverse clinical events. Changes in temperature variability can also be seen in figure 5.15 more than 12 hours before the adverse event. Furthermore, respiratory rate variability and systolic blood pressure variability were in the top four most important features for predicting adverse clinical events in the analysis in chapter 3. The development of non-invasive measurement methods for all 5 vital signs could be beneficial, as it would allow such physiological changes to be

measured in real-time without the immediate presence of clinical staff. We explore the use of non-contact methods of measuring heart rate and respiratory rate in an acute post-operative population in chapter 6.

We assessed different methods of computing diurnal variability from continuous vital-sign data to measure how well each method could discriminate between "normal" and "abnormal" patient data. Our approach, based on distributions of vital-sign variability values, was chosen to be suitable for the CALMS-2 continuous data records, which had a high amount of missing data. Ideally we would use a similar evaluation approach as in chapter 3 to directly measure the discriminative performance of continuous vital-sign variability features within an EWS framework. Further work is needed to collect reliable long-term recordings of continuous vital-sign data from post-operative patients, which could be used to explore other methods of computing vital-sign variability, and to evaluate vital-sign variability features for EWSs.

5.6 Conclusions

Diurnal variability computed from continuous vital-sign data can be used to distinguish between "normal" and "abnormal" patient data, as evidenced by the separation in the distributions, CDFs and trajectories of variability values that were evaluated as described in section 5.3.3. Vital-sign variability time-series show increased values preceding adverse clinical events for all five vital signs, often 24 hours in advance. Diurnal variability estimates computed from continuous data could track important physiological changes currently measured manually, without the additional clinical burden of increasing the frequency of recording manual observations. We conclude that diurnal variability features computed from continuous vital-sign data may be useful for detecting physiological deterioration.

Our analysis was limited by having to use the continuous vital-sign dataset collected during the CALMS-2 study, which had high amounts of missing data. This reflects the difficulty of performing ambulatory continuous vital-sign monitoring in post-operative care, as wearable sensors need to be reliable and acceptable to the patient for long time-periods. As described in section 1.4, very few studies have successfully demonstrated long-term continuous monitoring in post-operative wards. Improved continuous vital-sign monitoring methods need to be developed for ambulatory settings, as the benefits of using continuous monitoring systems are limited when vital-sign values, variability values and EWSs cannot reliably be computed. Long-term continuous vital-sign records should be collected in future studies to further explore the potential of diurnal

variability as a feature in EWSs for continuous vital-sign data.

We investigate the derivation of vital signs from video camera data in the remaining chapters of this thesis. In chapter 6 we focus on changes in physiology that are relevant in the context of diurnal variability and EWS computation. In chapter 7 we investigate spatial changes in respiration during physiological deterioration.

Chapter 6

Non-contact vital-sign monitoring

Previous chapters have explored the effectiveness of manual and continuous vital-sign measurements in detecting changes in physiology that precede adverse clinical events. Continuous vital-sign monitoring can sample vital-sign variability that occurs between clinical observations, thus building a richer description of physiological changes than manual methods would allow. Variability features improve the performance of EWS models (chapter 3), and further improvements may be achieved using continuous variability signals, such as those described in chapter 5. Continuous vital-sign monitoring can measure changes in physiology as they occur, so may allow more timely detection of physiological deterioration.

In practice, it is difficult to perform continuous monitoring reliably in post-operative wards as patients find it difficult to tolerate the long-term use of ECG electrodes and other sensors. An alternative approach may be to use non-contact methods that can continuously monitor patients without affecting their comfort or ability to ambulate. As introduced in chapter 1, non-contact estimates of vital-sign values can be derived from the output of a video camera. Most previously reported studies have been conducted in controlled laboratory settings with healthy subjects. Clinical environments present additional environmental challenges such as variable or flickering lighting conditions, human challenges such as patient movement or the presence of nurses or family members in the video, and physiological challenges such as compromised cardiovascular functionality. While previous clinical studies have been conducted, for example on pre-term infants in the NICU (Villarroel et al. [2014]) and patients undergoing haemodialysis treatment(Tarassenko et al. [2014]), further studies are needed to develop and test video-based monitoring in different hospital environments.

In this chapter, we present a clinical study designed to investigate the use of extended video-derived non-contact vital-sign monitoring in patients recovering from major surgery during elective post-operative admission to the ICU. We describe the motivation and objectives of the study, as well as the equipment set-up, clinical protocol, and data collection. We then describe methods for deriving non-contact

estimates of respiratory rate and heart rate from video camera signals, based on the analysis of data from the clinical study. While none of the patients in this study experienced adverse clinical events (as defined in previous chapters), we focus on changes in vital signs that are clinically relevant in the context of vital-sign variability and the calculation of EWSs. We present case studies of three patients who experienced such physiological changes and demonstrate the ability of non-contact monitoring to track these changes. The analyses presented in this chapter are additional to those required to fulfil the study objectives (described in Section 6.1), which were mostly to do with the feasibility and acceptance of the technology in an acute care setting.

6.1 The Non-Contact Vital-sign Monitoring study

The Non-Contact Vital-sign Monitoring (NVSM) study was conducted over a period of 10 months, from May 2016 to February 2017, in the Churchill Hospital ICU (Oxford, UK). The primary aim of the study was to establish the proportion of time that data of sufficient quality to derive vital signs could be acquired during non-contact monitoring in an acute clinical setting. Patients were to be monitored continuously for an extended period of up to 48 hours, during which time many different clinical scenarios could feasibly occur (e.g. different lighting, positions, physiological conditions, human interactions). This would allow the exploration of the use of non-contact technology in an acute care setting from a technical perspective (e.g. the data collected) and from a non-technical perspective (e.g. patient and staff interactions). A secondary goal of the study was to understand the human factors that may influence the acceptance of non-contact vital-sign monitoring in acute care. The final goal of the study was to collect a research dataset for the development of algorithms for extracting vital signs from video camera data in a clinical setting.

The NVSM study was designed in collaboration with research nursing staff and clinical staff at the Kadoorie Centre for Critical Care Research and Education (hereby referred to as the Kadoorie Centre). Patient ethics, data security and clinical safety were carefully considered in the design of the study because of the sensitive nature of video data and the acute environment of the ICU. The clinical set-up and protocol were designed to facilitate the collection of video camera data, reference vital-sign data from bedside monitors, and manual vital-sign observations from clinical staff.

Patient recruitment and ethics

Eligible patients were recruited from the pre-operative assessment clinics servicing the following specialities: maxillo-facial surgery, gastrointestinal surgery, hepatobiliary surgery, renal transplant, pancreatic transplant, urology, and gynaecology. Patients were initially screened by members of the clinical care team against inclusion/exclusion criteria, and eligible patients were then approached for informed consent by members of the research team. The following exclusion criteria were used:

- Patients under the age of 18.
- Patients whose anatomy precludes the use of the required monitoring.
- Patients who were judged to lack capacity at the time of interest, for example due to an illness, severe learning disability, intoxication caused by a drug, or otherwise.
- Patients unable to understand written English and for whom no translator could be found.

A convenience sample of 6-10 patients was sought for the pilot phase of the study, from whom data would be collected to tune the monitoring algorithms. An additional 16-20 patients were sought for the main part of the study, in which the optimised versions of the monitoring algorithms would be assessed. However, with only one set of camera equipment available, and limited research staff to perform recruitment, fewer patients were monitored than originally anticipated during the design of the study. The research team recruited 24 elective surgical patients. Nine of the recruited patients were unavailable for monitoring due to cancelled or delayed surgery, admission to the Churchill Overnight Recovery Unit (CORU) or admission to a bed-space that the monitoring equipment could not access. Thus, monitoring was undertaken for 15 patients.

The study was designed to be non-disruptive to care. Equipment (described in section 6.1.1) was chosen such that it would not meaningfully affect the behaviour or experience of the patient or clinical staff, and would be easily removable during an emergency. Vital-sign estimates from the non-contact monitoring software were computed retrospectively and were not visible during the study. Thus, non-contact vital-sign estimates did not influence clinical decisions. Ethical approval for this study was granted by the Wales Research Ethics Committee (REC 16/WA/0024).

Clinical Protocol

Video recording was commenced as soon as it was practical to do so upon the patient's arrival on the ICU, which typically occurred in the evening. Patients were monitored for the duration of their stay in the ICU, up to a maximum monitoring time of 48 hours. The equipment recorded data continuously for the duration of the study. While the patients or staff members could ask for the equipment to be turned off for any reason throughout the study, this did not occur.



Figure 6.1: Position of the equipment trolley at the end of the patient's bed.

The monitoring equipment was mounted on a bespoke trolley system positioned at the end of the patient's bed, as shown in figure 6.1. The trolley was easily movable so it would not affect access to the patient during an emergency. As is the usual practice, patients had access to a chair beside the bed, in which they would occasionally sit. Clinical research staff would periodically check on the patient and move the camera equipment to keep the patient within the focal plane of the camera if they had moved from the bed to the chair, or vice-versa, so that data were captured for both locations. An opaque 'privacy blind', i.e. a small sheet of plastic attached to a hinged fitting, was included in the monitoring set-up. This could be positioned to obscure the view of the patient from the camera focal plane, and patients would occasionally ask for the blind to be used to prevent the camera from recording dignity-compromising situations (e.g. routine hygiene).

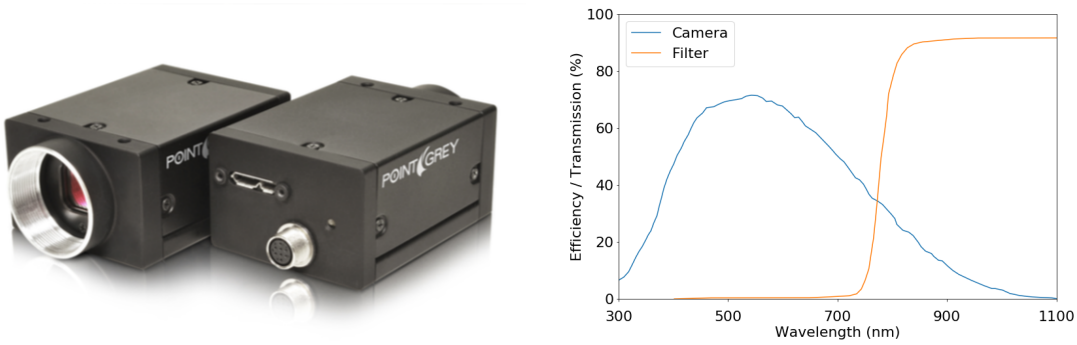
6.1.1 Equipment

Non-contact vital-sign monitoring equipment

Patients in the NVSM study were monitored with a 3.2 megapixel monochromatic camera (Point Grey Research Grasshopper3 GS3-U3-32S4M-C). This camera is shown in figure 6.2a. A 1.5 megapixel lens (non-polarised) with a 12.5 mm fixed focal length and an aperture range of f/1.4-16 was attached (Goyo Optical GMHR412514MCN), as well as an infrared-pass filter (Schneider Kreuznach 098). The response curves for the camera and the infrared-pass filter are shown in figure 6.2b. Raw uncompressed video data with 8 bits-per-pixel resolution were recorded with a frame size of 1024x768 pixels and a sampling rate of 100 Hz. An 850nm infrared illuminator was used to provide consistent illumination in a manner that did not influence the visible lighting conditions of the ward (Metaphase FL201-IRN-24 850 nm). The infrared illuminator, strobed at 50Hz to illuminate alternate video frames, was powered by a Metaphase ULC-2 LED controller connected to a Gardasoft CC320 sequence controller. A thermal video camera (FLIR Systems A65) was also used to monitor patients, with a frame rate of 9 frames/second, a precision of 0.1°C, an infrared resolution of 640 x 512 pixels, and a 14-bit temperature output. Note that the thermal camera was used for temperature measurements in another part of the study, and thus these measurements were not analysed for this thesis.

The sampling rate of 100 Hz was chosen to allow the exploration of algorithms for deriving non-contact estimates of pulse transit time (PTT), which is the time taken for a blood pressure wave to travel between two points along an arterial path (Daly [2016]). PTT has been proposed as a surrogate measurement for blood pressure (Geddes et al. [1981]). PTT estimation was not conducted in this thesis but it may be a future research direction that can be explored with the NVSM dataset (as discussed in chapter 8).

The use of strobing (of the infrared illumination) was originally chosen to support the derivation of non-contact estimates of SpO_2. As described in Section 1.1.1, traditional SpO_2 measurement methods from contact pulse oximeters derive estimates using two separate light sources of different wavelengths. In the NVSM study it was intended to facilitate research into non-contact estimation of SpO_2 by using strobing to illuminate the environment with infrared light sources of different wavelengths in alternate frames. However, it was not possible to use a second wavelength illuminator (at 750 nm) without it giving a faint red glow and hence causing confusion or worry for the patients during their recovery.



(a) (b)

Figure 6.2: Technical specifications of the monochromatic camera equipment used in the NVSM study. (a) Point Grey Research Grasshopper3 GS3-U3-32S4M-C Camera (source: FLIR Systems, Inc.). (b) Sensor response curve for the monochromatic camera (source: FLIR Systems, Inc.) and the infrared filter (source: Jos. Schneider Optische Werke GmbH).

The monochromatic video camera, the thermal video camera, and the infrared illuminator were mounted on top of the trolley (figure 6.3). The trolley set-up included an 8 TB Hewlett-Packard (HP) microserver, as well as a computer keyboard, mouse and screen. The microserver was used to record and store all data from the monochromatic and thermal cameras. The computer equipment was set up with proprietary software installed on the microserver, allowing research staff to initialise the recording software and to check periodically that the camera equipment was correctly recording data.

Reference monitoring equipment

Traditional vital-sign monitoring was provided in this study as per the clinical standard of care in the ICU, using a Philips Intellivue patient monitoring system (Philips Healthcare Intellivue MP70). Patients were fitted with ECG electrodes, which recorded raw ECG waveforms and impedance pneumography waveforms. From these, the Philips monitor computed estimates of heart rate and respiratory rate. Patients also wore a pulse oximetry finger probe (Philips M1941A), which recorded the PPG waveform and computed estimates of SpO_2 and heart rate. Blood pressure was measured using oscillometry, or, for patients of sufficiently high acuity, an invasive ABP line was fitted.

In addition to continuous monitoring, manual vital-sign measurements of heart rate, respiratory rate, SpO_2, blood pressure and temperature were periodically collected from patients, typically with 1-hour intervals as per the standard of care. Clinical research staff took additional measurements of respiratory rate and temperature on

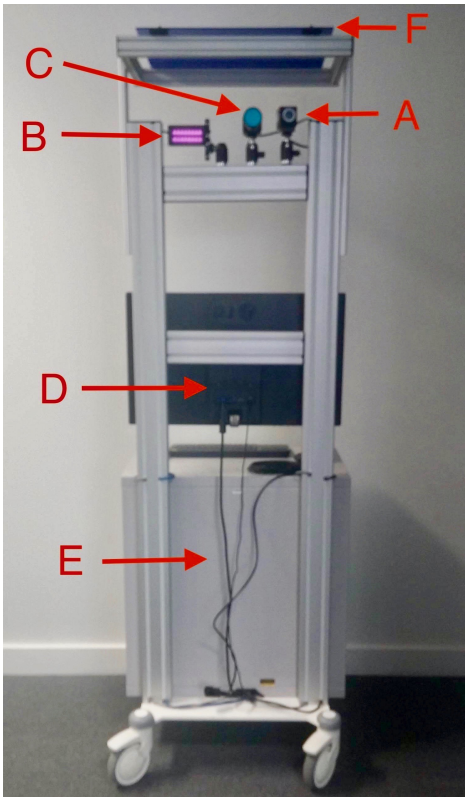


Figure 6.3: Set-up of the trolley containing the non-contact vital-sign monitoring equipment. The components are shown in their respective positions – A) optical camera; B) infrared illuminator; C) thermal camera; D) computer equipment; E) Microserver (in cabinet); F) privacy blind (folded up).

an ad-hoc basis up to twice daily. For respiratory rate, clinical research staff were instructed to count the number of breaths in a 60-second time period to keep the reference manual measurement as consistent as possible. Temperature readings were taken from both ears, with two different tympanic thermometers (Covidien PLC Genius 2).

6.1.2 Data Collection

Data from the Philips patient monitor were transferred to the microserver via the ixTrend software system (ixellence GmbH). Raw waveform data for ECG (250 Hz), PPG (125 Hz), impedance pneumography (62.5 Hz) and ABP (125 Hz) were recorded on the microserver, as well as the second-by-second measures of heart rate from ECG and PPG, respiratory rate from impedance pneumography, SpO_2 from pulse oximetry, and systolic and diastolic blood pressure from ABP. The microserver was replaced every 12 hours with an identical microserver, due to the large storage requirements of

the video data (approximately 6.7 GB per minute for the digital video data).

Summary measures of heart rate, respiratory rate and SpO_2, as well as oscillometric readings of blood pressure and tympanic temperature readings, were saved to the Carevue clinical information system (Philips Healthcare, Shabot [1997]) used by the Churchill Hospital ICU. Selected clinical information was documented by clinical research staff on paper-based clinical research forms (CRFs), including patient demographics, admission category, recording start and end times, Fitzpatrick score[1], Apache II score[2] and free-text clinical notes. The respiratory rate and temperature readings measured by the clinical research staff were also documented on the CRFs.

Security

The microserver used a RAID 5 configuration for data protection and fault tolerance. The microserver was locked in the trolley cabinet during monitoring, and was then transported from the ICU to the Kadoorie Centre. Within the Kadoorie Centre, the video data and the Philips monitor data were transferred from the microservers to a secure study server situated behind two locked doors. The CRF data, which were documented on paper forms in the ICU, were also transported to the Kadoorie Centre where they were double-entered into an electronic format with differences reconciled by two researchers. The vital-sign data collected manually were extracted from the Carevue clinical information system by a Kadoorie Centre researcher. The CRF data and the Carevue data were then transferred from Kadoorie Centre computers to the secure study server.

Privacy

Throughout the data collection process, only non-identifiable patient data were accessible to research team members who were not clinically authorised. This was facilitated by replacing the patient's name and hospital ID with a study ID, which was generated sequentially based on the order in which patients were monitored. The format of the study ID was 'NVSMxx' where 'xx' was a two-digit number beginning from 01. The study ID was written on the CRF forms and was used for identifying the vital-sign data collected during the study. The hospital ID corresponding to each study ID was

[1]The Fitzpatrick Scale is a numerical skin classification system. There are six skin types from Type I (pale white, always burns, never tans) to Type VI (darkest brown, never burns, never tans).

[2]Acute Physiology And Chronic Health Evaluation II (APACHE II) is a numerical severity of disease classification system based on initial values of routine physiological measurements, age, and previous health status (Knaus et al. [1985]). It is applied within 24 hours of ICU admission and has values that range from 0 (low severity) to 71 (high severity).

kept separately from the remaining study documentation, and was only accessible by the research nurses and the senior clinical team. The hospital ID was subsequently used by the research nurses to identify and obtain the relevant vital-sign records in the Carevue clinical information system.

Before the video camera data were available for analysis, a research nurse reviewed the data using a proprietary software tool (Oxehealth Ltd, Oxford, UK) and deleted all sections of video that contained situations that compromised the patient's dignity, such as accidental partial nudity. Furthermore, segments of the video sometimes contained images of people (e.g. family members, nurses) who had not consented to being filmed in the study. These segments were removed by the research nurse, who deleted camera data outside of a bounding box drawn around the image of the patient. In instances where the patient and the non-consenting individual could not be separated in the video frame (e.g. because the person obscured part of the view of the patient), the entire frame was deleted.

Pre-processing/blurring

While the raw video data were kept on the secure study server in the Kadoorie Centre for long-term storage, it was necessary to transfer segments of video data to the Institute of Biomedical Engineering (IBME), University of Oxford, for data analysis. However, as part of the ethics agreement of the study, non-clinical staff were not approved to access the raw data outside of the Kadoorie Centre. Thus, before transferring the video data to the IBME, a blurring algorithm was applied to the data to protect the privacy of the patient.

The blurring algorithm operates on each video frame by iterating through every non-overlapping 10x10 square region of pixels, computing the sum of the pixel values in the region, and assigning this value as a 'super pixel' in the new blurred video frame. The algorithm begins at the top left corner of the image and iterates across and down the image. For frame heights and widths that are not perfectly divisible by 10 (which was the case in our data), the bottom-most and right-most regions contain fewer than 100 pixels and therefore the corresponding *super pixel* values are generally lower than those from the rest of the video frame (as fewer original pixels are included in the sum). This has no practical effect as the edge pixels do not typically contain important data. Equation (6.1) describes how the blurred video frame is computed from the original video frame.

$$P^*(w,h) = \left\{ \begin{array}{ll} \sum_{i=10w-9}^{10w} \sum_{j=10h-9}^{10h} P(i,j), & w = 1,...W^* - 1,\ h = 1,...H^* - 1 \\ \sum_{i=10w-9}^{10w} \sum_{j=10h-9}^{H} P(i,j), & w = 1,...W^* - 1,\ h = H^* \\ \sum_{i=10w-9}^{W} \sum_{j=10h-9}^{10h} P(i,j), & w = W^*,\ h = 1,...H^* - 1 \\ \sum_{i=10w-9}^{W} \sum_{j=10h-9}^{H} P(i,j), & w = W^*,\ h = H^* \end{array} \right\} \quad (6.1)$$

where P are the pixels in the original frame and P^* are the *super pixels* in the blurred frame. W is the width of the original frame and W^* is the width of the blurred frame, with $W^* = \lceil W/10 \rceil$.[1] Likewise, H is the height of the original frame and H^* is the height of the blurred frame, with $H^* = \lceil H/10 \rceil$.

The resulting blurred video frames are approximately $1/10^{\text{th}}$ of the size of the original video frame in each direction, and $1/100^{\text{th}}$ of the size overall. For example, the blurred data in the NVSM study had 103x77 *super pixels* from the original frame size of 1024x768. The blurring algorithm did not affect the use of the data for vital-sign estimation for this thesis because the summation of pixel values over a region of interest (ROI) is the first step in our vital-sign estimation algorithms. An example video frame is shown in figure 6.4a, with the face obscured to protect patient privacy, while the blurred video frame is shown in figure 6.4b.



(a) **(b)**

Figure 6.4: Example raw (a) and blurred (b) video frames. The pixels in the blurred frame are computed from the sum of 10x10 pixel blocks in the raw frame.

[1] $\lceil \rceil$ is the ceiling operator, which returns the next integer greater than or equal to the input.

6.1.3 Study Summary

Overview

Table 6.1: Demographic and monitoring details of patients in the NVSM study. Note that ID refers to the patient identifier, and GI means gastrointestinal. Day time was considered to be between 8am and 8pm for the purpose of calculating the monitoring time during day and night. The Fitzpatrick scale was used to measure the skin type.

ID	Cancer type	Gender	Age	APACHE II score	Skin type	Monitoring time (hours:minutes)		
						Day	Night	Total
01	Lower GI	Female	69	13	1	06:12	11:58	18:10
03	Maxillofacial	Male	61	17	2	17:12	12:25	29:37
05	Maxillofacial	Male	65	14	2	06:34	00:00	06:34
06	Maxillofacial	Male	79	22	2	09:38	12:00	21:38
07	Maxillofacial	Female	58	19	2	01:20	00:00	01:20
08	Lower GI	Female	68	16	2	06:27	12:00	18:27
09	Maxillofacial	Male	51	16	2	07:53	00:46	08:39
10	Maxillofacial	Female	32	7	3	12:45	12:00	24:45
13	Upper GI	Female	59	17	2	10:59	12:00	22:59
15	Gynaecological	Female	79	22	1	24:48	26:06	50:54
16	Maxillofacial	Male	58	20	2	15:39	16:14	31:53
17	Upper GI	Male	56	10	2	10:14	11:55	22:09
19	Maxillofacial	Male	51	13	2	08:09	09:15	17:24
23	Maxillofacial	Male	72	16	2	05:05	11:53	16:58
24	Maxillofacial	Male	75	21	2	07:37	01:25	09:02
		Male (%)	Mean (Standard deviation)					
All		9 (60)	62.2 (12.4)	16.2 (4.3)	2 (0.5)	10:02 (05:33)	09:59 (06:47)	20:01 (11:37)

Demographics of the 15 patients monitored in the NVSM study are given in table 6.1, as well as the details of the monitoring they received. One patient (NVSM15) received monitoring for the full 48 hours. NVSM01 had their bedside monitoring equipment removed by clinical staff for reasons unrelated to the study, and four patients (NVSM05, NVSM07, NVSM15, NSVM24) were not monitored with the camera equipment until the morning after their arrival. For patient NVSM24, this was because the camera equipment was accidentally unplugged by clinical staff as the bed-space was being set up, while the other three were delayed due to availability of research staff. Furthermore, problems with the camera monitoring software caused the video monitoring to fail for NVSM06 and NVSM07, hence only 6:34 and 1:20 (hours:minutes) of data were

recorded for these patients, respectively. Otherwise, the remaining 9 patients were monitored for the duration of their stay, but were discharged early from the ward (within 48 hours of admission).

To our knowledge, the NVSM study is the first video-based non-contact monitoring study in an acute adult care setting for extended (>24 hours) periods of time. The extended nature of this study means that clinical activities were unlikely to be influenced by the presence of the trolley system, unlike shorter studies where clinical staff may tolerate making short-term accommodations for the sake of the study. While not all patients received more than 24 hours of monitoring, this was mostly due to unforeseen events such as early discharge, and thus this last statement holds true as clinical staff would assume that the camera would remain for an extended period of time. Data were obtained during the daytime and overnight, allowing algorithms to be tested in varying ambient light conditions. Thus, the study set-up provided a good approximation to how the video camera system may be used in future deployments.

Patient and staff perceptions

Semi-structured interviews and questionnaires were used by research nurses to collect feedback from ICU staff members on their perceptions of the non-contact vital-sign monitoring technology. This feedback formed the basis of a qualitative analysis by Ede et al. [In preparation, 2018], which identified themes that emerged around the staff members' experience of the technology. A key theme of this analysis was the perceived trade-off between patient comfort and acuity of monitoring. Patients often experience discomfort wearing wired vital-sign monitoring equipment. For example, patients often become agitated during delirium causing them to remove wires or lines, and relatives may avoid contact with patients due to the perceived physical barrier of wires and probes from monitoring equipment. There was a consensus amongst staff that non-contact monitoring may improve the experience of patients and relatives by reducing discomfort and anxiety caused by monitoring equipment.

A negative aspect of non-contact monitoring identified by Ede et al. [In preparation, 2018] was the perception that the video camera could leave staff with feelings of professional exposure, as the video footage might be used to review their actions. Furthermore, troubleshooting continuous monitoring equipment is an important aspect of the role of clinical staff. The relative autonomy of non-contact equipment may lead to clinical staff lacking confidence in the technology, as it reduces their ability to confirm that monitoring is operating correctly. Finally, clinical staff perceive invasive monitoring to be more accurate and associate invasive monitoring with higher levels of

patient acuity. Thus, there may be concerns from clinical staff in terms of the ability of non-contact vital-sign monitoring to meet their expectations of care.

Research nurses also collected ad-hoc "impressions" from patients regarding their experience of non-contact vital-sign monitoring. This was not systematically collected and thus must be considered as anecdotal. Most patients had a positive experience of the system, finding that it did not affect their stay on the ward. However, NVSM15 expressed anxiety when the monitoring equipment was removed after 48 hours, thinking that the removal of the camera equipment may have been due to clinical staff not wishing to record an adverse clinical event. While non-contact vital-sign monitoring primarily uses video cameras as sensors for vital-sign estimation, this response highlights the challenges of overcoming preconceptions related to more common uses of video cameras (for capturing footage).

Overall, the feedback from staff and patients highlighted the potential positive benefits of non-contact monitoring. Patients may have improved comfort and mobility due to the removal of wires and probes, while still receiving continuous vital-sign monitoring. Future uses of the technology must consider the human factors that may influence clinical acceptance.

Analysis

The analyses reported in the following sections of this chapter describe preliminary analyses of the NVSM dataset with patient case studies during times of clinically relevant changes in physiology. The patient case studies were chosen based on clinical events of interest (such as immediately preceding an X-ray) or based on the magnitude of changes in vital-sign values. The video footage of the candidate cases were then reviewed for obvious periods of movement that could cause vital-sign changes due to artefact, and then were reviewed by a clinician to confirm whether they were indeed clinically relevant. We underwent this process to focus our analyses on vital-sign changes that would need to be tracked by a non-contact vital-sign monitoring system in order to detect physiological deterioration. Analysis of the full NVSM dataset was considered to be too time-consuming within the scope of this thesis, and thus will be conducted as future work.

6.2 Patient Case Studies

Three patient case studies are presented in this chapter. We selected five vital-sign segments – three from NVSM15, and one each from NVSM19 and NVSM23. Details

of the vital-sign time segments are given in table 6.2.

Table 6.2: Details of the five vital-sign segments chosen from patients in the NVSM study for analysis of non-contact vital-sign monitoring algorithms.

Segment	Reason chosen	Start time (minutes)	Duration (minutes)
NVSM15-1	Preceding x-ray	279	24
NVSM15-2	Tachycardia	967	30
NVSM15-3	Tachycardia	2026	30
NVSM19-1	Tachycardia	386	30
NVSM23-1	Tachycardia	107	30

NVSM15 was a 79-year old female with Fitzpatrick type 1 skin, who was recovering from gynaecological cancer surgery. NVSM15 suffered a right-sided pneumothorax, which is an abnormal collection of air in the pleural space between the lung and the chest wall. Pneumothoraces are a medical emergency requiring prompt recognition (Yarmus and Feller-Kopman [2012]). If untreated, a pneumothorax may progress to a tension pneumothorax, whereby the pleural air build-up exceeds atmospheric pressure, producing adverse effects such as collapsed lung and cardiac arrest. Diagnosis is typically made by physical examination, noting acute onset of breathlessness, chest pain, tachycardia and hypotension, as well as medical imaging such as chest X-rays. In most patients, pneumothoraces are managed by draining air from the pleural space with a small-bore catheter, although tension pneumothoraces require immediate intervention by inserting a large needle into the suspected side for decompression.

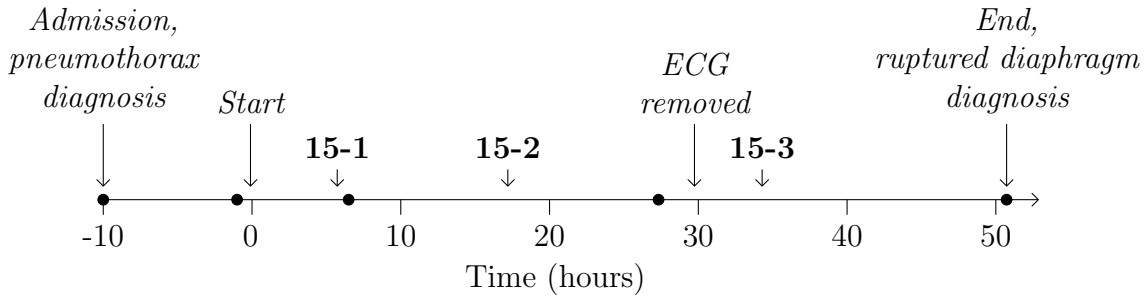


Figure 6.5: Timeline of the study segments and important events for patient NVSM15. The times of chest X-rays are shown as black dots.

The timeline of events for NVSM15 is shown in figure 6.5. NVSM15's pneumothorax was diagnosed by chest X-ray shortly after admission to the ICU. Several chest X-rays were taken during the study period, each confirming the persistent pneumothorax. An X-ray taken at the end of the recording revealed that NVSM15 had suffered a

ruptured diaphragm, causing the liver to herniate into the chest. This condition was not present in the previous X-ray taken 24 hours earlier, so the onset must have occurred during the final day of recording. From the Philips monitor recordings, it is clear that NVSM15 experienced several instances of physiological deterioration. For this reason, this chapter includes three vital-sign segments from NVSM15.

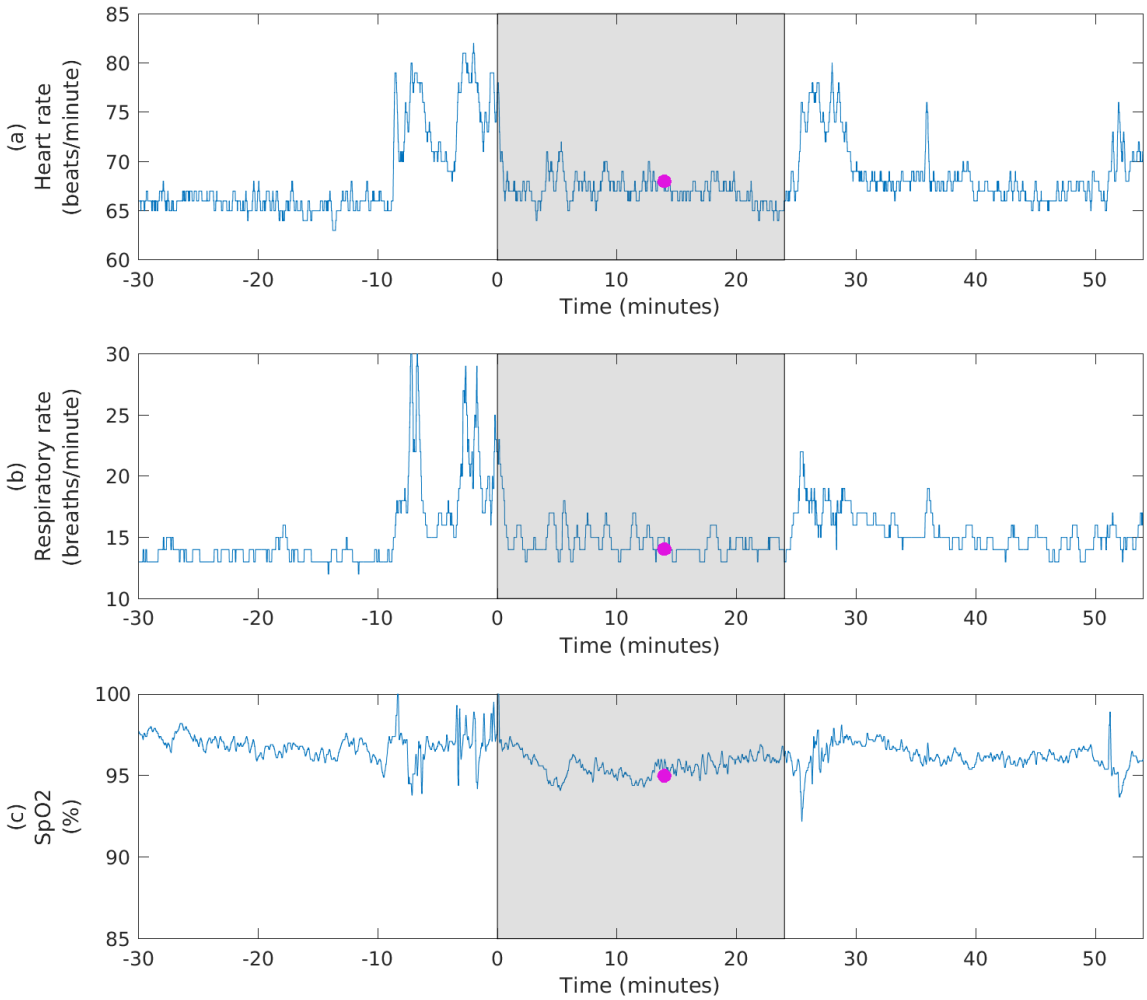


Figure 6.6: Reference heart rate (a), respiratory rate (b), and SpO_2 (c) for segment NVSM15-1 (shaded in grey). Manual vital-sign estimates are shown as pink dots and bedside monitor estimates are shown as blue lines. Data are also shown 30 minutes before and after the segment to provide physiological context. NVSM15 was a 79 year-old female patient recovering from gynaecological cancer surgery.

NVSM15-1 (figure 6.6) is a 24-minute segment during which time the patient did not move and had stable vital-sign values. Before this segment there was a period of high levels of patient movement, and immediately after this segment clinical staff visited and took a chest X-ray (this could be seen in the video footage). The high

levels of patient movement caused measurement artefact in the reference signals, which can be seen as momentary spikes in heart rate and respiratory rate before and after the segment (see minutes -10 to 0 and 24 to 30 in figure 6.6). A manual vital-sign observation was recorded during the segment, which was consistent with the bedside monitor values, thus confirming the reliability of the reference vital-sign time-series.

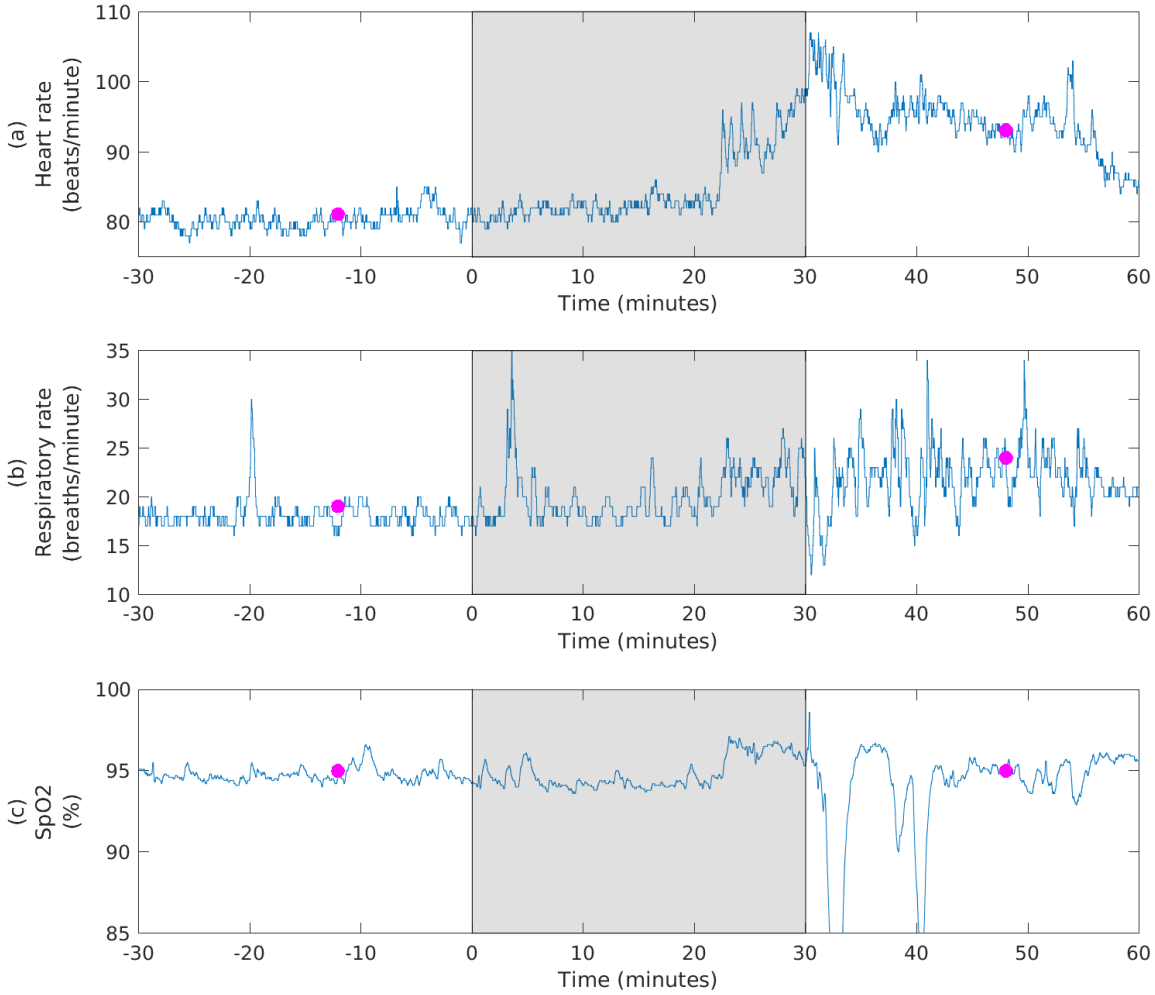


Figure 6.7: Reference heart rate (a), respiratory rate (b), and SpO_2 (c) for segment NVSM15-2 (shaded in grey). Manual vital-sign estimates are shown as pink dots and bedside monitor estimates are shown as blue lines. Data are also shown 30 minutes before and after the segment to provide physiological context. NVSM15 was a 79 year-old female patient recovering from gynaecological cancer surgery.

NVSM15-2 (figure 6.7) is a 30-minute segment during which time there was an increase in heart rate of 20 beats/minute, an increase in respiratory rate of approximately 5 breaths/minute, and stable SpO_2 around 95-97%. Preceding this segment there were stable vital-sign values as seen by the flat heart rate, respiratory

rate and SpO_2 time-series. Following the segment the heart rate and respiratory rate remained elevated for 20 minutes before returning to normal values, and there were two brief moments of measurement artefact in the SpO_2 time-series. Manual vital-sign observations were recorded 10 minutes before the start of the section, and 17 minutes after the end of the section, which were consistent with the bedside monitor values.

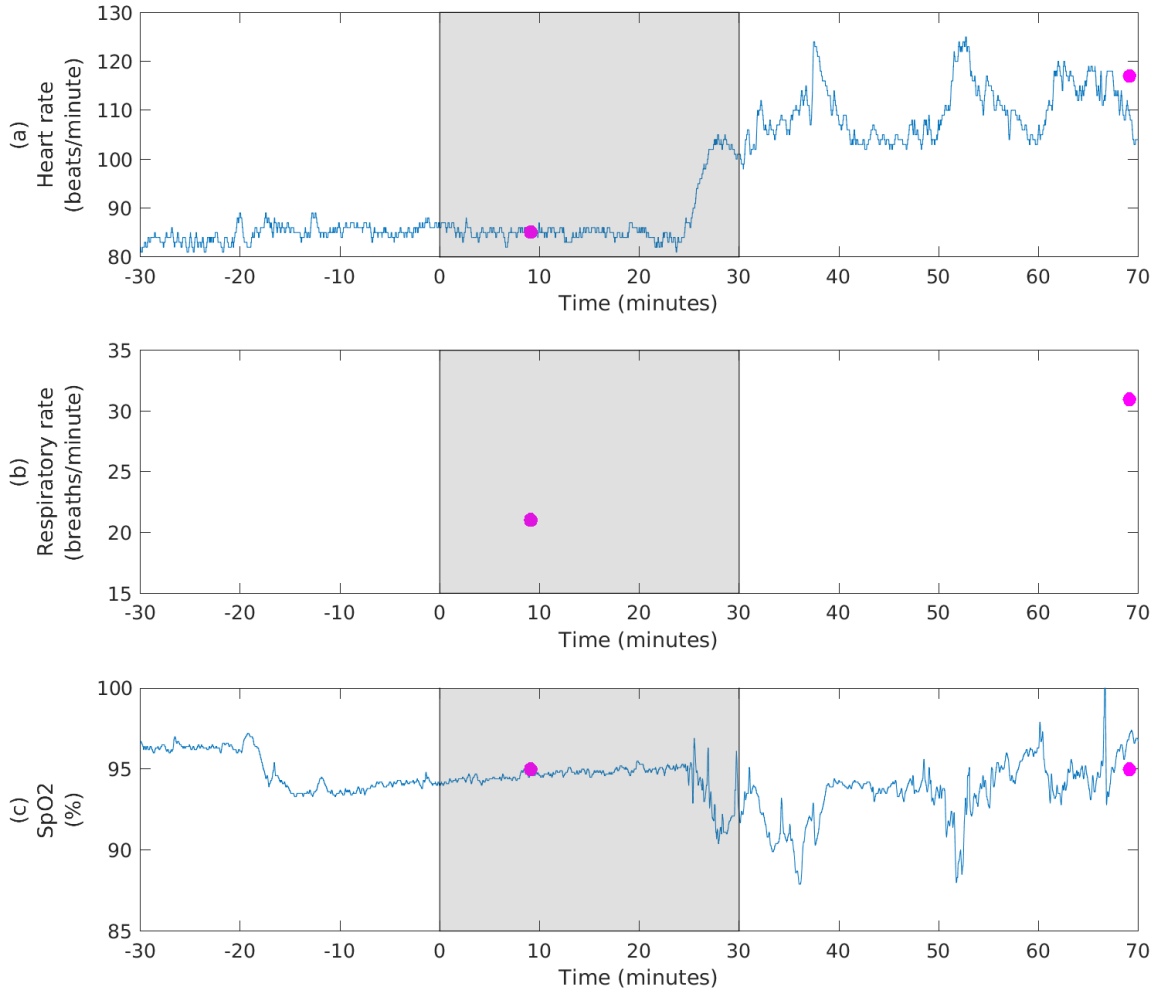


Figure 6.8: Reference heart rate (a), respiratory rate (b), and SpO_2 (c) for segment NVSM15-3 (shaded in grey). Manual vital-sign estimates are shown as pink dots and bedside monitor estimates are shown as blue lines. Data are also shown 30 minutes before and 40 minutes after the segment to provide physiological context. NVSM15 was a 79 year-old female patient recovering from gynaecological cancer surgery.

NVSM15-3 (figure 6.8) is a 30-minute segment with an increase in heart rate of 20 beats/minute. There was no reference respiratory rate signal during this time as the patient's ECG electrodes had been removed (so that there was no IP signal available), although a manual vital-sign estimate was taken during the segment. The

privacy blind was put down at the end of this segment and a 40-minute period of tachycardia followed, as well as periods of decreased SpO_2. It is noteworthy that a manual respiratory rate of 30 breaths/minute, which is highly abnormal, was recorded 40 minutes after the end of the segment.

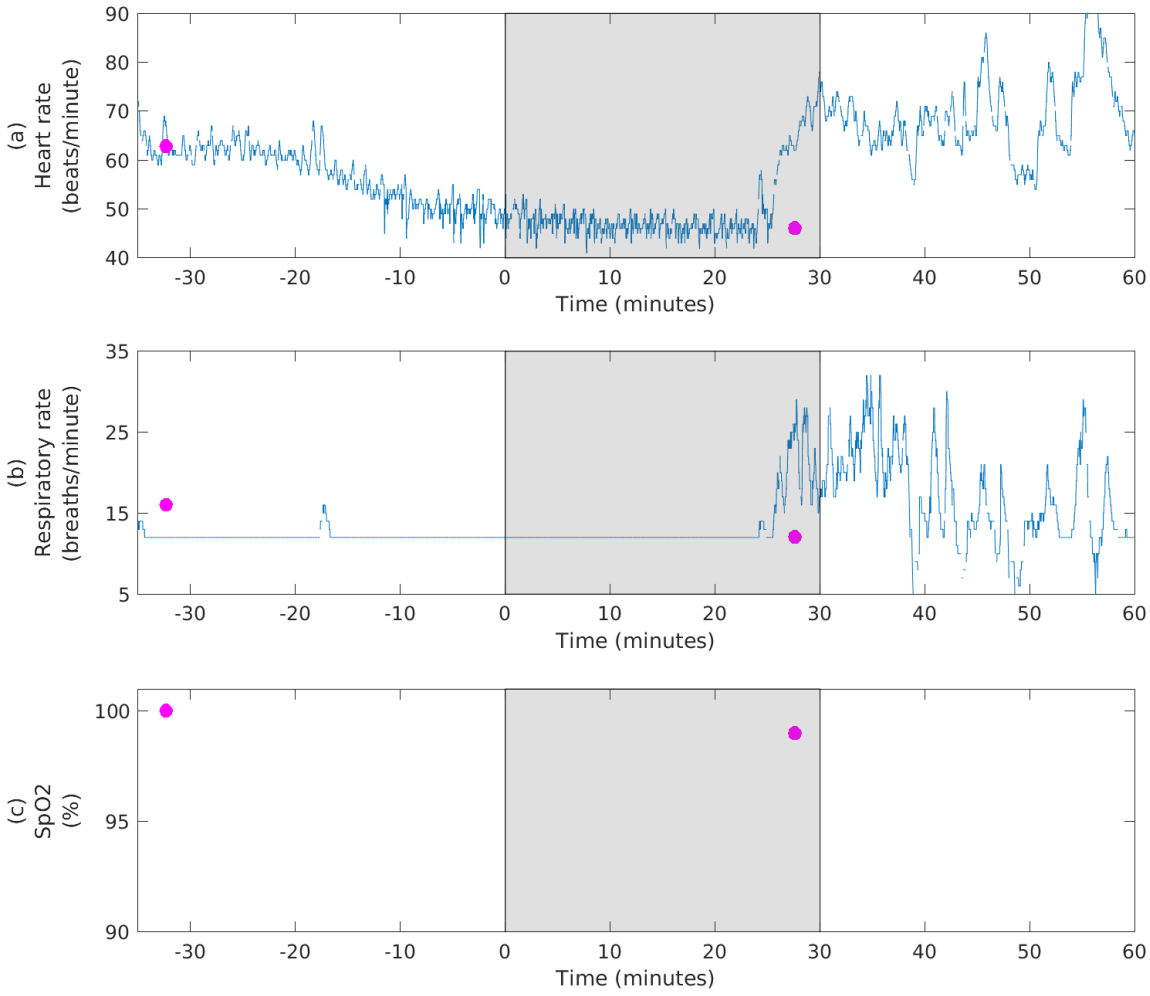


Figure 6.9: Reference heart rate (a), respiratory rate (b), and SpO_2 (c) for segment NVSM19-1 (shaded in grey). Manual vital-sign estimates are shown as pink dots and bedside monitor estimates are shown as blue lines. Data are also shown 35 minutes before and 30 minutes after the segment to provide physiological context. NVSM19 was a 51 year-old male patient recovering from maxillofacial cancer surgery.

NVSM19 was a 51-year old male with Fitzpatrick type 2 skin, who was recovering from maxillofacial cancer surgery. NVSM19-1 (figure 6.9) is a 30-minute segment that followed a period of stable respiratory rate during which time the patient was receiving breathing support through a ventilator mask, with a concomitant decrease in heart rate. The patient experienced a bradycardia during the first 25 minutes of

the segment. In the final 5-minutes of the segment there was an increase in heart rate of 30 beats/minute, an increase in respiratory rate of 10-15 breaths/minute, as well as periods of patient movement. The patient was not attached to a pulse oximeter probe so SpO_2 was not continuously measured (but was recorded manually by the nurse).

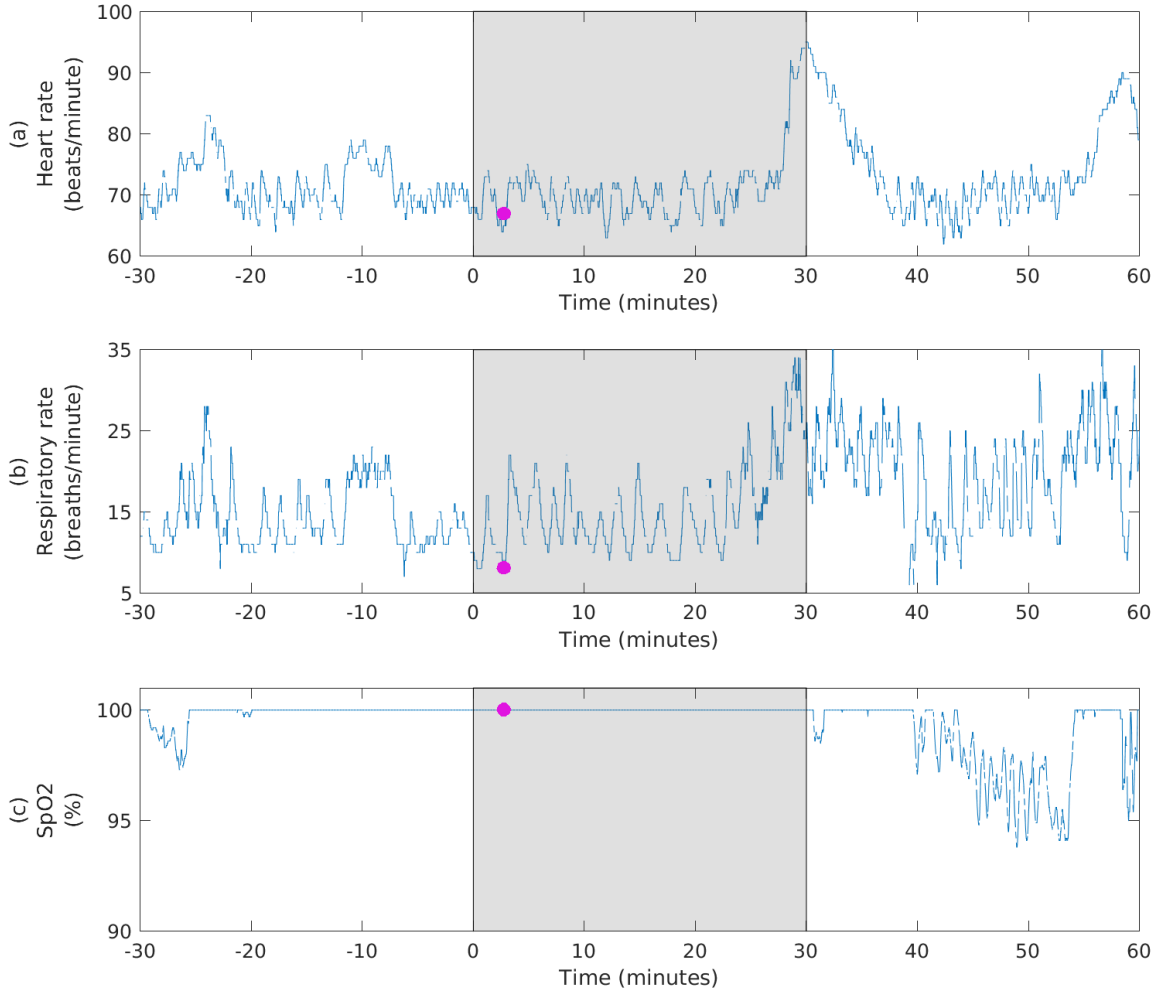


Figure 6.10: Reference heart rate (a), respiratory rate (b), and SpO_2 (c) for segment NVSM23-1 (shaded in grey). Manual vital-sign estimates are shown as pink dots and bedside monitor estimates are shown as blue lines. Data are also shown 30 minutes before and after the segment to provide physiological context. NVSM23 was a 72 year-old male patient recovering from maxillofacial cancer surgery.

NVSM23 was a 72-year old male with Fitzpatrick type 2 skin, who was recovering from maxillofacial cancer surgery. During periods of the study, NVSM23 received supplemental oxygen through a nasal canula and also had a tracheostomy[1], which can be seen in the video footage. NVSM23-1 (figure 6.10) is a 30-minute segment with

[1]A tracheostomy is an opening created at the front of the neck so a tube can be inserted into the windpipe (trachea) to assist with breathing.

an increase in heart rate of 20 beats/minute and an increase in respiratory rate of 10-15 breaths/minute in the final 5 minutes. The SpO_2 value was measured to be constant at 100% during this segment, as a result of the patient's supplemental oxygen support. The respiratory rate time-series showed an unusual oscillatory pattern with cyclical periods of low respiration followed by increased respiration, which may be related to respiratory difficulties from the maxillofacial surgery and the associated therapies of supplemental oxygen and tracheostomy. The final 5 minutes of the segment had increased levels of patient movement. There were periods of stable physiology preceding the segment, with normal heart rate values and a constant SpO_2 of 100%.

6.3 Heart rate estimation

6.3.1 Overview

An algorithm was developed to estimate heart rate from signals extracted from pixel values in each video frame. The steps of the algorithm are summarised in the flow chart presented in figure 6.11 and are described in detail in sections that follow. An overview of the algorithm is as follows:

1. A square ROI is selected on the patient's face, with size and location as described in section 6.3.2.

2. The video data are segmented into time-windows from which the algorithm computes heart rate estimates. We chose overlapping 15-second windows with a step size of 1 second. Thus, heart rate estimates are computed every second.

3. Within each time-window, a raw signal is extracted by averaging the pixel values within the ROI in each video frame. This signal is a univariate time-series (one value per video frame) called the raw ROI time-series.

4. The raw ROI time-series is band-pass filtered and the power spectral density (PSD) is computed using the Fast Fourier Transform (FFT), as described in section 6.3.3.

5. A signal quality index (SQI) is calculated based on the signal-to-noise ratio (SNR) of the peak of the PSD, and the estimate is discarded for SQI values below a threshold, as described in section 6.3.5.

6. The heart rate is computed by identifying the spectral peak with maximum power in the PSD, as described in section 6.3.4. Estimates are smoothed with median filtering and large changes between estimates are discarded.

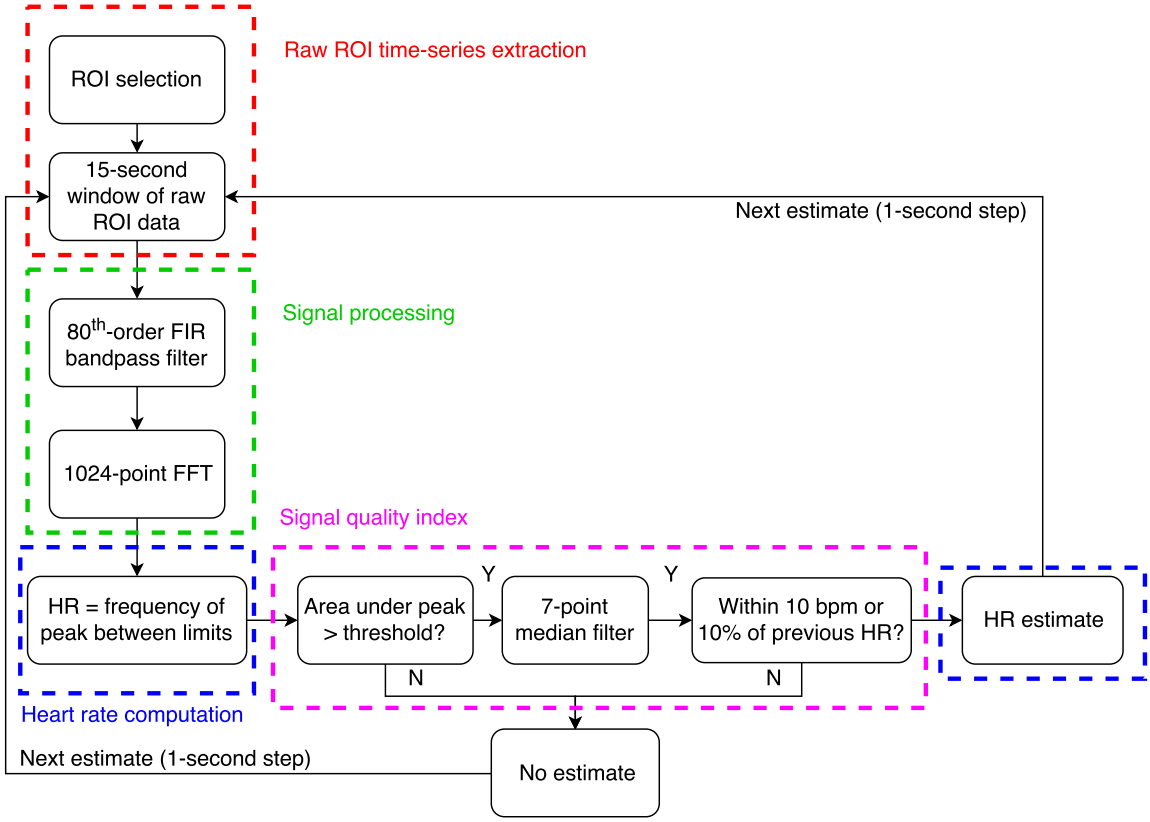


Figure 6.11: Overview of the algorithm used to estimate the heart rate in overlapping windows of raw ROI signal. Zero-phase band-pass filtering was performed with cut-off frequencies of 0.7 Hz (42 beats/minute) and 3 Hz (180 beats/minute). The heart rate was computed as the peak of the FFT power spectrum between frequency limits of 0.66 Hz (40 beats/minute) and 3 Hz (180 beats/minute).

6.3.2 ROI selection

The first step in non-contact vital-sign estimation of heart rate is the selection of an ROI on an exposed area of skin. The location of the ROI should be chosen in areas where the skin is relatively hairless, the area is well perfused, well illuminated, and ideally not affected by high levels of movement. Villarroel et al. [2014] reported that ROI time-series from large grids typically contain a stronger cardiac frequency component than ROI time-series from smaller grids. Thus, we selected candidate ROIs of different sizes from locations on the face with exposed skin, such as the cheeks,

forehead and neck. An example video frame from NVSM23-1 is shown in figure 6.12, showing the location of an 80x80 ROI on the left cheek.

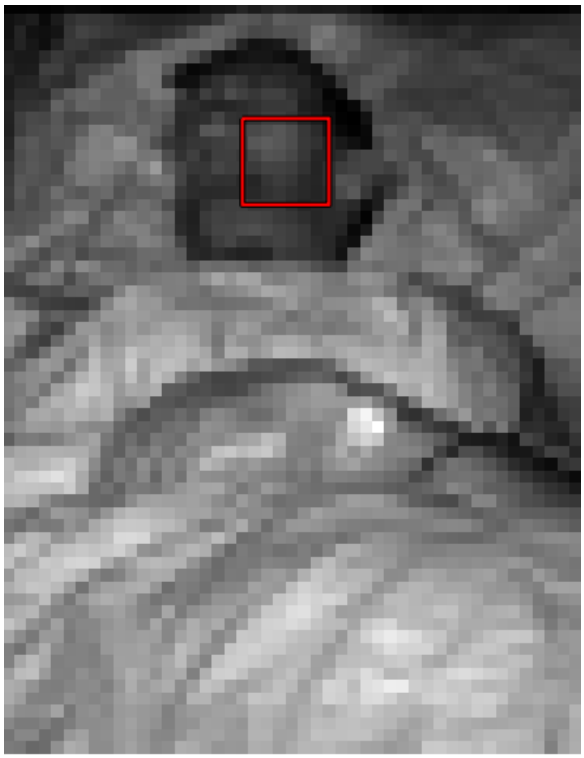


Figure 6.12: Example video frame from segment NVSM23-1, showing (in red) the position of an 80x80 ROI for heart rate estimation on the left cheek.

To extract the raw ROI time-series we took the mean of the pixel values within the ROI, doing this for every second frame to account for the strobed illumination (selecting those frames for which the illuminator is switched on). The raw ROI time-series were sampled at 50 Hz (half of the camera frame rate of 100 Hz). Example 1-minute segments of raw ROI time series from ROIs of different sizes on the face of NVSM15-1 are shown in figure 6.13. The IP signal and the PPG signal for this time segment are also shown for comparison. It can be seen that respiratory and cardiac frequency components both exist in the signal, with the relative strength of these components affected by factors such as the size of the ROI and the number of edges in the ROI. The presence of edges can cause an ROI time-series to be more influenced by pixel value changes due to motion than due to cardiac-synchronous light reflectance.

ROI time-series from different sizes and locations were examined for each time segment to establish the optimal settings for extracting heart rate. We tested ROI sizes of 70x70, 80x80, 90x90 and 100x100 pixels. ROI time-series were examined both before and after filtering (signal processing as described in section 6.3.3).

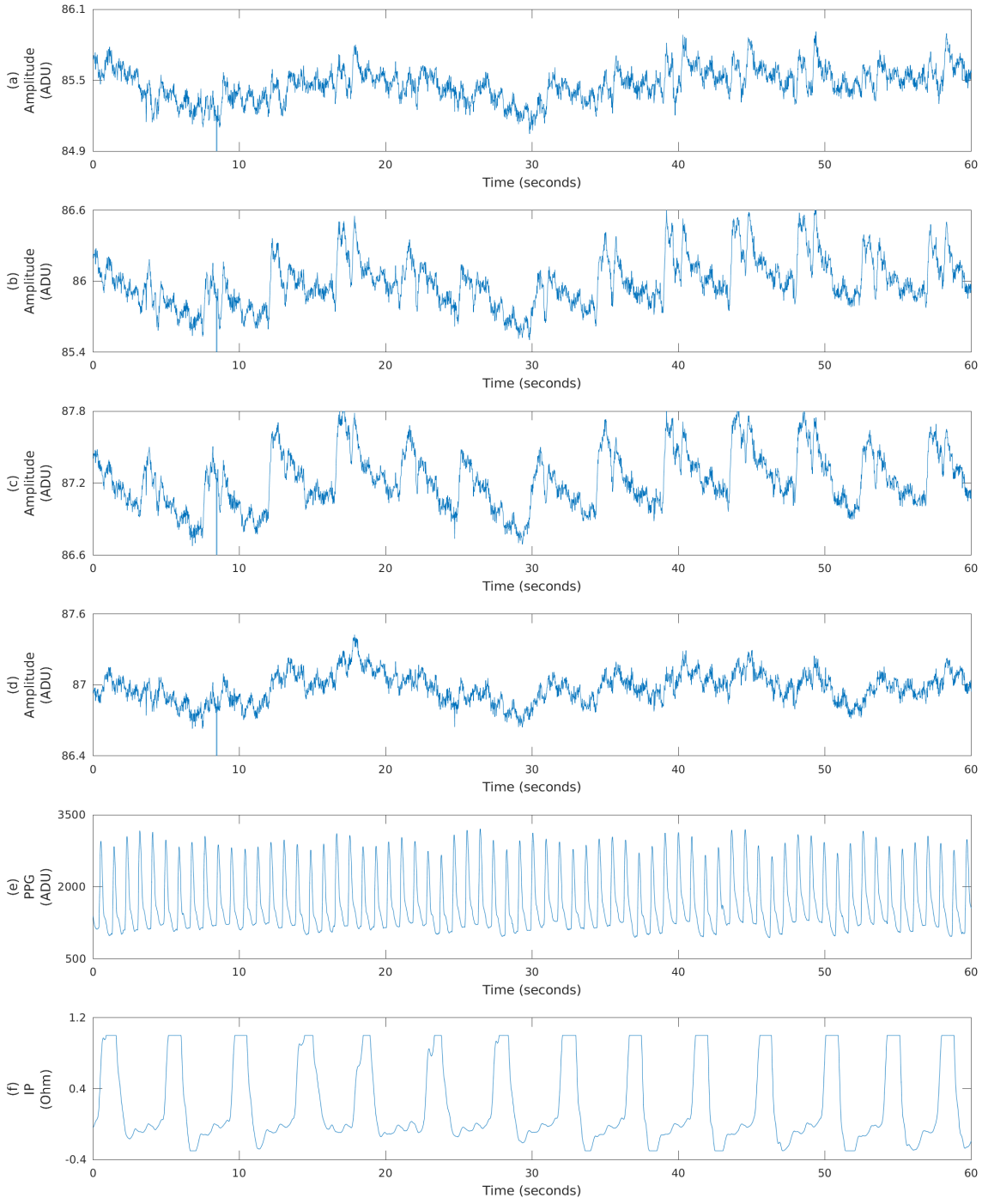


Figure 6.13: Example 1-minute segment of raw ROI time-series data during NVSM15-1, extracted from ROIs on the right cheek of size (a) 70x70 pixels, (b) 80x80 pixels, (c) 90x90 pixels and (d) 100x100 pixels. Contemporaneous segments of PPG signal from the finger pulse oximeter (e) and respiratory signal from the Impedance Pneumography electrodes (f) are also shown.

The chosen size and location for extracting ROI time-series for heart rate estimation from each segment are shown in table 6.3. The optimal specifications differed for each patient as several factors influenced the quality of the ROI time-series. Some segments had a larger area of exposed skin than others, thus allowing larger ROIs to be chosen. In addition, the distance of the patient to the camera differed between segments, which affected the relative size of the patient skin area captured by each pixel (meaning an ROI of 70x70 pixels in one segment could be equivalent to, for example, an ROI of 80x80 pixels in another segment).

Table 6.3: ROI size and location for extracting heart rate from each segment.

Segment	ROI size	ROI location
NVSM15-1	100x100	Right cheek
NVSM15-2	90x90	Forehead
NVSM15-3	70x70	Right cheek
NVSM19-1	70x70	Forehead
NVSM23-1	80x80	Left cheek

6.3.3 Signal processing

Once the raw ROI time-series were extracted for each window, band-pass filtering was used to remove non-cardiac frequency components from the signal. We used an 80th-order finite impulse response (FIR) band-pass filter with pass-band limits of 0.7 Hz (42 beats/minute) and 3 Hz (180 beats/minute) to remove frequency components outside of the expected distribution of heart rate values (mostly the component of the respiratory rate). Zero-phase filtering was implemented to avoid the introduction of delays. We used the MATLAB [2016] *filtfilt* function, which processes the signal in both the forwards and reverse directions, thus effectively doubling the filter order.

We calculated the PSD of the ROI time-series using the FFT. The FFT is an established method of computing the Discrete Time Fourier Transform (DTFT) to find a frequency-domain representation of a time-domain signal, from which the power of the frequency components of the signal can be estimated. The frequency resolution of the DTFT is given by:

$$f_r = \frac{f_s}{NFFT} \tag{6.2}$$

where f_r is the frequency resolution, f_s is the sampling frequency, and $NFFT$ is the number of samples in the signal. The FFT can be computed efficiently using

standard software in MATLAB [2016], with a faster implementation available when NFFT is a power of 2. Hence, in this work we use zero-padding to increase the length of the signal to the next power of 2 when computing the FFT.

To find the PSD for heart rate computation, the signal was first detrended and a Hamming window was applied to the signal to reduce edge effects. A 1024-point FFT PSD estimate of the signal was computed (the signal was zero-padded from 750 to 1024 points). An example window of ROI time-series before and after signal-processing is shown in figure 6.14 for NVSM15-1.

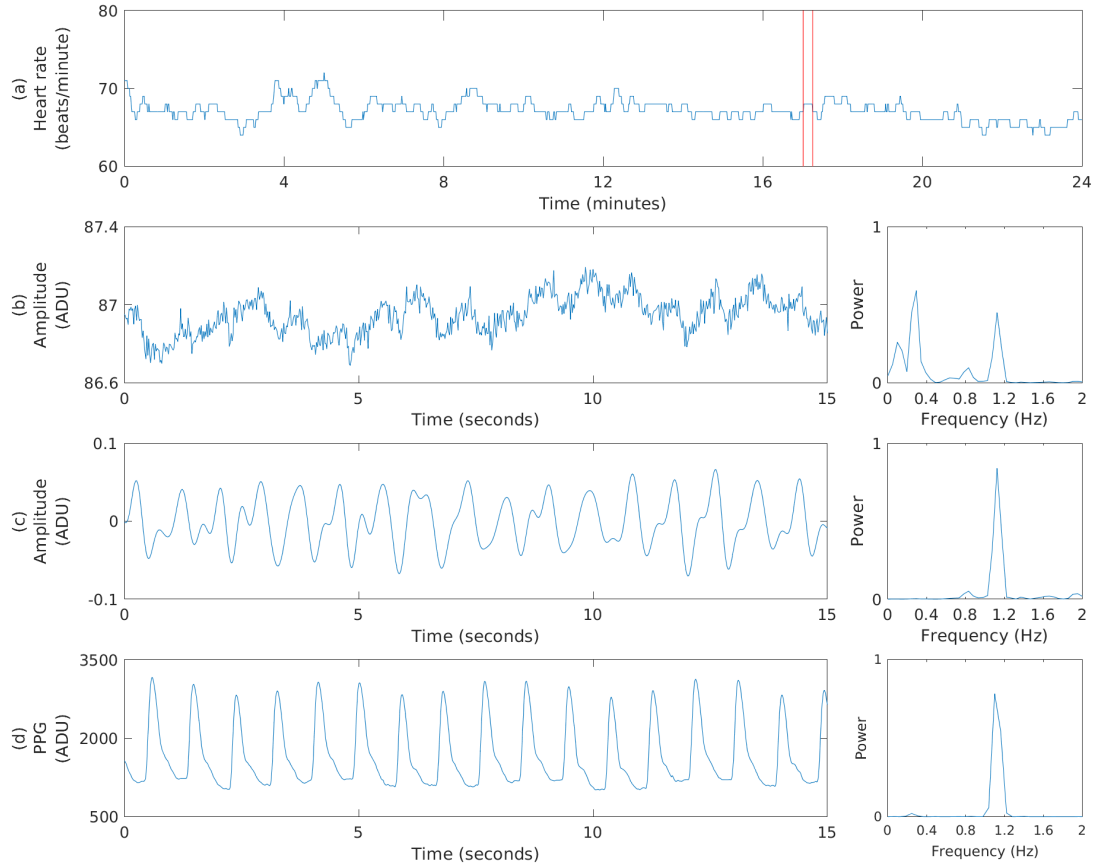


Figure 6.14: Example 15-second window of ROI time-series taken from segment NVSM15-1. The timing of the window is shown in the heart rate time-series from the Philips bedside monitor with vertical red lines (a). The raw ROI time-series is shown (b) with a 1024-point FFT power spectrum (zero-padded from $50 \times 15 = 750$ samples), as well as the processed ROI time-series after band-pass filtering (c) with a 1024-point FFT power spectrum. A contemporaneous segment of PPG signal from the finger pulse oximeter is shown (d) with a 2048-point FFT power spectrum (zero-padded from $125 \times 15 = 1875$ samples).

6.3.4 Computing heart rate estimates

The initial heart rate estimate for each time-window was computed as the frequency corresponding to the peak of the PSD within a frequency range of 0.66 to 3 Hz (40 to 180 beats/minute). These frequencies were chosen based on the expected distribution of heart rate to ensure peaks corresponding to low-frequency (e.g. due to respiration) or high-frequency noise were not chosen. Heart rate was initially computed as follows:

$$i^* = \{\, i^* | i \in \mathbb{Z} \wedge i < NFFT/2 \wedge 0.66 \leq x_i \leq 3 \wedge P(x_{i^*}) > P(x_i) \,\forall\, i \neq i^* \,\} \tag{6.3}$$

$$HR_k = 60 \times x(i^*) \tag{6.4}$$

where i^* is the index of the spectral peak and $x(i^*)$ is the frequency (in Hz) at the spectral peak. The peak is found in the first $NFFT/2$ samples as the FFT PSD is symmetrical around the Nyquist frequency ($f_s/2$). HR_k is the (initial) heart rate estimate of the kth time-window in beats per minute.

6.3.5 Signal quality index

An SQI was applied to discard estimates derived from poor-quality ROI time-series. We used an SQI that found the ratio of the total peak spectral power (area under the PSD curve) within ± 0.05 Hz (3 beats/minute) of the estimated heart rate frequency to the total spectral power between 0.66 Hz (40 beats/minute) and 3 Hz (180 beats/minute). The computation of the SQI is given in equation (6.5) as follows:

$$SQI_k = \int_{f^*-0.05}^{f^*+0.05} F(f)df \Big/ \int_{0.66}^{3} F(f)df \tag{6.5}$$

$$HR_k = \left\{ \begin{array}{ll} HR_k, & SQI_k \geq 0.10 \\ None, & SQI_k \leq 0.10 \end{array} \right\} \tag{6.6}$$

where SQI_k is the SQI of the kth time-window. f is the frequency component (in Hz), $F(x)$ is the power of the frequency component within the signal computed by the FFT, and f^* is the frequency at the spectral peak (at index i^* computed in equation (6.3)). If the SQI is less than a given threshold, the heart rate estimate is discarded. We used an SQI threshold of 10% in our experiments. Empirical tests showed that 10% provided a good trade-off between removing erroneous estimates and retaining correct estimates. An example PSD of high signal quality is shown in

figure 6.15 and an example PSD of poor signal quality is shown in figure 6.16.

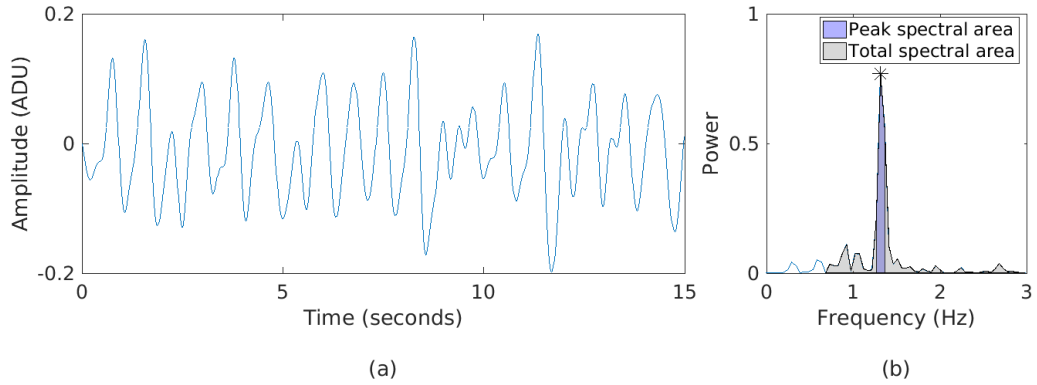


Figure 6.15: Example window of high-quality bandpass filtered ROI time-series (a) with a 1024-point FFT power spectrum (b). The peak spectral area and the total spectral area are shown, which are used to compute the SQI. The SQI value is 0.46, which is above the specified threshold of 0.10. The spectrum has a prominent peak power from which a heart rate estimate can be computed.

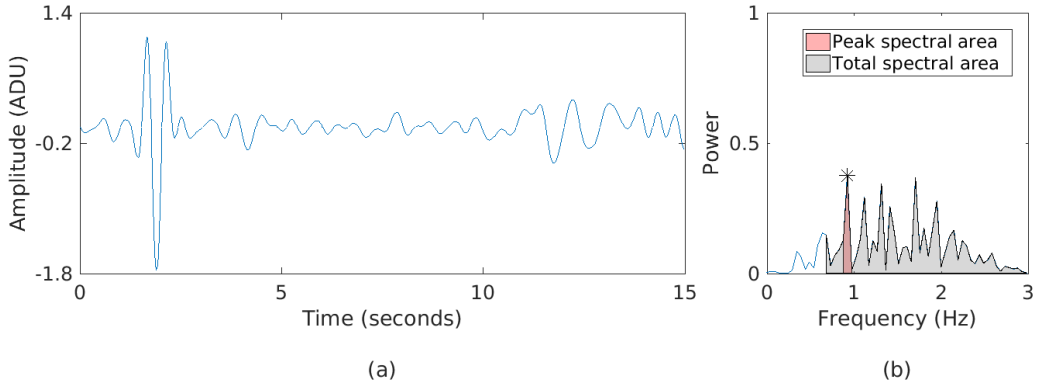


Figure 6.16: Example window of poor-quality bandpass filtered ROI time-series (a) with a 1024-point FFT power spectrum (b). The peak spectral area and the total spectral area are shown, which are used to compute the SQI. The SQI value is 0.09, which is below the specified threshold of 0.10. The spectrum has several peaks at low power and no clear dominant frequency from which a heart rate estimate can be computed.

A 7-point median filter was applied to the heart rate estimates. In addition, if the difference between the heart rate estimate and the estimate from the previous window was 10 beats/minute or was greater than 10% of the previous heart rate value, the estimate was discarded. This process placed a smoothness criterion on the heart rate estimates to remove large changes between windows (which are physiologically implausible). The computation of the final heart rate estimates was as follows:

$$HR_k = median(HR_j), \quad \{\, j | j \in \mathbb{Z} \wedge k - 6 \leq j < k \,\} \tag{6.7}$$

$$HR_k = \left\{ \begin{array}{ll} HR_k, & k = 1 \\ HR_k, & \{\, k | k \in \mathbb{Z} \wedge 2 \leq k < K \wedge |HR_k - HR_{k-1}| \leq thresh_k \,\} \\ None, & otherwise \end{array} \right\} \tag{6.8}$$

$$thresh_k = max(10, 0.1 \times HR_{k-1})$$

where HR_k is the final estimate of heart rate and $thresh_k$ is the smoothness criterion used to establish whether to discard the estimate, for each time-window.

6.4 Respiratory rate estimation

6.4.1 Overview

An algorithm was developed to estimate respiratory rate from signals extracted from pixel values in each video frame. The steps of the algorithm are summarised in the flow chart presented in figure 6.17 and are described in detail in sections that follow. An overview of the algorithm is as follows:

1. A square ROI is selected on the patient's chest, with size and location as described in section 6.4.2.

2. The video data are segmented into time-windows from which the algorithm computes respiratory rate estimates. We chose overlapping 30-second windows with a step size of 1 second. Thus, respiratory rate estimates are computed every second.

3. Within each time-window, a raw signal is extracted by taking the mean of the pixel values within the ROI in each video frame. This signal is a univariate time-series (one value per video frame) called the raw ROI time-series.

4. The raw ROI time-series is band-pass filtered and the onsets and peaks of the respiratory waveform are found with a peak detection algorithm (section 6.4.3).

5. Two respiratory rate estimates (in breaths/minute) are calculated as 60 divided by the median period (in seconds) between successive onsets (RR_{onset}) and

also successive peaks (RR_{peak}). If RR_{onset} and RR_{peak} differ by more than 4 breaths/minute no estimate is reported, otherwise the respiratory rate is calculated as $(RR_{onset} + RR_{peak})/2$.

6. No estimate is reported if the difference between the respiratory rate estimate and the estimate from the previous window is greater than 4 breaths/minute.

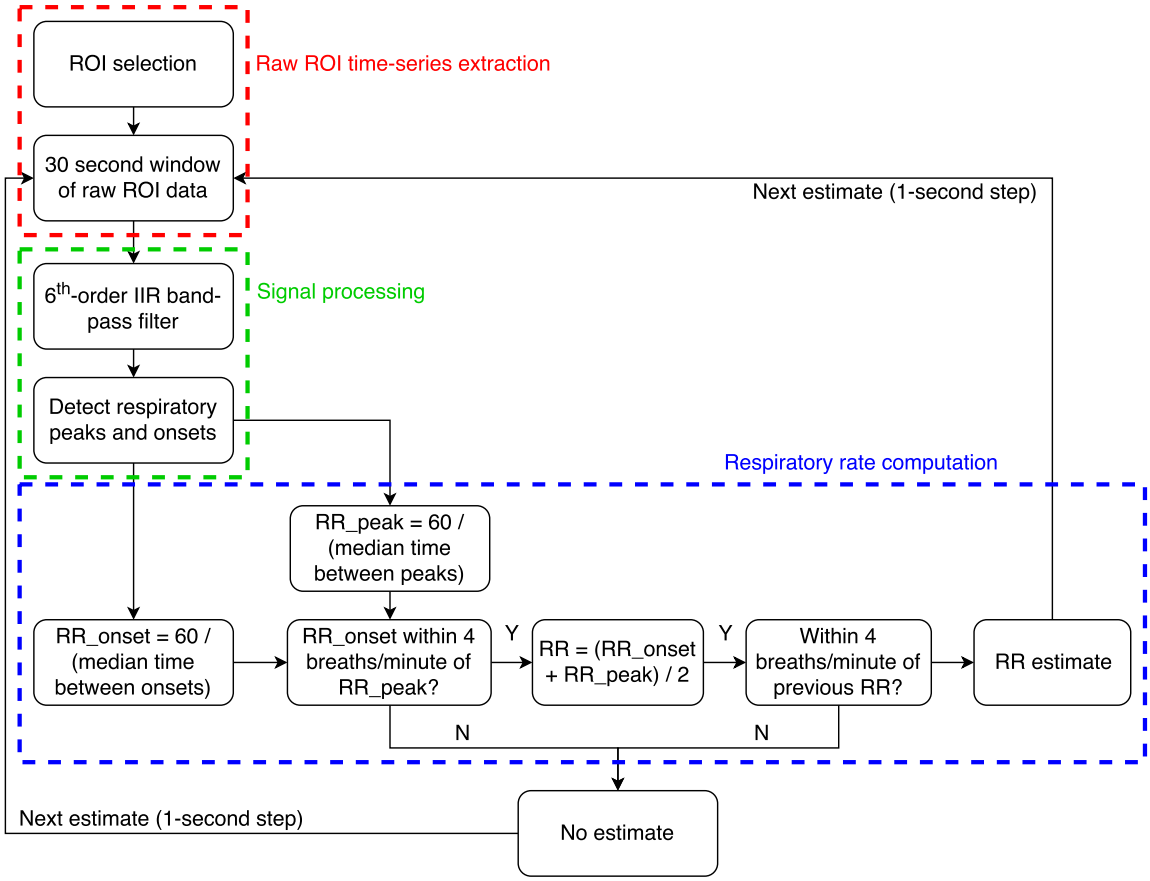


Figure 6.17: Overview of the algorithm used to estimate the respiratory rate in overlapping 30-second windows of raw ROI signal. Zero-phase band-pass filtering was performed with cutoff frequencies of 0.1 and 0.7 Hz (6 and 42 breaths/minute). The onsets and peaks of the respiratory waveforms were computed with a peak detection algorithm. RR_{onset} was computed as 60 divided by the median time between successive onsets and RR_{peak} was computed as 60 divided by the median time between successive peaks. The respiratory rate was the average of RR_{onset} and RR_{peak}.

6.4.2 ROI selection

The first step in non-contact vital-sign estimation of respiratory rate is the selection of an ROI. We manually selected candidate square ROIs of different sizes from locations

on the chest, to extract signals whose amplitude varied with the movement of the chest during respiration. As reported by Villarroel et al. [2014], ROIs for extracting respiratory rate are typically smaller than ROIs for extracting cardiac information as respiratory frequencies are more dominant in smaller ROIs. An example video frame from NVSM15-3 with the location of a 30x30 ROI on the chest is shown in figure 6.18.



Figure 6.18: Example video frame from segment NVSM15-3, demonstrating the position of a 30x30 ROI for respiratory rate estimation, which is drawn with a red border on the middle of the chest.

The process for extracting raw ROI time-series was the same as described in section 6.3.2 for heart rate estimation (pixel values were averaged within the ROI for the illuminated video frames). Example 1-minute segments of raw ROI time series from ROIs of different sizes are shown in figure 6.19. The IP signal from this time segment is also shown for comparison. A clear pulsatile waveform can be seen in the raw ROI time-series, with a frequency that is close to that of the respiratory signal from IP. We tested ROI sizes of 10x10, 20x20, 30x30 and 40x40 pixels. ROI time-series were examined both before and after filtering (signal processing as described in section 6.4.3). The chosen size and location for extracting ROI time-series for respiratory rate estimation from each segment are shown in table 6.4. The optimal specifications differed for each patient, which may have been due to factors such as the distance from the camera to the patient, the pose of the patient, or the amount of clothing or bedding that was covering the patient.

Table 6.4: Respiratory rate estimation ROI size and location for each segment.

Segment	ROI size	ROI location
NVSM15-1	10x10	Right side
NVSM15-2	20x20	Right side
NVSM15-3	30x30	Middle
NVSM19-1	40x40	Middle
NVSM23-1	20x20	Middle

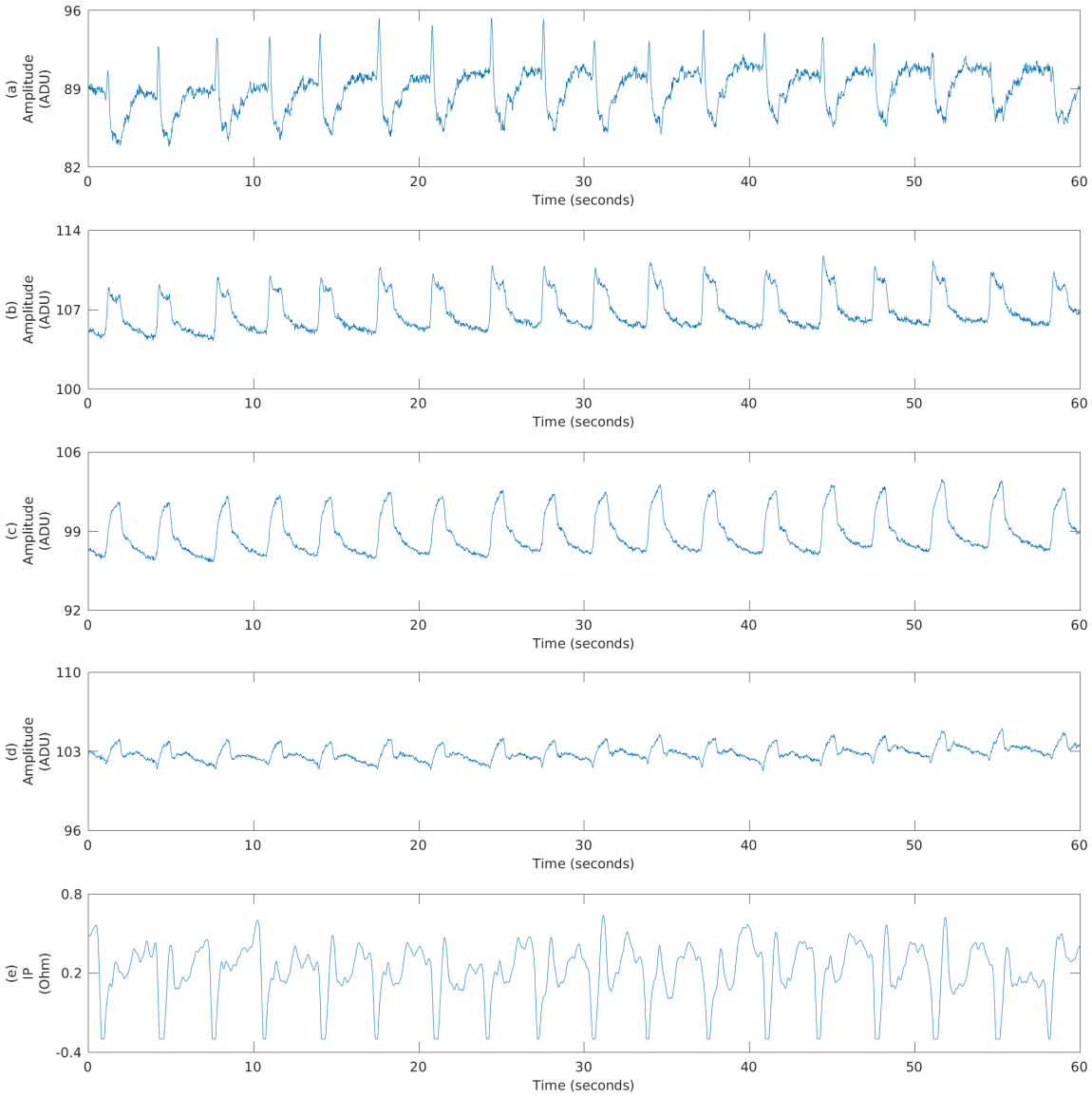


Figure 6.19: Example 1-minute segment of raw ROI time-series data during NVSM15-2, extracted from ROIs on the chest of size (a) 10x10 pixels, (b) 20x20 pixels, (c) 30x30 pixels and (d) 40x40 pixels. A contemporaneous segment of respiratory signal from the Impedance Pneumography electrodes (e) is also shown.

6.4.3 Signal processing

Once the raw ROI time-series data were extracted for each window, band-pass filtering was used to remove any cardiac frequency components and low frequency noise. We used a 6th order infinite impulse response (IIR) Butterworth band-pass filter with pass-band limits of 0.1 Hz (6 breaths/minute) and 0.7 Hz (42 breaths/minute). We applied zero-phase filtering with the MATLAB [2016] *filtfilt* function, thus effectively doubling the filter order (as described in section 6.3.3). An example window of ROI time-series before and after filtering is shown in figure 6.20 for NVSM15-1.

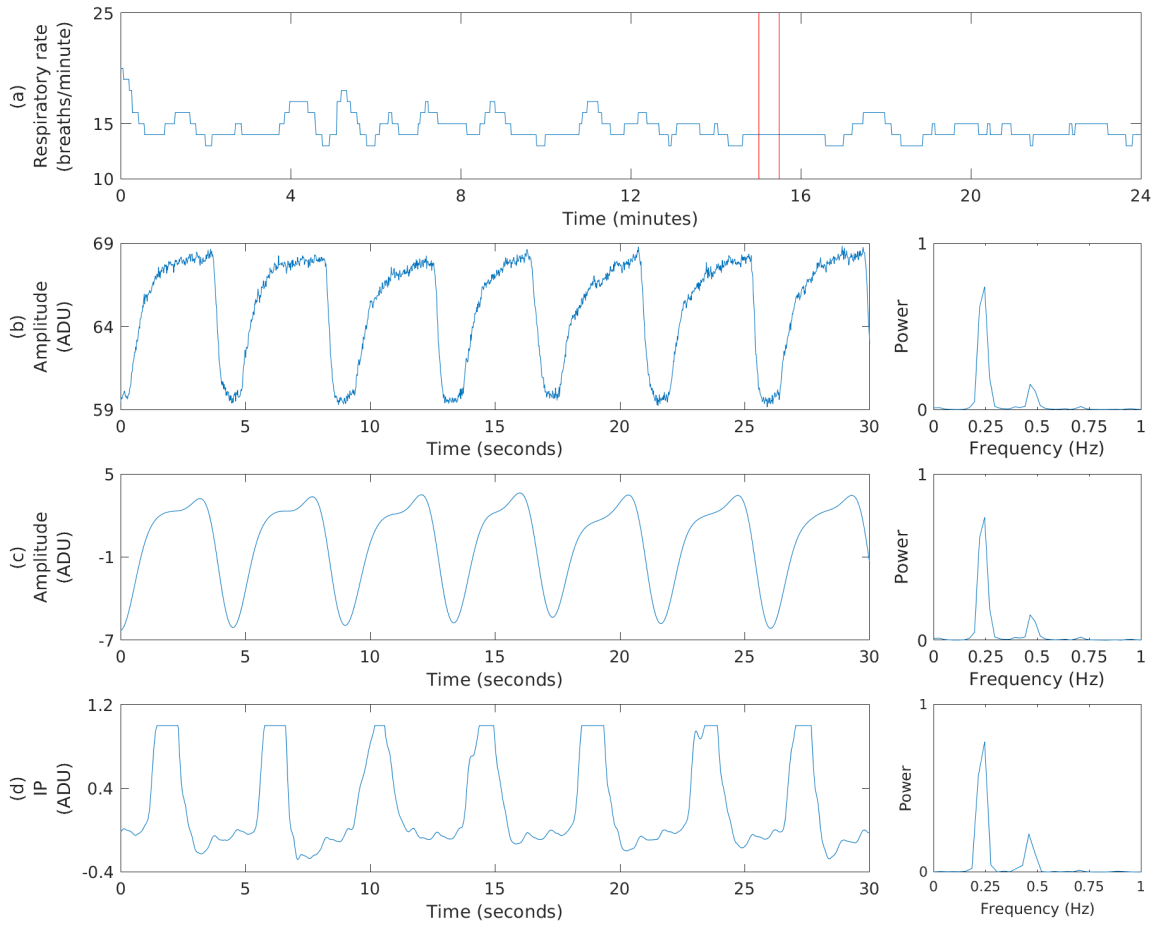


Figure 6.20: Example window of ROI time-series from segment NVSM15-1. The timing of the window is shown with vertical red lines in the respiratory rate time-series from the bedside monitor (a). The raw ROI time-series is shown (b) as well as the filtered ROI time-series after low-pass filtering (c), with 2048-point FFT PSDs (zero-padded from $50 \times 30 = 1500$ samples). A contemporaneous segment of respiratory signal from the impedance pneumography electrodes is shown (d) with a 2048-point FFT PSD (zero-padded from $62.5 \times 30 = 1875$ samples).

Peak detection

We applied a peak detection algorithm to find the onsets and peaks of the respiratory waveforms. The algorithm was designed to find the maximum and minimum values of the respiratory signal within local sections either side of a baseline signal. To find the baseline signal we applied an exponential moving average filter to the respiratory signal. The baseline signal was calculated recursively as follows:

$$B_i = \left\{ \begin{array}{ll} S_i, & i = 1 \\ \lambda \times S_i + (1 - \lambda) \times B_{i-1}, & i > 1 \end{array} \right\} \tag{6.9}$$

where B_i is the baseline signal, S_i is the respiratory signal, i is the sample index and λ is the weighting coefficient. We used $\lambda = 0.01$, which we empirically found to produce a low-amplitude baseline signal, with the onsets and peaks of the respiratory signal consistently above and below this signal.

We then computed the sample indices where the baseline signal and the respiratory signal crossed. To do so, we computed a difference signal $D_i = S_i - B_i$ and found the sample indices where the differences changed sign. This was computed as follows:

$$C^{down} = \{\, i | i \in \mathbb{Z} \wedge i < N \wedge D_i > 0 \wedge D_{i+1} < 0 \,\} \tag{6.10}$$

$$C^{up} = \{\, i | i \in \mathbb{Z} \wedge i < N \wedge D_i < 0 \wedge D_{i+1} > 0 \,\} \tag{6.11}$$

where C^{down} is the set of indices where the respiratory signal moves from above the baseline signal to below the baseline signal, and C^{up} is the set of indices where the respiratory signal moves from below the baseline signal to above the baseline signal. N is the number of samples in the signal.

The onsets and peaks of the respiratory signal were found as the extrema between the crossing points, which were calculated as follows:

$$\begin{aligned} i^{onsets} = \{\, i^* | j \in C^{down} \wedge k^* = \{k^* | k \in C^{up} \wedge t_k > t_j\} \wedge \\ i^* = \{i^* | i \in \mathbb{Z} \wedge t_j < t_i < t_{k^*(1)} \wedge S_{i^*} < S_i \, \forall \, i \neq i^*\}\} \end{aligned} \tag{6.12}$$

$$\begin{aligned} i^{peaks} = \{\, i^* | k \in C^{up} \wedge j^* = \{j^* | j \in C^{down} \wedge t_j > t_k\} \wedge \\ i^* = \{i^* | i \in \mathbb{Z} \wedge t_k < t_i < t_{j^*(1)} \wedge S_{i^*} > S_i \, \forall \, i \neq i^*\}\} \end{aligned} \tag{6.13}$$

where i^{onsets} are the indices of the waveform onsets and i^{peaks} are the indices of the waveform peaks. i^* is the index of the local extremum between the crossing points j in C^{down} and k in C^{up}. $k^*(1)$ is the index of the first upward crossing after each downward crossing, while $j^*(1)$ is the index of the first downward crossing after each upward crossing. t is the time of the sample.

We applied a post-processing step to remove incorrect onsets or peaks caused by noise in the respiratory signal erroneously crossing the baseline signal. If the next peak after an onset occurred within 33 samples then the onset and peak were both removed. This ensured that each (peak-onset) time corresponded to an equivalent respiratory rate less than 45 breaths/minute. An example 3-minute segment of filtered ROI time-series with onsets and peaks detected is shown in figure 6.21.

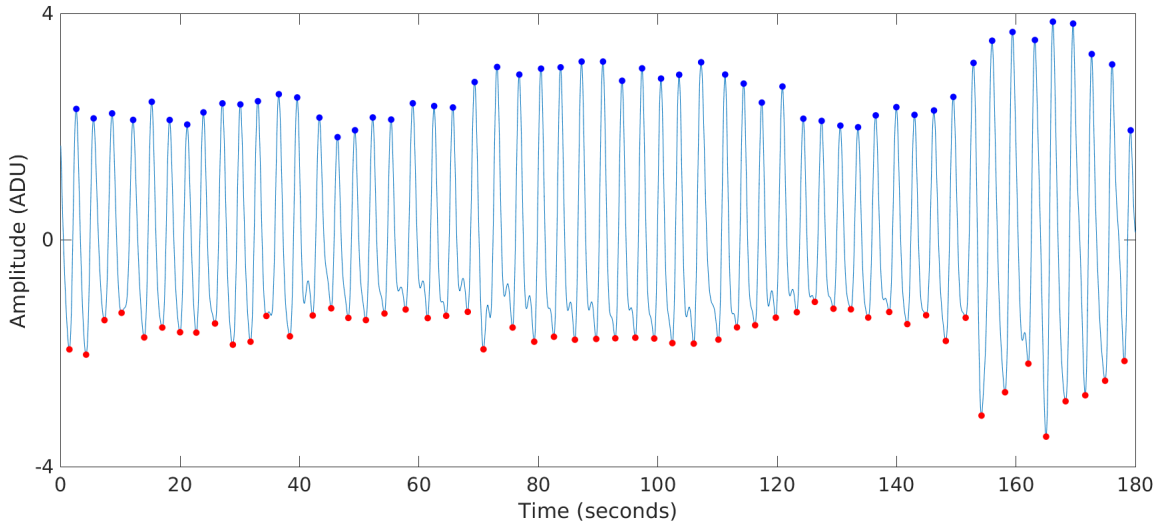


Figure 6.21: A 3-minute period of filtered 20x20 ROI time-series from the right side of the chest during segment NVSM15-2. The peaks of the respiratory waveforms are shown as blue dots and the onsets are shown as red dots.

6.4.4 Computing respiratory rate estimates

A respiratory rate estimate was computed for each window of ROI data based on the time between consecutive respiratory waveform peaks and onsets, using the following steps. First, the periods between consecutive peaks were calculated from which the median period for the window was computed. The respiratory rate estimate (in breaths/minute) based on the peaks, RR^{peaks}, was calculated as 60 divided by the median period between peaks (in seconds), as follows:

$$RR^{peaks} = \frac{60}{median(t(i_{j+1}^{peaks}) - t(i_j^{peaks}))}, \quad j = 1, 2, ..., M-1 \tag{6.14}$$

where t is the time of the sample and i^{peaks} are the indices of the M peaks in the time-window. This process was repeated using the median period between consecutive onsets, to compute a respiratory rate estimate from the onsets, RR^{onsets}.

To ensure that the peaks and onsets detected were consistent, a respiratory rate was only computed from a window of ROI data if RR^{peaks} and RR^{onsets} were within 4 breaths/minute. If the two estimates agreed, the respiratory rate was calculated as $(RR^{peaks} + RR^{onsets})/2$. As with heart rate estimation (section 6.3.5), we used a smoothness criterion to remove large changes between consecutive windows. If the absolute difference between the respiratory rate estimate and the estimate from the previous window was greater than 4 breaths/minute, the estimate was discarded. The final respiratory rate computation is described in equations (6.15) and (6.16), as follows:

$$RR_k = \left\{ \begin{array}{ll} \dfrac{RR_k^{peaks} + RR_k^{onsets}}{2}, & \left| RR_k^{peaks} - RR_k^{onsets} \right| \leq 4 \\ None, & otherwise \end{array} \right\} \tag{6.15}$$

$$RR_k = \left\{ \begin{array}{ll} RR_k, & k = 1 \\ RR_k, & \{\, k | k \in \mathbb{Z} \wedge 2 \leq k < K \wedge |RR_k - RR_{k-1}| \leq 4 \,\} \\ None, & otherwise \end{array} \right\} \tag{6.16}$$

where RR_k is the final respiratory rate estimate calculated from RR_k^{peaks} and RR_k^{onsets} for the kth time-window. K is the total number of time-windows in the segment.

6.5 Results

The performance of the heart rate and respiratory rate estimation algorithms was evaluated by comparing the camera-derived estimates with the equivalent reference values from the Philips bedside monitor. We calculated the Mean Absolute Error (MAE) using equation (6.17), and the Root Mean Squared Error (RMSE) using equation (6.18), between the K estimates (est) and reference values (ref) in each segment. Reference values during times of discarded camera-derived estimates were not included in the calculation. RMSE is similar to MAE but penalises erroneous

values more severely. We also calculated the percentage number of estimates that were discarded during each segment, and the time gaps between consecutive estimates.

$$MAE = \frac{1}{K}\sum_{k=1}^{K}|est_k - ref_k| \tag{6.17}$$

$$RMSE = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(est_k - ref_k)^2} \tag{6.18}$$

Table 6.5 gives the results comparing the heart rate and respiratory rate estimates derived from analysis of the camera signals to the reference values from the monitor. Heart rate and respiratory rate time-series plots for each segment are shown in figures 6.22 to 6.26. The distribution of time gaps between consecutive heart rate and respiratory rate estimates are shown in figures 6.27a and 6.27b.

Table 6.5: Results for the comparison of camera-derived estimates of heart rate and respiratory rate with reference values from the Philips bedside monitor for each segment. Respiratory rate estimation performance is not given for segment NVSM15-3 as there were no reference values available for comparison.

Heart rate				
NVSM segment	**MAE beats/minute**	**RMSE beats/minute**	**Monitoring time minutes:seconds**	**Estimation time minutes:seconds (% of monitoring time)**
15-1	0.8	0.9	23:45	21:47 (91.8)
15-2	0.8	1.1	29:45	26:41 (89.7)
15-3	1.2	1.7	29:45	27:06 (91.1)
19-1	2.2	3.1	29:45	15:03 (50.6)
23-1	2.8	3.4	29:45	20:16 (68.1)
Respiratory rate				
15-1	0.6	0.8	23:30	23:30 (100.0)
15-2	1.5	2.4	29:30	29:28 (99.9)
15-3	NA	NA	29:30	29:30 (100.0)
19-1	0.4	1.3	29:30	25:52 (87.7)
23-1	2.0	2.9	29:30	19:48 (67.1)

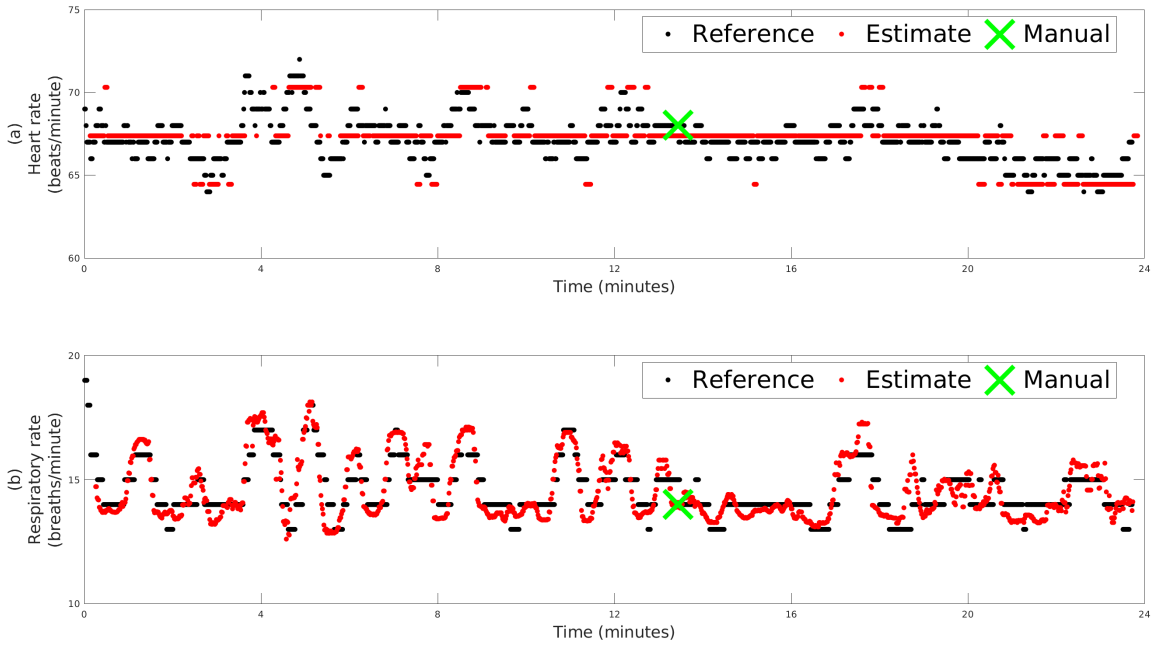


Figure 6.22: Camera-derived heart rate and respiratory rate estimates compared with reference values for segment NVSM15-1.

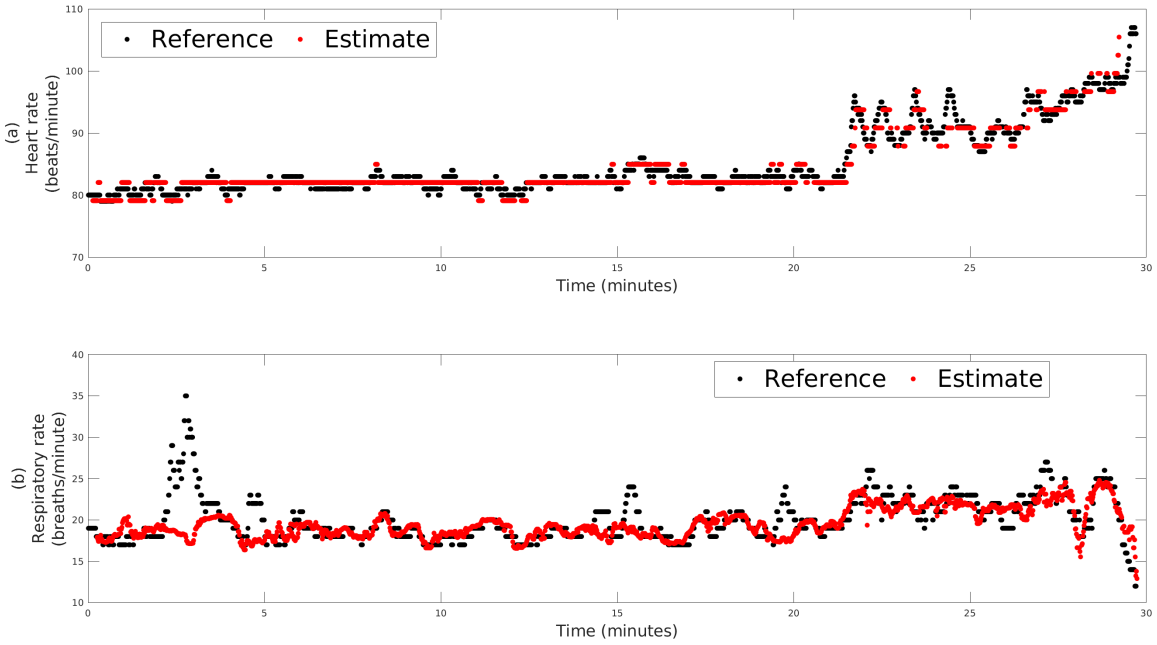


Figure 6.23: Camera-derived heart rate and respiratory rate estimates compared with reference values for segment NVSM15-2.

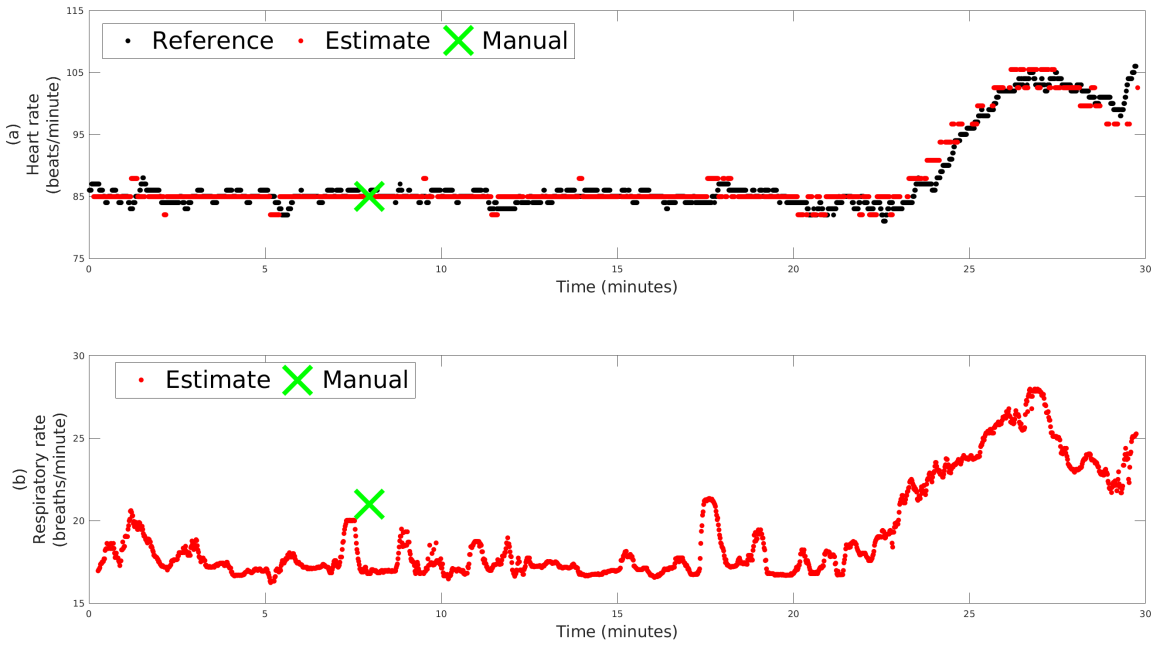


Figure 6.24: Camera-derived heart rate and respiratory rate estimates compared with reference values for segment NVSM15-3. Reference continuous respiratory rate values were not available as there was no impedence pneumography signal.

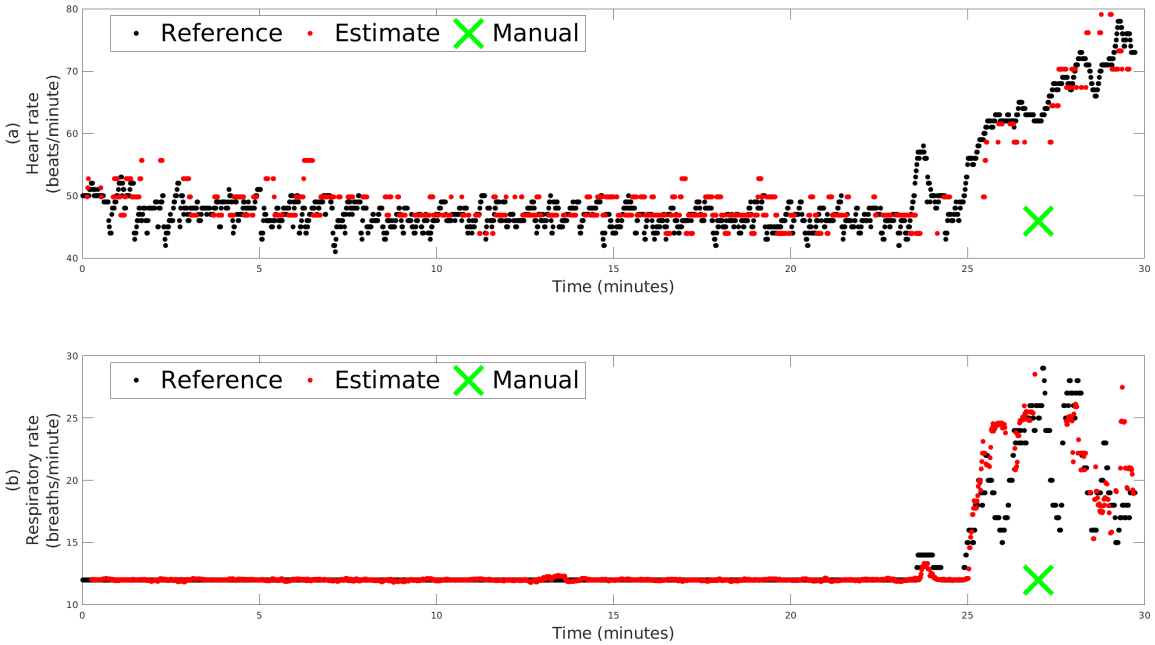


Figure 6.25: Camera-derived heart rate and respiratory rate estimates compared with reference values for segment NVSM19-1.

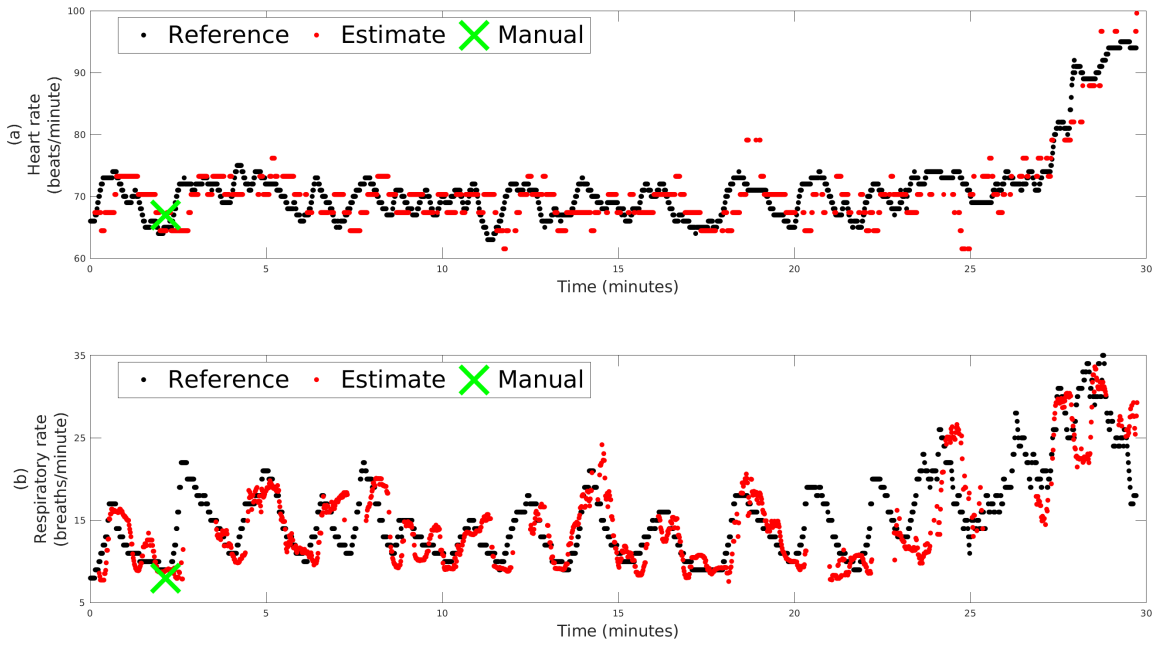
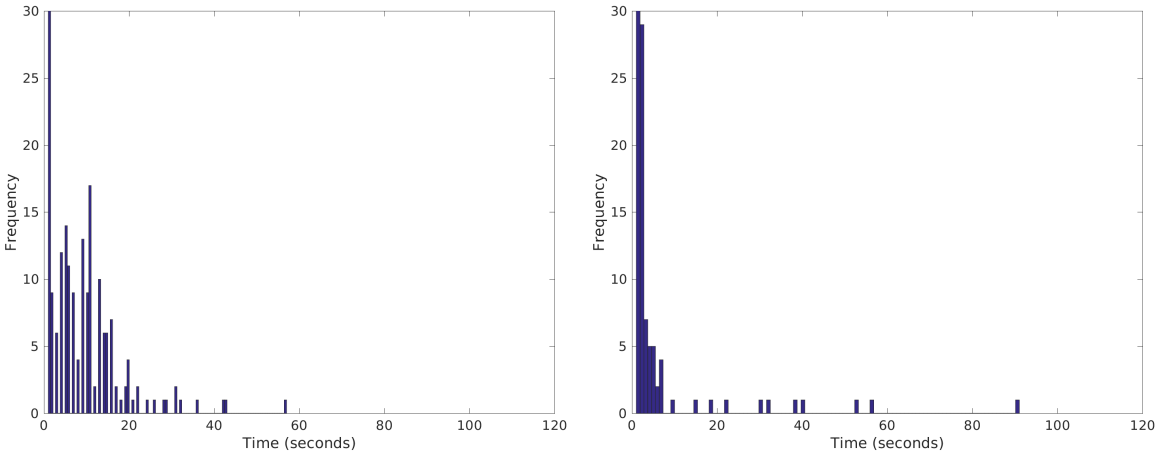


Figure 6.26: Camera-derived heart rate and respiratory rate estimates compared with reference values for segment NVSM23-1.



(a) Heart rate **(b)** Respiratory rate

Figure 6.27: Distributions of time gaps between successive heart rate (a) and respiratory rate (b) estimates for all 5 study segments. The majority of estimates were separated by 1-second gaps, hence the y-axis was truncated to focus on gaps due to missing values.

6.6 Discussion

The heart rate and respiratory rate estimates derived from the video camera signals had a high level of accuracy within 3 beats/minute and 2 breaths/minute of the reference value for all five segments. As reported in table 6.5, the MAE for heart rate varied from 0.8 to 2.8 beats/minute and the MAE for respiratory rate varied from 0.4 to 2.0 breaths/minute, compared to the Philips bedside monitor values in each segment. Estimates were more accurate for the three segments taken from NVSM15. This is a reflection of the level of variance in the signals. NVSM23-1 had noticeably high variance as shown in figure 6.26, particularly for respiratory rate, which was likely caused by abnormal respiratory physiology related to the maxillofacial cancer surgery or the associated therapies such as supplemental oxygen support and tracheostomy, as described in section 6.2. Periods of high variance often occurred as a result of patient movement, such as the final 5 minutes of respiratory rate in NVSM19-1 (figure 6.25). The non-contact estimates also matched the manual observations shown in figures 6.23 to 6.25. The manual observations lagged the continuous vital-sign reference values and non-contact estimates in figure 6.26, which may have been due to a delay in the time of manual documentation.

The non-contact estimates of heart rate and respiratory rate followed the trends in the reference data during each segment. In figures 6.22 and 6.24 to 6.26 there were increases in heart rate that were tracked by the non-contact estimates. During these times there were fewer estimates reported by the camera. The vital-sign changes often occurred during times of increased patient movement so the camera algorithm could not always report a reliable estimate. The short-term respiratory rate fluctuations during NVSM23-1 (figure 6.26) were mostly tracked by the non-contact estimates. There was also an increase in respiratory rate during the final 5 minutes of NVSM19-1, although this corresponded to a high level of noise, perhaps due to patient movement that was noticeable in the video (figure 6.25). In this instance, the non-contact estimates also increased but did not always match the reference values.

The non-contact estimation algorithms were designed to track smooth changes in heart rate and respiratory rate. The discarding of estimates with large changes (10% for heart rate or 4 breaths/minute for respiratory rate) eliminated instances of movement artefact. However, these rules could also lead to genuine short-term physiological changes being ignored. One notable physiological change that existed in the reference vital-sign time-series but not the non-contact estimation time-series occurred approximately 3 minutes into the respiratory rate time-series in figure 6.23.

Rerunning the algorithm without the smoothness criterion and with a smaller window size of 10-seconds (shown in figure 6.28) revealed that most of the non-contact estimates during this brief period were similar to those shown in figure 6.23. Thus, the smoothing rules did not cause the estimation algorithm to miss this short term increase in respiratory rate. The reason that the non-contact estimates missed this physiological change is unknown although it is possible that this was an error in the reference signal.

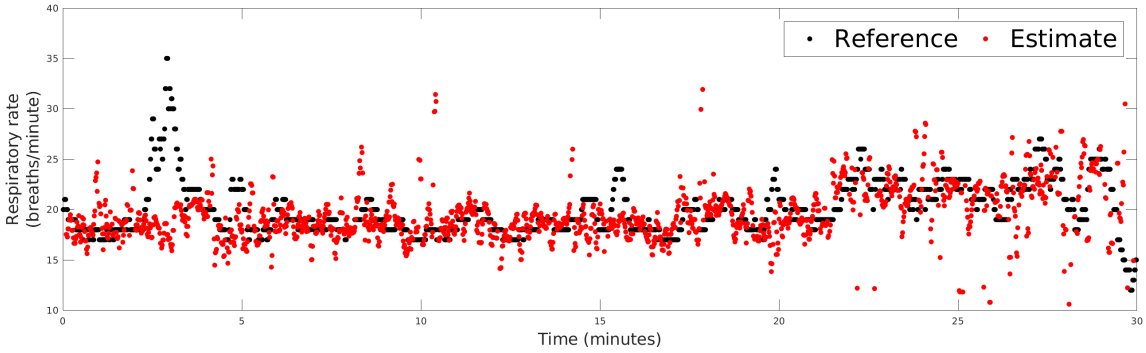


Figure 6.28: Camera-derived respiratory rate estimates compared with reference values for segment NVSM15-2. The estimation algorithm used 10-second sliding windows and did not include a smoothness criterion.

While the non-contact estimates were occasionally sparse (e.g. 49.4% of heart rate estimates were discarded in NVSM19-1, as shown in table 6.5), there were several non-contact estimates reported during all major physiological changes. Indeed, the distribution of time-gaps between estimates given in figures 6.27a and 6.27b shows that there was only 1 instance where respiratory rate estimates were separated by more than 1 minute (91 seconds), while there were no such instances for heart rate estimation. Thus, aside from short occurrences of patient movement, non-contact estimates were consistently reported throughout the recording sessions.

Non-contact monitoring with a video camera may be a viable method of continuously monitoring patients during post-operative care. Non-contact estimates tracked changes in physiology that occurred between manual vital-sign estimates. The ability to provide continuous heart rate and respiratory rate estimates in combination with manual observations of the other vital signs may be a useful addition to clinical care. Indeed, processing of the segments presented in this chapter produced non-contact vital-sign estimates during times when clinically significant changes in vital-sign values occurred between manual observations.

Non-contact methods may be most useful in clinical scenarios with intermediate levels of acuity, such as wards with acuity between that of critical care (requires

granular, beat-by-beat heart rate recordings) and low-risk general wards (requires periodic manual observations usually every 4 to 8 hours). Indeed, with estimates every minute or less, non-contact methods can provide vital-sign values much more often than manual observations. Post-operative care is an example of a setting with sufficiently high levels of adverse clinical events where camera-derived vital-sign estimation could be a cost-effective method of monitoring patients without requiring them to wear sensors for prolonged periods of time.

The ability to track vital-sign changes between manual observations is important in the context of EWSs. Considering that an elevated EWS score is an institutionally-defined criterion for escalating care, it is important that all times of increased EWS are noted and not just during times of manual observations. NVSM15-2, NVSM15-3, NVSM19-1 and NVSM23-1 all correspond to periods of time with an EWS value of 1 or more (see EWS cut-offs in table 3.4). In all of these instances, there was either a delay before a manual observation was taken or there was no manual observation at all. This demonstrates the potential usefulness of non-contact continuous vital-sign monitoring, allowing the detection of physiological deterioration in real-time.

Chapter 3 showed how vital-sign variability is an important feature in algorithms for detecting physiological deterioration. Furthermore, chapter 5 demonstrated how variability can be calculated from continuous vital-sign time-series. Here we demonstrate that vital-sign changes that contribute to the calculation of diurnal vital-sign variability can be tracked with non-contact methods. An EWS that combines heart rate, heart-rate variability, respiratory rate and respiratory-rate variability could be used with non-contact estimates to predict physiological deterioration continuously. This score could be periodically updated with blood pressure, temperature and SpO_2 values from manual observations. Such a monitoring system would combine the benefits of a predictive algorithm with (potentially) high sensitivity and specificity, with the ability to monitor a patient without requiring nursing staff to be present and without the discomfort of traditional contact sensors.

There were many limitations to the NVSM study. The mean age of the patient cohort was 62 years (table 6.1), which may limit the generalisability of the results to more elderly populations. The study population did not contain any people with skin colour above 3 on the Fitzpatrick scale. The performance of heart rate estimation algorithms may be affected by skin tone and thus results derived from this study will require further validation using data obtained from a more diverse cohort. Furthermore, the sample size obtained in this study is insufficient to provide gold-standard evidence for any hypotheses, so conclusions are of necessity limited to proof-of-principle level.

The use of the privacy blind in the NVSM study tended to overlap with the times of most interesting physiology. In choosing the segments for this chapter, many of the segments were immediately followed by times with the privacy blind down, thus removing the ability to track the patient throughout this time. For example, it would have been useful to have video footage beyond NVSM15-3 during the period of tachycardia (shown in figure 6.8). However, the fact that the privacy blind was down provided further validation that this was indeed a period during which the patient experienced physiological discomfort.

The reported results were obtained in 5 short segments from 3 different patients. This is insufficient evidence to draw conclusions on the efficacy of the methodology used. However, the purpose of this analysis was to demonstrate the feasibility and potential usefulness of non-contact monitoring in the clinical context of post-operative care. Indeed, several methodological improvements are needed before non-contact monitoring can be validated on the full NVSM dataset. Firstly, the ROI regions were hand-selected in this chapter, which was possible because there was very little variation in the patients' position throughout the recordings. ROIs will need to be automatically selected and tracked so that monitoring can continue after a patient changes position. This has been done with computer vision techniques and deep learning, but this requires tuning for each particular dataset (Chaichulee et al. [2017]). Furthermore, the optimal ROI changes throughout a recording and hence it is necessary to be able to change ROI and/or fuse estimates from multiple ROIs, with methods such as multi-output Kalman filters (Villarroel et al. [2017]).

6.7 Conclusions

We have presented a clinical study that was conducted to assess the use of extended non-contact monitoring in an acute care setting. We investigated five segments recorded during this study, during which time the patients experienced changes in physiology. We demonstrated that clinically-relevant changes in vital-sign values can be accurately tracked with camera-based non-contact methods.

We described how non-contact monitoring may be a suitable method for providing continuous vital-sign monitoring of patients in wards of intermediate acuity, for whom continuous monitoring is currently not practical. When combined with an EWS with vital-sign variability features, non-contact monitoring may provide automated early warning of deterioration.

Before non-contact monitoring can be fully deployed in a clinical setting, further

work is required to ensure that robust estimates are reported to clinical staff. Necessary improvements include better means of automatically finding ROIs, improved signal processing methods for obtaining estimates, and methods to fuse multiple estimates together to provide a robust estimate. Further studies are needed to obtain data from a more general population (e.g. in terms of skin type) containing many more examples of interesting physiological events.

Chapter 7

Localised respiration mapping

Non-contact vital-sign monitoring with a video camera allows simultaneous physiological estimates to be obtained from different areas of the body. ROIs can be chosen from several locations and estimation algorithms may be run in parallel to derive multiple non-contact measurements. Spatial variations in physiology can therefore be visualised to provide clinical information such as the local effects of certain pathologies or the locations of regions of unusual physiological activity. This information represents an additional benefit of non-contact monitoring over traditional contact methods.

In this chapter we describe a method of visualising spatial variations in respiration over ROIs on the chest. We revisit the case study of patient NVSM15 (presented in chapter 6), who suffered a pneumothorax and a ruptured diaphragm. We investigate whether spatial differences in respiration can be detected during a period of abnormal respiratory physiology.

We begin by providing an overview of previous work on spatial analyses of physiology with a video camera (section 7.1), before describing our data (section 7.2). We then propose a method of estimating the accuracy of non-contact estimates of respiratory rate in the absence of reference data, with which we support the analyses presented in this chapter. We then assess the temporal variations in the respiratory signal from different ROIs (section 7.3) before focussing on the spatial variations in the respiratory signal at time-points of interest (section 7.4). Finally, we describe a method of representing the spatial variations in respiration as a colour-coded physiological map and present results for patient NVSM15 (section 7.5).

7.1 Previous work

Methods for visualising spatial distributions of physiological parameters have previously been described in the non-contact monitoring literature. Spatio-temporal maps of haemodynamic data have been computed, allowing the visualisation of blood perfusion and blood volume changes (Daly [2016]; Kamshilin et al. [2011, 2013]; Zaunseder

et al. [2018]). The origins of these methods began with early work demonstrating that changes in pixel intensity over an area of skin from near-infrared charge-coupled device (CCD) cameras could be detected during venous reflow after physical stimuli (Such [1996]; Such et al. [1997]; Wu [2003]. Spatial maps of blood perfusion with high spatial and temporal resolution have since been produced (Kamshilin et al. [2011, 2013]) and these have been used to develop new hypotheses about the physiological origins of the PPGI signal (Kamshilin et al. [2015, 2016]). Finally, spatio-temporal maps of blood perfusion amplitude and phase have been presented as first steps in the development of algorithms for non-contact estimation of PTT (Daly [2016]; Zaunseder et al. [2018]).

Spatial maps of cardiac and respiratory information have been presented as supplemental analyses in several non-contact monitoring papers that have a primary focus on heart rate and respiratory rate estimation (Tarassenko et al. [2014]; Verkruysse et al. [2008]; Wieringa et al. [2005]). These maps are generally created by computing the strength of the physiological frequency in small ROIs distributed across the image (from signal PSDs), which are then overlaid onto the image with a colour-coded representation. ROI locations can then be selected based on areas of high signal strength with respect to the physiological frequency of interest (Fallet et al. [2016]).

Depth sensing cameras such as the Microsoft Kinect have been used to measure respiratory-induced changes in physiological surface motion for the purpose of target tracking during radiotherapy (Ortmüller et al. [2015]; Tahavori et al. [2014]; Wentz et al. [2012]; Wijenayake and Park [2017]). Three-dimensional models of chest movement can be developed from the camera signal, which allow the displacement of organs to be predicted throughout the respiratory cycle. Similar methods have been presented to monitor respiratory motion for more general purposes (Bernal et al. [2015]), with proposed applications such as in-home sleep monitoring (Procházka et al. [2016a,b]) and breathing analyses during exercise (Povšič et al. [2015]). Two-dimensional colour maps of chest displacement have been presented in many of these studies. Finally, Smith et al. [2016] developed a system for estimating respiratory flow rate using a geometric model of chest motion developed from monochromatic video data with multiple infrared light sources.

Perhaps the most interesting applications of physiological spatial maps are those applied to patients with underlying physiological abnormalities. For example, Verkruysse et al. [2008] assessed blood perfusion in a discolouration of the skin known as a port-wine stain, while Zaproudina et al. [2013] mapped abnormal facial blood perfusion patterns in patients who suffer from migraines. Tarassenko et al. [2014] mapped the

distribution of cardiac and respiratory frequency strength in facial ROIs in patients undergoing haemodialysis treatment. Finally, Wentz et al. [2012] assessed depth-based respiration modelling methods on three cancer patients. However, despite most authors proposing clinical applications for their methodologies, the majority of these studies have analysed healthy subjects in controlled laboratory environments. Further opportunities exist to demonstrate novel use cases for spatial maps of physiology applied to visualise phenotypes of physiological abnormalities.

7.2 Patient case study

In this chapter we revisit the case study of patient NVSM15, which was first described in section 6.2. NVSM15 was a 79-year old female with Fitzpatrick type 1 skin, who was recovering from gynaecological cancer surgery. NVSM15 suffered a right-sided pneumothorax, which was diagnosed on admission to the ICU, and a ruptured diaphragm, which was diagnosed at the end of the recording period. The patient's liver herniated into the chest as a result of the ruptured diaphragm.

Here we focus on the time between the penultimate chest X-ray that was taken during the recording, which showed no sign of a ruptured diaphragm, and the final X-ray from which the ruptured diaphragm was diagnosed. The diaphragm must have ruptured during the final 24 hours, but it is not known exactly when this occurred. As shown in figure 6.5, the penultimate X-ray occurred 24 hours before the end of the study recording. A detailed timeline for this segment is shown in figure 7.1.

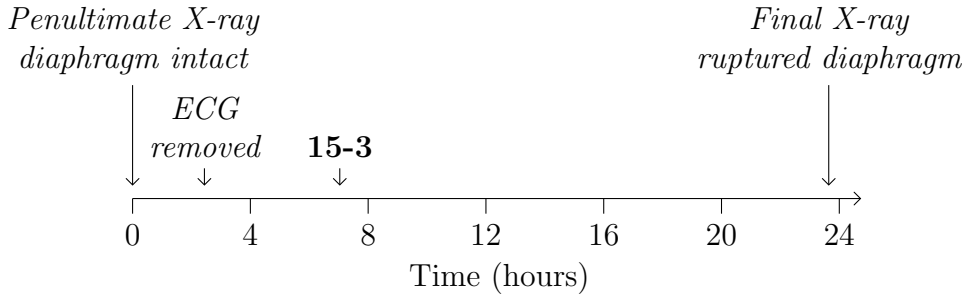


Figure 7.1: Timeline of the final day of recording for patient NVSM15. The patient had a right-side pneumothorax throughout this period. The diaphragmatic rupture was diagnosed at the end of the recording, thus the onset must have occurred during the final 24 hours. 15-3 refers to segment NVSM15-3 (see chapter 6).

As described in section 6.2, NVSM15-3 is a 30-minute segment that occurred approximately 7 hours after the penultimate X-ray, with reference vital-sign values

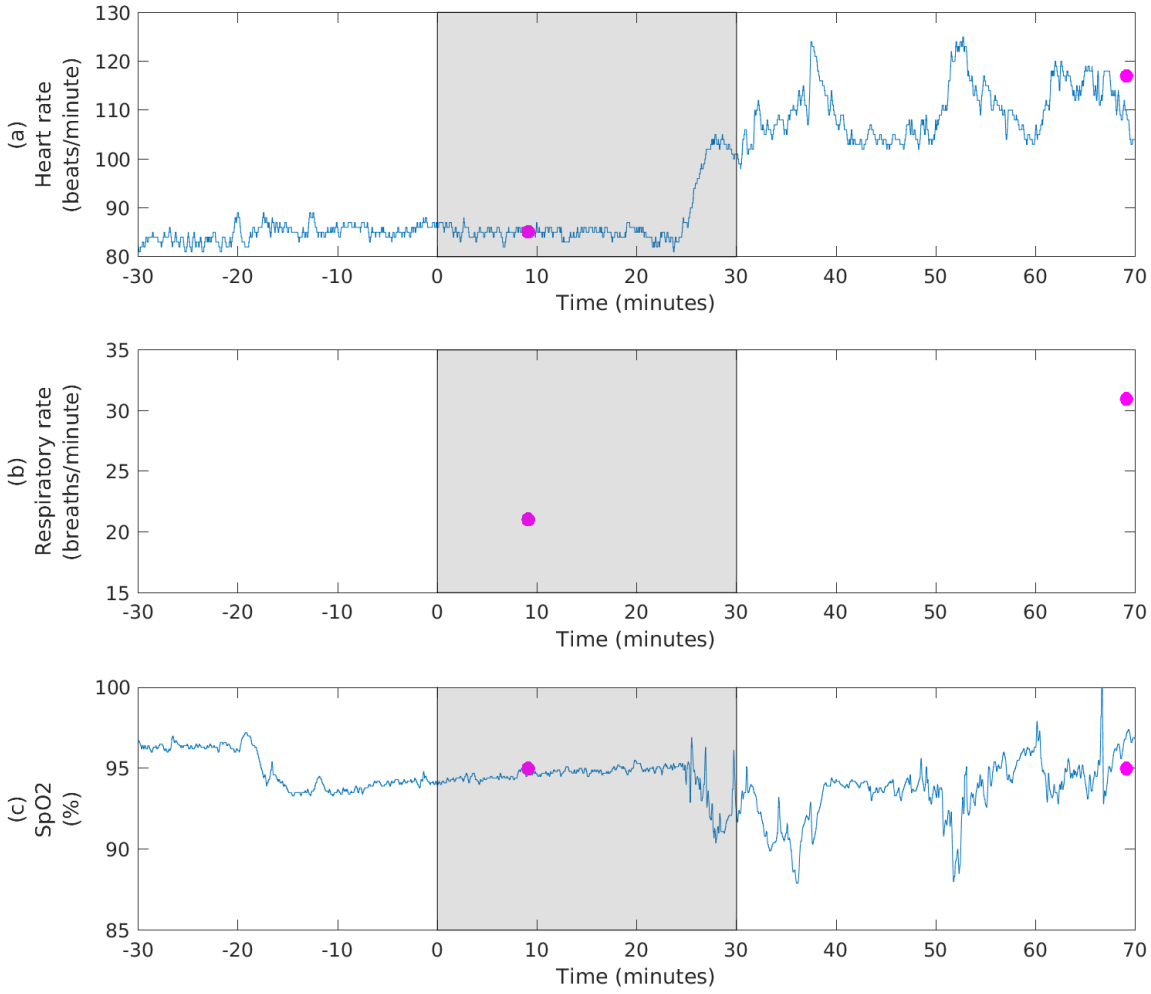


Figure 7.2: Reference heart rate (a), respiratory rate (b), and SpO_2 (c) for segment NVSM15-3 (shaded in grey). Manual vital-sign estimates are shown as pink dots and bedside monitor estimates are shown as blue lines. Data are also shown 30 minutes before and 40 minutes after the segment to provide physiological context. NVSM15 was a 79 year-old female patient recovering from gynaecological cancer surgery. Re-plotted from figure 6.8.

shown in figure 7.2 (re-plotted from figure 6.8). During this segment, the heart rate increased by 20 beats/minute and the SpO_2 decreased by approximately 5%. There was no reference respiratory rate signal as the patient's ECG electrodes had been removed (so that there was no IP signal available), although the camera-derived respiratory rate estimates showed an increase of 10-15 breaths/minute that occurred at the same time as the increase in heart rate. The privacy blind was put down at the end of segment NVSM15-3, so there was no video information available to track the patient's physiology after NVSM15-3. However, the reference vital-sign values showed that the tachycardia was sustained for 40 minutes after the end of NVSM15-3, at which

time there was also a manual respiratory rate measurement of 30 breaths/minute.

The data in NVSM15-3 represent the onset of physiological deterioration, with symptoms of respiratory distress such as elevated respiratory rate, elevated heart rate, and decreased SpO_2. Thus, the NVSM15-3 data segment provides an interesting case study for assessing the spatial distribution of respiratory signals over the patient's chest. We note that this data segment occurred in a 24-hour period when the diaphragm ruptured and the liver herniated into the chest. It is therefore of interest to discover whether abnormal spatial patterns of respiration were noticeable in the camera signal.

7.2.1 Validating non-contact respiratory rate estimates

The non-contact estimates of respiration shown for NVSM15-3 in figure 6.24 suggest that NVSM15 increased her respiratory rate by 10-15 breaths/minute towards the end of the segment. However, this is unconfirmed due to the lack of a gold-standard reference respiratory rate signal. While there was a manual measurement confirming that NVSM15 had a highly abnormal respiratory rate of 30 breaths/minute, this value was taken 40 minutes after the end of NVSM15-3 (see figure 7.2). We do note, however, that the respiratory rate estimation algorithm provided accurate estimates for the other 4 segments analysed in chapter 6 when compared with respiratory rate values provided by the Philips bedside monitor.

We propose a method of validating the accuracy of camera-based non-contact respiratory rate estimates in the absence of reference values. Our method is based on the observation that a respiratory motion component is present in ROI time-series taken from the chest, as well as other locations such as the shoulders, the neck and the face. This is intuitive, as the diaphragm, thorax and shoulders rise and fall during the act of breathing, which translates to motion through the neck and head. Respiratory motion will also occur at the nose and mouth during inhalation and exhalation. This effect was noted by Tarassenko et al. [2014] as they derived non-contact respiratory rate estimates from ROIs on the face, and it can be seen in figure 6.14, as the PSD of the raw ROI time-series taken from the cheek has a strong respiratory frequency component. Thus, ROI time-series taken from ROIs on the face and the chest will have correlated respiratory motion components. Other motion components due to random movement are less likely to be as well correlated because of the distance between the face and the chest. For example, if the patient turns her head, this would have more of an effect on the ROI time-series from the face than those from the chest. We therefore propose that if non-contact estimates of respiratory rate derived from

separate ROIs from the chest and the face match, then it is likely that the estimates computed from the camera relate to the underlying respiratory rate of the patient. We used this method to establish the validity of the non-contact respiratory rate estimates derived from the video signal during the NVSM15-3 data segment.

Using the methods described in section 6.4, we extracted raw ROI time-series data from locations on the forehead and on the left and right sides of the chest, and then computed separate respiratory rate estimates from the three ROI time series. We used ROIs of size 30x30 ROIs pixels, which is the same size as the ROI used to estimate the respiratory rate for segment NVSM15-3 in chapter 6 (table 6.4). We chose ROIs from both sides of the chest to assess whether differences in respiratory rate estimates could exist because of the right-sided pneumothorax or the ruptured diaphragm. We assessed the results quantitatively in table 7.1 by reporting the MAE and the RMSE (as described in section 6.5).

Table 7.1: Comparison of the camera-derived respiratory rate estimates from a 30x30 pixel ROI on the forehead (the "reference") with the camera-derived respiratory rate estimates from 30x30 pixel ROIs on the right and left sides of the chest.

Side of chest	**MAE breaths/minute**	**RMSE breaths/minute**
Right	0.282	0.548
Left	0.335	0.639

The respiratory rate estimates from all three locations are similar, with MAEs of 0.282 and 0.335 breaths/minute between the estimates from the forehead and those from the right and left sides of the chest, respectively (table 7.1). The estimate from the forehead tracks the estimates from both sides of the chest, as shown in figure 7.3. Minor discrepancies can be seen near the end of the segment (approximately 29 minutes) when there were periods of patient movement. The estimates all track the brief increases in respiratory rate that occurred after approximately 2 and 17 minutes, as well as the significant rise in respiratory rate that occurred in the final 7 minutes of the segment. The similarity of the respiratory rate estimates derived from ROIs on both sides of the chest is expected, as the mechanisms driving respiration do not affect the rate in each lung separately, even during the presence of an injury that may have an asymmetric effect (e.g. a right-sided pneumothorax). This analysis provides evidence that the non-contact estimates of respiratory rate derived for data segment NVSM15-3 are accurate and provides support for our analysis of respiratory signals in this chapter.

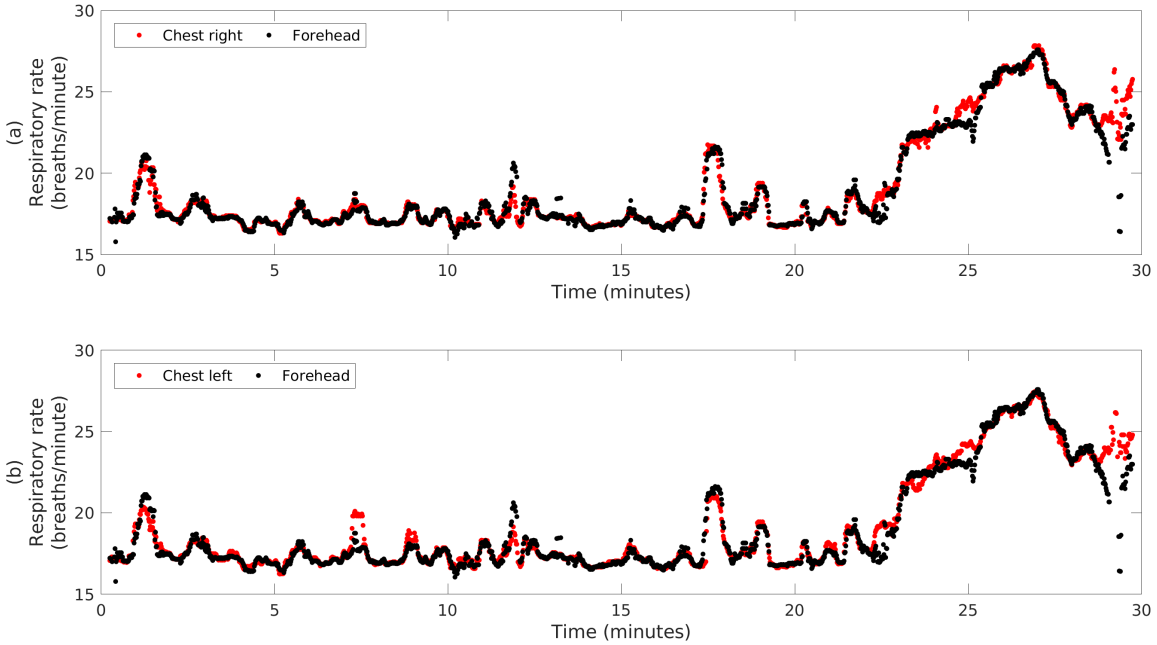


Figure 7.3: Non-contact respiratory rate estimates derived from 30x30 pixel ROIs on the right (a) and left (b) sides of the chest compared to respiratory rate estimates derived from a 30x30 pixel ROI on the forehead, during segment NVSM15-3.

7.3 Temporal variations in respiration

7.3.1 Overview

We began our analysis by assessing the temporal variations in the respiratory signals extracted from various ROIs distributed over the chest. We aimed to visualise whether similar trends in the respiratory signal would occur in different areas of the chest, in the clinical context of the right-sided pneumothorax, the ruptured diaphragm, or the respiratory distress of the patient in general. We focussed on the final 15 minutes of segment NVSM15-3, when the most interesting changes in respiratory rate occurred.

7.3.2 Respiratory signal extraction

ROI time-series data were extracted from square ROIs with 30x30 pixels, which were distributed in locations over the chest as shown in figure 7.4. We chose ROIs located vertically on the top, middle and bottom of the chest as well as horizontally on the far-right, mid-right, mid-left and far-left of the chest. We grouped ROIs by their horizontal location for the purpose of visualising the respiratory signal.

Once the raw ROI time-series were extracted from each location, the signal

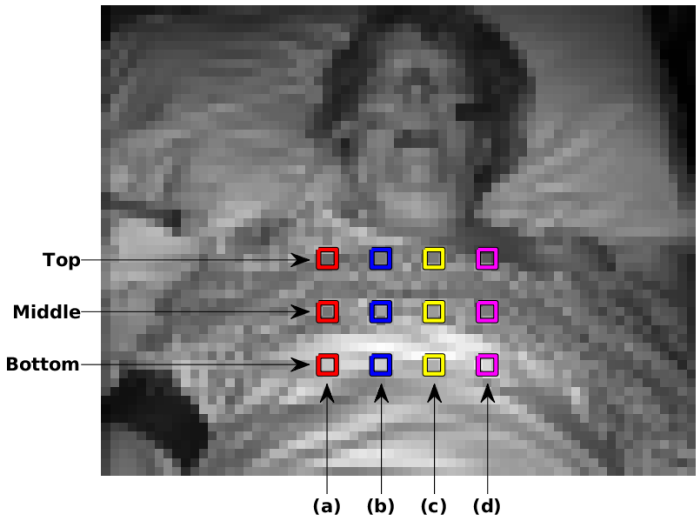


Figure 7.4: Positions of the 30x30 pixel ROIs from which respiratory signals were extracted for the analysis of temporal variations in respiratory signal from different locations. ROIs were chosen from the top, middle and bottom of the chest in horizontal locations on the far-right (a), mid-right (b), mid-left (c) and far-left (d) of the chest.

processing steps described in section 6.4.3 were used to compute the respiratory signal. A 6th order IIR Butterworth band-pass filter with pass-band limits of 0.1 Hz (6 breaths/minute) and 0.7 Hz (42 breaths/minute) was used to remove any cardiac frequency components and low-frequency noise. Zero-phase filtering was implemented using the MATLAB [2016] *filtfilt* function.

We calculated the amplitude of the waveforms in each respiratory signal, using the peak detection algorithm described in section 6.4.3 to find the onsets and peaks of the respiratory waveform. The amplitude of each respiratory waveform was calculated as the difference in signal value between the peak and onset of the waveform.

7.3.3 Respiratory signal visualisation

Figure 7.5 shows the respiratory signals extracted from ROIs on the far-right, mid-right, mid-left and far-left of the chest. We present the respiratory signals in the grid formation shown in figure 7.4 to allow comparison. Likewise, figure 7.6 shows the waveform amplitudes of the respiratory signals extracted from ROIs on the far-right, mid-right, mid-left and far-left of the chest.

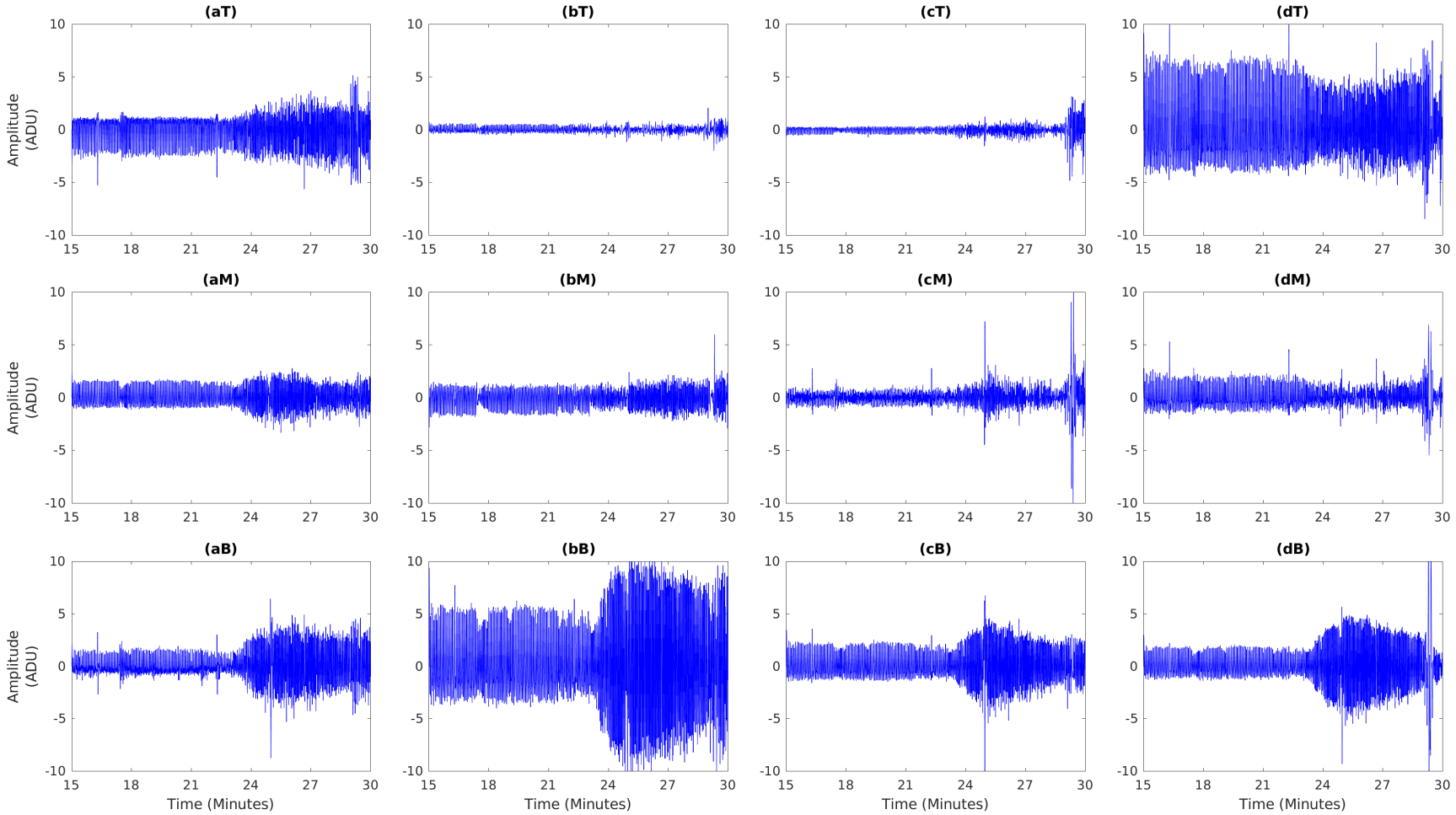


Figure 7.5: Respiratory signals during the last 15 minutes of data segment NVSM15-3 extracted from ROIs located horizontally on the far-right (a), mid-right (b), mid-left (c) and far-left (d) of the patients chest, at vertical locations at the top (T), middle (M) and bottom (B) of the chest. The location of the ROIs was shown in figure 7.4 and the location of the plots mirrors the grid formation.

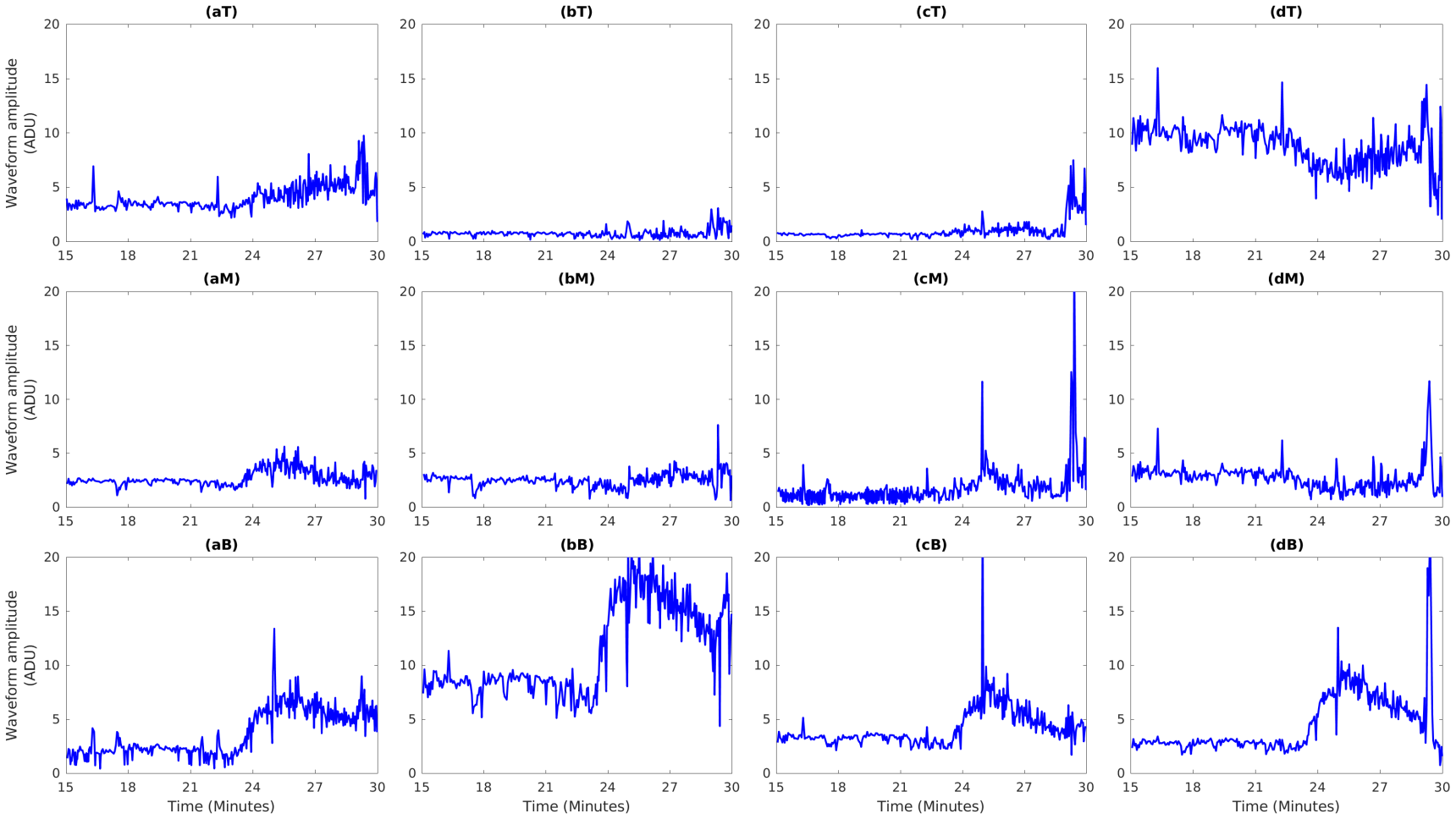


Figure 7.6: The waveform amplitudes of the respiratory signals during the last 15 minutes of data segment NVSM15-3 extracted from ROIs located horizontally on the far-right (a), mid-right (b), mid-left (c) and far-left (d) of the patients chest, at vertical locations at the top (T), middle (M) and bottom (B) of the chest. The location of the ROIs was shown in figure 7.4 and the location of the plots mirrors the grid formation.

7.3.4 Discussion

Figure 7.5 shows that camera-derived respiratory signals vary according to the location of the ROI from which they were extracted. For example, the signal extracted from the top of the far-left side of the chest (figure 7.5dT) had a much larger signal variance (oscillating between -5 to 8 ADU) than the signal derived from the middle of the far-left side of the chest (figure 7.5dM, oscillating between -1 to 2 ADU). We calculated the amplitude of the respiratory signal waveforms to quantify the signal variance, as shown in figure 7.6. The waveform amplitude calculated from respiratory signals extracted from different ROIs varied in terms of the baseline waveform amplitude and the amount the amplitude changed throughout the 15 minutes. For example, the waveform amplitude in figure 7.6bT was a constantly low value (1 ADU) while the waveform amplitude in figure 7.6bB began at approximately 9 ADU before increasing to 20 ADU at around the 25th minute.

The respiratory signal variance is a measure of the level of respiration-induced motion. The motion causes the pixel intensities within each ROI to change value as the chest moves closer to or further away from the video camera, or as different parts of the chest move in and out of the ROI. When there is more respiration-induced motion this effect is larger, which causes the level of signal variance to be higher. This can be confounded by differences in the respiratory signal due to imaging effects. For example, parts of the chest will be exposed to different levels of lighting, and will have different angles relative to the camera. This can affect the sensitivity of the ROI signal to changes in motion, as ROIs with more contrast (such as those containing an edge on the patient's body or clothing) will change pixel intensity more with motion. For example, the signal extracted from the ROI at the top of the far-left of the chest (figure 7.5dT) had higher variance compared to the signal extracted from ROIs below or to the right, which can be explained by its position on the edge of the patient's clothing, as shown in figure 7.4. However, imaging effects do not explain all of the variance between respiratory signals as the temporal changes that can be seen occurred within the same ROI at the same location, and these changes were more pronounced in some ROIs than others. Furthermore, these effects will stay relatively constant unless the patient moves position (the video footage showed that the patient did not move position until the final minute). It is therefore likely that the camera detected true physiological differences in respiratory motion in different locations.

In section 7.2.1 it was shown that the respiratory rate estimates derived from different ROI locations on either side of the chest (and the forehead) were almost

identical, with low errors between estimates. Thus, the respiratory signal amplitude may be affected by the ROI location even if the rate of respiration is not. When the patient increased her respiratory rate in the final 7 minutes of the segment (figure 7.3), the amplitude of the respiratory waveforms changed, which is most noticeable in the signal derived from the bottom of the middle-right side of the chest (figure 7.5bB) with the waveform amplitude increasing from 10 to 20 ADU. The respiratory signals in figure 7.5dT-dM had a slight decrease in waveform amplitude during this time, but to a lesser degree (1-5 ADU). Interestingly, respiratory rate changes were not necessarily always accompanied by significant concomitant changes in respiratory signal amplitude. For example there is a brief rise in respiratory rate around minutes 17 to 18 in figure 7.3, which is accompanied by very small changes in respiratory signal amplitude (maximum of 2-3 ADU in figure 7.5bB).

The increase in respiratory signal amplitude at minute 24 suggests that the patient took larger breaths causing an increase in respiratory motion. In effect, the camera therefore detected the total effort of respiration both in terms of the rate and size of chest movements as the patient breathed. This is additional information to the respiratory rate, since other increases in respiratory rate (such as at minute 18) did not also have a concomitant increase in respiratory amplitude (so changes in rate and amplitude can occur independently). The combined measurements of respiratory amplitude and respiratory rate from the video camera may be a useful indicator of the respiratory distress that the patient was suffering.

Previous authors have used depth-sensing cameras to demonstrate large differences in depth-of-motion at the diaphragm and at the chest between periods of chest breathing and periods of diaphragmatic breathing (Tahavori et al. [2014]). Here we demonstrate that variations in motion can also be detected with a monochromatic video camera (albeit without absolute depth values). We found that respiratory signals extracted from ROIs located at the bottom of the chest typically had larger waveform amplitudes than those extracted from ROIs located at the top or middle of the chest. Furthermore, ROIs located at the bottom of the chest were more sensitive to changes in respiratory motion. In this instance, it is unclear whether the variations were caused by normal respiratory mechanisms or the underlying respiratory pathology of the patient (e.g. right-sided pneumothorax). Further investigation of the spatial variations in respiratory signal amplitude are described in the sections of this chapter that follow.

7.4 Spatial variations in respiration

7.4.1 Overview

The previous analysis revealed that the amplitude of the respiratory signal depends on the location of the ROI from which the signal was extracted, even if the associated respiratory rate does not. Respiratory signals from some ROIs had a constant amplitude throughout the segment, while signals from other ROIs were highly sensitive to changes in respiratory motion. In particular, signals extracted from ROIs at the bottom of the chest had an increase in respiratory waveform amplitude that occurred with a concomitant rise in respiratory rate of 10-15 breaths/minute (23rd minute). Interestingly, there was also a brief rise in respiratory rate that did not coincide with an increase in respiratory waveform amplitude (17th minute).

In the following analysis, we present an investigation of the amplitude of the respiratory signal in different spatial locations, at three different time-points during the NVSM15-3 data segment. We assessed how the respiratory amplitude changed as the location of the ROI moves horizontally and vertically, and also how the respiratory amplitude is distributed across the right and left sides of the chest. Our aim was to assess potential asymmetries in respiratory motion.

7.4.2 Respiratory signal extraction

For the analysis of spatial variations in respiratory amplitude, we extracted ROI time-series from multiple ROIs on the chest. To analyse how the respiratory signal amplitude varied as a function of horizontal location on the chest, we extracted ROI time-series from every overlapping 30x30 pixel ROI in three horizontal rows, which are shown in figure 7.7. All 33 possible overlapping ROIs were selected in each row. Furthermore, to analyse the respiratory signal amplitude variation as a function of vertical location, we extracted ROI time-series from every overlapping 30x30 pixel ROI in four vertical columns, which are shown in figure 7.8. All 15 possible overlapping ROIs were selected in each column.

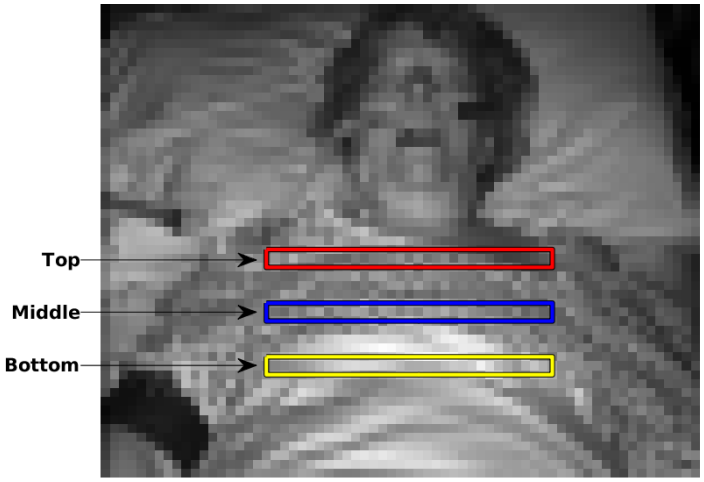


Figure 7.7: Positions of the horizontal rows of overlapping ROIs on the top, middle and bottom of the chest, from which ROI time-series were extracted to assess how respiratory signal amplitude varied with horizontal position. We extracted ROI time-series from 33 overlapping ROIs in each row. The rows shown are $33 \times 10 = 330$ pixels wide and 30 pixels high (the border is on top of the outer *super pixels*).

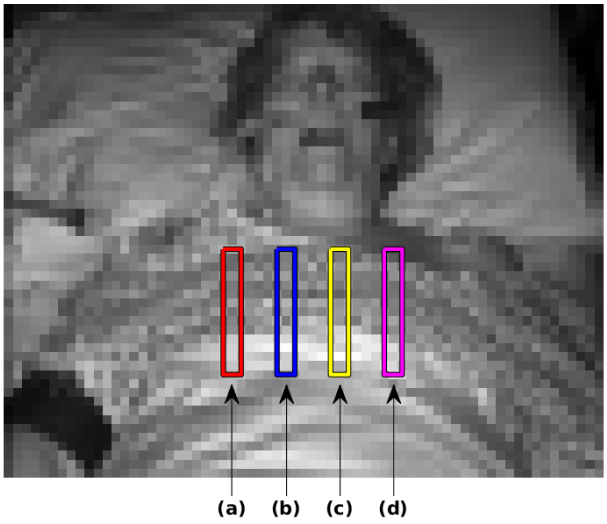


Figure 7.8: Positions of the columns of overlapping ROIs on the far-right (a), mid-right (b), mid-left(c), and far-left (d) of the chest, from which ROI time-series were extracted to assess variations in respiratory amplitude with vertical position. We extracted ROI time-series from 15 overlapping ROIs in each column. The columns shown are 30 pixels wide and $15 \times 10 = 150$ pixels high (the border is on top of the outer *super pixels*).

We also consider the distribution of the respiratory signal amplitudes on either side of the chest. To do so, we extract ROI time-series from every possible overlapping

30x30 ROI in 150x150 square regions on either side of the chest. The locations of the square regions are shown in figure 7.9. All $15 \times 15 = 225$ possible overlapping ROIs were selected in each region.

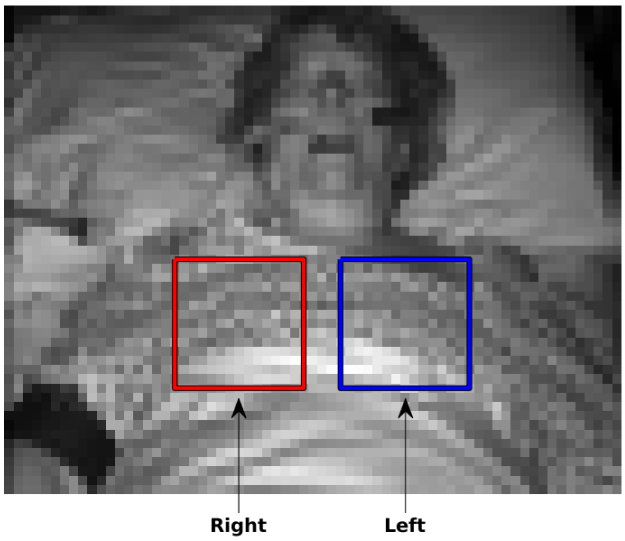
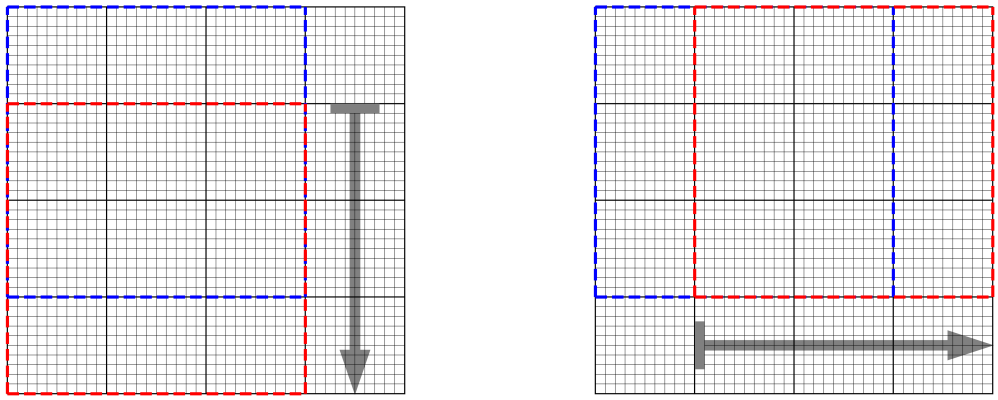


Figure 7.9: Location of the 150x150 regions on the right (a) and left (b) sides of the chest. ROI time-series were extracted from all 225 possible 30x30 overlapping ROIs in each region.

The selection of overlapping 30x30 ROIs is shown schematically for the vertical (figure 7.10a) and horizontal (figure 7.10b) arrangements of ROI locations. The first ROI is shown in blue and the next overlapping ROI is shown in red. As described in section 6.1.2, the blurring algorithm applied to the video data created a format in which each pixel in the blurred video frame was the sum of a 10x10 square grid of the original video pixels (that we call a *super pixel*). Consequently, the minimum amount by which the ROIs could overlap was every 10th pixel.

We extracted data at three time-points during segment NVSM15-3. For each time-point, 1-minute signals were extracted. We selected the time-points to represent three distinct respiratory patterns noticeable in the respiratory signals presented in section 7.3. The location of the three time-points relative to the non-contact respiratory signal (computed from a 30x30 ROI on the chest) are shown in figure 7.11. The first time-point, S1, was chosen from minutes 15 to 16 as a time with a relatively steady respiratory rate (approximately 16 breaths/minute) and a low respiratory signal amplitude (as shown in figure 7.6aT). The second time-point, S2, was chosen from minutes 17.2 to 18.2 as a time with an increase in respiratory rate (of approximately 5 breaths/minute) and a low respiratory signal amplitude (as shown in figure 7.6bB). The final time-point, S3, was chosen from minutes 26 to 27 as the time of peak



(a) **(b)**

Figure 7.10: Selection of overlapping ROI locations relative to the original pixels and the 10x10 *superpixels* (outlined) used in the blurred video frames. The locations of the first (blue) and second (red) overlapping 30x30 ROIs are shown.

respiratory rate (between 27 to 28 breaths/minute) and a high respiratory signal amplitude (as shown in figure 7.6bB).

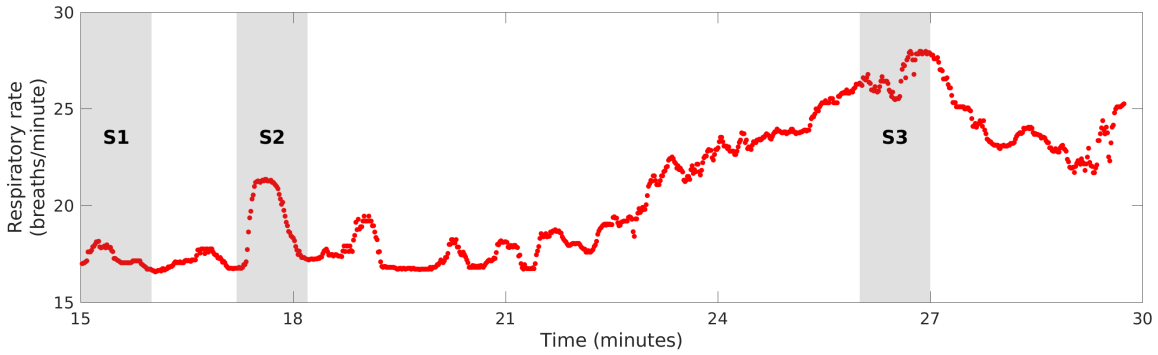


Figure 7.11: The locations of the three time-points extracted during segment NVSM15-3 for analysing the spatial variations in respiratory signal amplitude. The camera-derived respiratory estimates were computed from a 30x30 pixel ROI on the chest. The three time-points were extracted from minutes 15 to 16 (S1), 17.2 to 18.2 (S2), and 26 to 27 (S3), and are annotated in their respective locations.

For each ROI location in figures 7.7 to 7.9 and time-point in figure 7.11, the 1-minute section of raw ROI time-series was extracted. The respiratory signal was then computed by applying a 6th order IIR band-pass filter (as described in section 6.4.3 and section 7.3.2). We again used the respiratory peak detection algorithm described in section 6.4.3 to find the onsets and peaks of the respiratory waveform and calculated the waveform amplitude as the difference in signal value between the waveform peak and onset. The respiratory signal amplitude for the given ROI and time-point was then

calculated as the median of the waveform amplitudes within the 1-minute section of data. We analyse the spatial variations in respiration by visualising how the respiratory amplitude changes with ROI location in each row or column (results in section 7.4.3).

Respiratory amplitude distributions

We estimated the distributions of respiratory amplitudes computed from the overlapping 30x30 ROIs within the 150x150 regions on either side of the chest (figure 7.9). We used non-parametric Gaussian kernel density estimation with log-transformation to bound support to positive values (Charpentier and Flachaire [2015]). The density estimate was computed as follows:

$$f(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\log(x)-\log(X_i)}{\sigma}\right)^2\right] \tag{7.1}$$

where $f(x)$ is the density estimate at the data point x, n is the size of the dataset X (of waveform amplitudes), and σ is the kernel bandwidth.

When computing the kernel density estimation, we used a kernel bandwidth σ calculated with *unbiased cross-validation by least squares* (Charpentier and Flachaire [2015]), which we described in chapter 5. To do so, we used the MATLAB [2016] constrained optimisation function *fminbnd* (based on the algorithm of Brent [2013] using golden section search with inverse parabolic interpolation) to minimize the objective function described in equation (5.3), which we repeat as follows.

$$g(\sigma) = \int f^2(x)dx - \frac{2}{n}\sum_{i=1}^{n} f_{-i}(X_i) \tag{7.2}$$

where g is the objective function for values of the kernel bandwidth σ. f is the kernel density estimation function described in equation (7.1) and f_{-i} is the kernel density estimation function based on the dataset containing all of the observations except for X_i. We constrained *fminbnd* to search for the minimum of g for $0.01 < \sigma < 5$, which we empirically found to reliably contain the global minimum.

7.4.3 Visualising spatial variations in respiratory amplitude

Figures 7.12 to 7.14 show the spatial variations of respiratory amplitude for time-points S1, S2, and S3, respectively. These include the distribution of respiratory amplitudes on either side of the chest. Figures 7.12 to 7.14 also show the variations in respiratory amplitude over the vertical and horizontal positions described in figures 7.7 and 7.8.

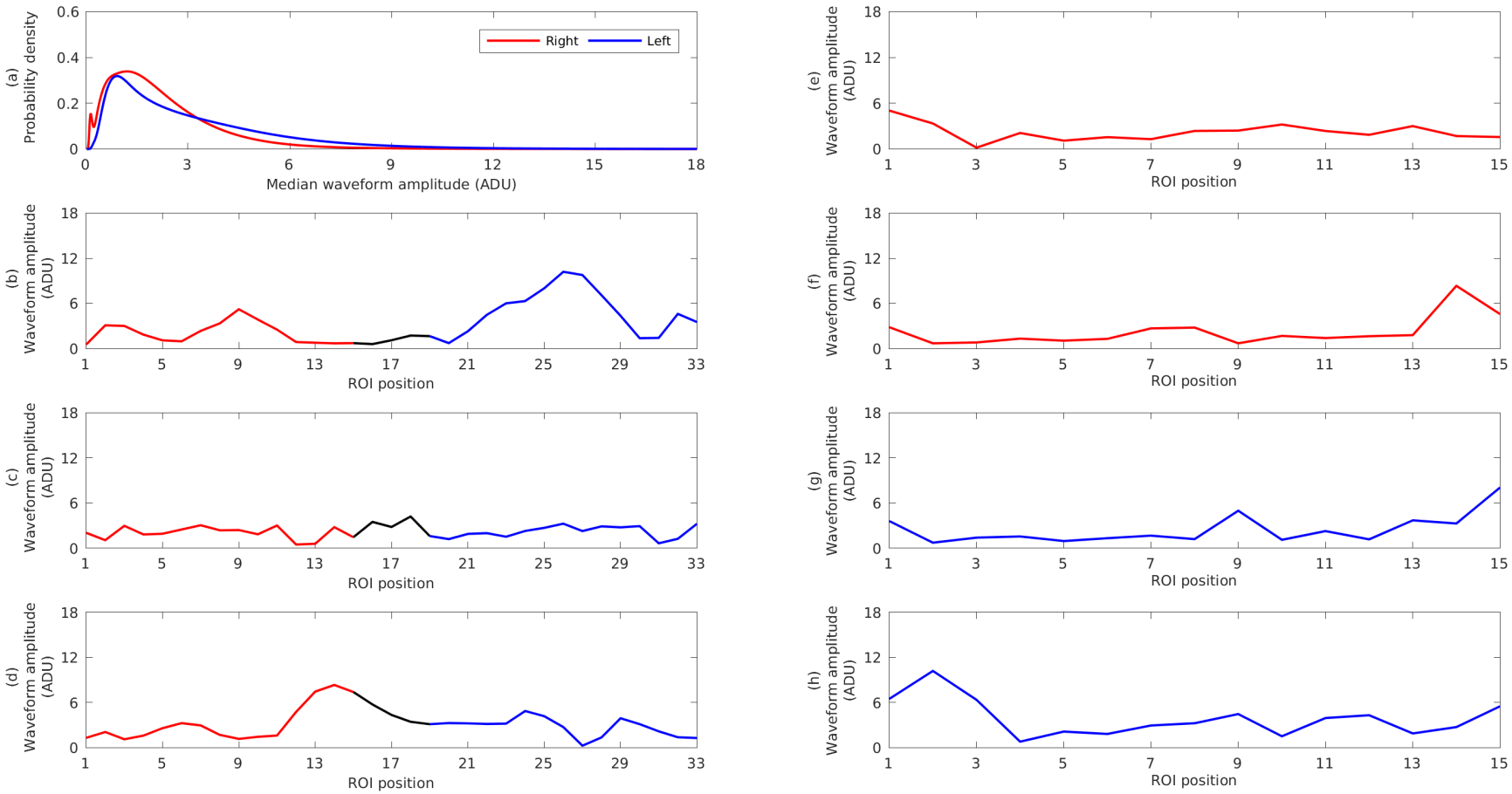


Figure 7.12: Spatial variations in respiratory signal amplitude at time-point S1. The distributions of respiratory signal amplitudes on the right and left sides of the chest (see figure 7.9) are shown (a). Variations in respiratory signal amplitude with horizontal location are shown for the rows along the top (b), middle (c) and bottom (d) of the chest (see figure 7.7). Variations in respiratory signal amplitude with vertical location are shown for columns down the far-right (e), mid-right (f), mid-left (g) and far-left (h) of the chest (see figure 7.8). Data points from the right (red), left (blue) and centre (black) are coloured accordingly.

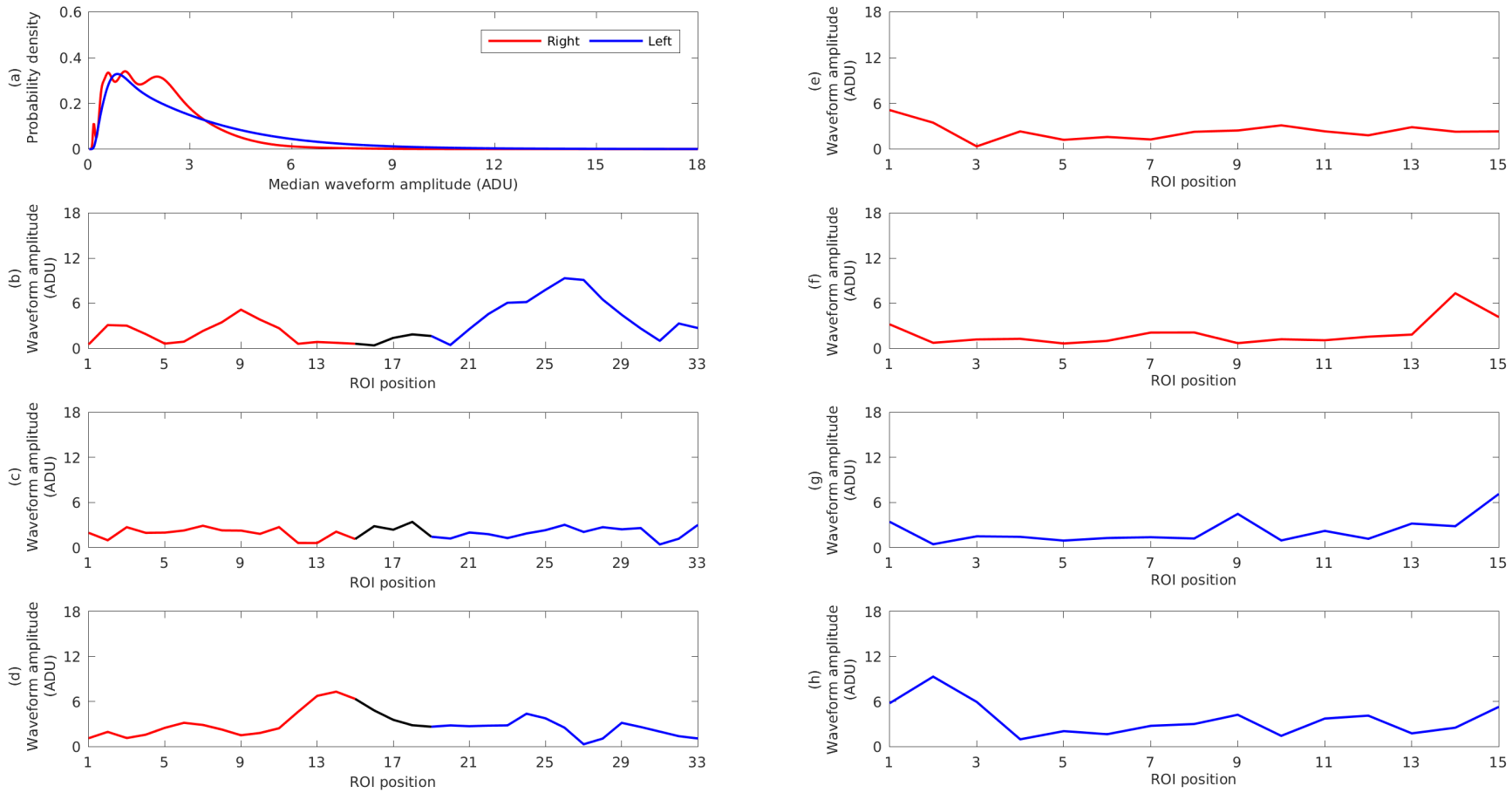


Figure 7.13: Spatial variations in respiratory signal amplitude at time-point S2. The distributions of respiratory signal amplitudes on the right and left sides of the chest (see figure 7.9) are shown (a). Variations in respiratory signal amplitude with horizontal location are shown for the rows along the top (b), middle (c) and bottom (d) of the chest (see figure 7.7). Variations in respiratory signal amplitude with vertical location are shown for columns down the far-right (e), mid-right (f), mid-left (g) and far-left (h) of the chest(see figure 7.8). Data points from the right (red), left (blue) and centre (black) are coloured accordingly.

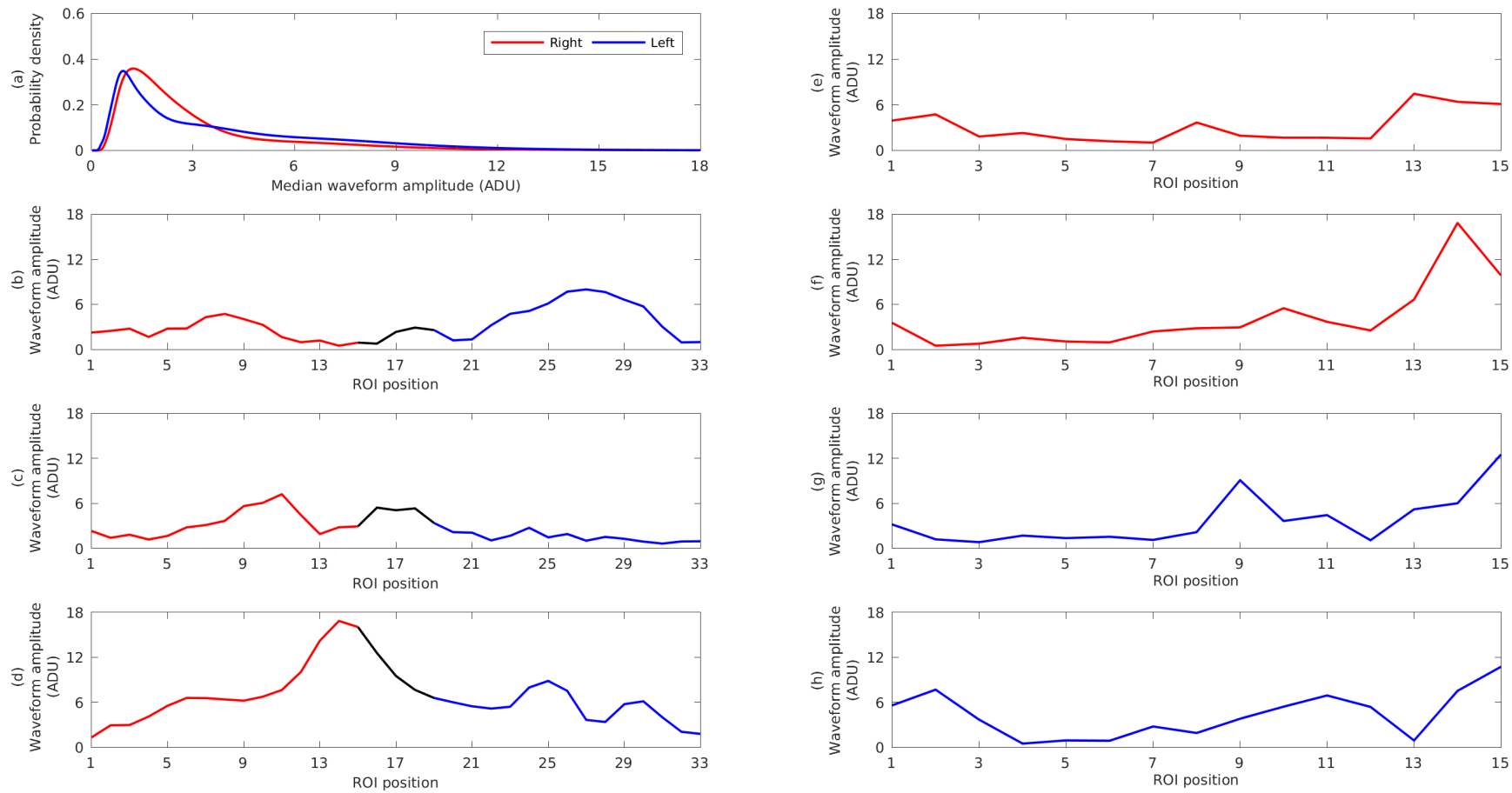


Figure 7.14: Spatial variations in respiratory signal amplitude at time-point S3. The distributions of respiratory signal amplitudes on the right and left sides of the chest (see figure 7.9) are shown (a). Variations in respiratory signal amplitude with horizontal location are shown for the rows along the top (b), middle (c) and bottom (d) of the chest (see figure 7.7). Variations in respiratory signal amplitude with vertical location are shown for columns down the far-right (e), mid-right (f), mid-left (g) and far-left (h) of the chest (see figure 7.8). Data points from the right (red), left (blue) and centre (black) are coloured accordingly.

7.4.4 Discussion

The analyses presented in figures 7.12 to 7.14 demonstrate spatial variations in respiratory signal amplitude. At time-point S1 (figure 7.12), the respiratory signal varied from almost zero amplitude in ROIs with little chest movement, to amplitudes of up to 10 (ADU) in ROIs with higher levels of chest movement. In general, there was more horizontal variation in signal amplitude than vertical variation, although there was almost no variation in respiratory signal amplitude from the middle of the chest (see figure 7.12c,g). There were two locations of high signal amplitude, one at the centre of the bottom of the chest (figure 7.12d) and one at the top left of the chest (figure 7.12b), which was located on the edge of the patient's clothing as discussed in section 7.3.4. The distributions of respiratory signal amplitudes were similar on both sides of the chest, which suggests that the right-sided pneumothorax did not affect the respiratory amplitude in terms of creating an asymmetry between the two sides of the chest.

The major variations in respiratory signal amplitudes at time-point S2 (figure 7.13) occurred at the same locations as those at time-point S1, although the values were slightly lower in some locations (e.g. ROI positions 13-17 in figure 7.13d). Interestingly, the respiratory rate at time-point S2 was elevated at approximately 5 breaths/minute higher than the respiratory rate at time-point S1. Thus, the patient elevated her respiratory rate with very little associated changes in the range of chest movement.

There were major spatial variations in respiratory signal amplitudes at time-point S3 (figure 7.14), which was when the patient had a highly elevated respiratory rate of approximately 26-28 breaths/minute. The locations of highest respiratory amplitude were the same as at time-points S1 and S2 – the lower part of the chest and along the edge of the patient's clothing – although the amplitudes in this case were 1.5-2 times as large. The respiratory amplitudes along the bottom of the chest in particular had increased since the previous time-points. The patient had therefore increased her chest motion significantly at the same time as increasing her respiratory rate. Again, the distributions of respiratory amplitudes in either side of the chest were similar, suggesting that the increase in chest motion applied evenly to both sides of the chest.

The spatial differences in respiratory signal amplitude are likely to be due to both imaging reasons and physiological reasons. As discussed in section 7.3.4, many small spatial variations in respiratory signal amplitude can be attributed to the fact that some ROIs are located in areas for which the camera can more easily detect respiratory motion, due to the presence of edges, the location of lights, or the angle of

the camera. However, the increase in respiratory amplitude at time-point S3 suggests that there can be large changes in amplitude during times of respiratory distress, for physiological reasons such as the patient taking larger breaths or demonstrating an abnormal pattern of breathing. This has more of an effect in some ROI locations than others. The right-sided pneumothorax seems to have little effect on the respiratory amplitude at any of the three time-points in terms of the symmetry of amplitudes on either side of the chest. It is unclear whether the diaphragm had ruptured at this point, and indeed what effect this would have in any case. However, it is noteworthy that the respiratory signal from ROIs closer to the diaphragm have a large amplitude and undergo the most change between time-points S2 and S3.

7.5 Respiratory amplitude maps

7.5.1 Overview

The analyses presented in this chapter have established that the respiratory amplitude in camera-derived respiratory signals can change during times of abnormal respiratory physiology. Furthermore, the spatial location of the ROI affects the respiratory signal amplitude, and the extent to which this changes as a result of abnormal respiratory physiology. In the final section of this chapter we present a method to map the signal amplitude to the location of the ROI in the video frame. We present the video frame at the time-points S1, S2 and S3, with the respiratory amplitude for each ROI mapped to the position on the chest.

In order to explore whether the spatial and temporal variations in respiratory signal amplitude may have been related to the ruptured diaphragm and liver herniation suffered by NVSM15, we also present respiratory amplitude maps at the time of the two X-rays described in figure 7.1. The patient's diaphragm was intact during the first X-ray and was ruptured during the time of the second X-ray. It is unclear whether the patient's diaphragm was intact at the three time-points (S1, S2 and S3) analysed during segment NVSM15-3.

7.5.2 Data segments

Following section 7.4, we used data from the time-points S1, S2 and S3. We extracted ROI time-series from multiple ROIs on the chest. We did the same for two additional time-points taken before the penultimate X-ray and the final X-ray, as follows.

Time-point S0 was chosen as the section of video that immediately preceded the first X-ray, when the patient's diaphragm was intact. The reference vital signs (from the Philips bedside monitor) around this time-point are shown in figure 7.15, and were mostly stable with minimal changes in heart rate and SpO_2, and a brief elevation in respiratory rate (although this may have been due to signal artefact). The reference heart rate at time-point S0 was 83 beats/minute; the reference respiratory rate was 20 breaths/minutes and the reference SpO_2 was 95%.

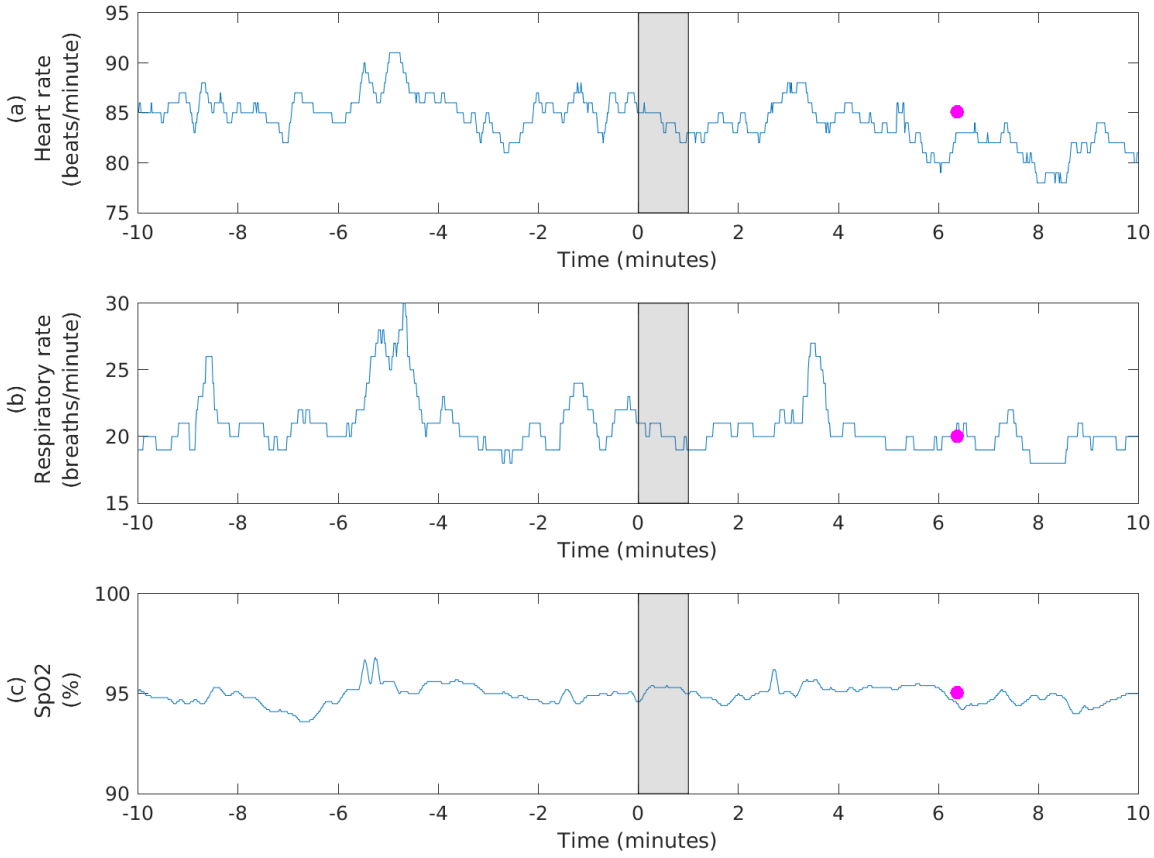


Figure 7.15: Reference vital-sign values for heart rate (a), respiratory rate (b) and SpO_2 (c) for patient NVSM15 10 minutes before and after the time-point S0 (shaded in grey). Manual vital-sign estimates are shown as pink dots and bedside monitor estimates are shown as blue lines. This segment immediately preceded a chest X-ray that confirmed the presence of a right-sided pneumothorax, and that the patient's diaphragm was intact. NVSM15 was a 79 year-old female patient recovering from gynaecological cancer surgery.

Time-point S4 was chosen as the section of video that immediately preceded the final X-ray, when the patient's diaphragm was found to have been ruptured. The reference vital signs around this time-point are shown in figure 7.16, and were mostly stable although with relatively high heart rate values (90 beats/minute) and low SpO_2

values (fluctuating between 85 to 90%). There was no reference respiratory rate as the patient's ECG electrodes had been removed (so that there was no impedance pneumography signal), and so estimates were derived from ROIs on the head and the chest. The camera-derived estimates from two independent locations matched, thus providing confidence that these values were indeed correct (as discussed in section 7.2.1). The respiratory rate was stable at approximately 15 breaths/minute.

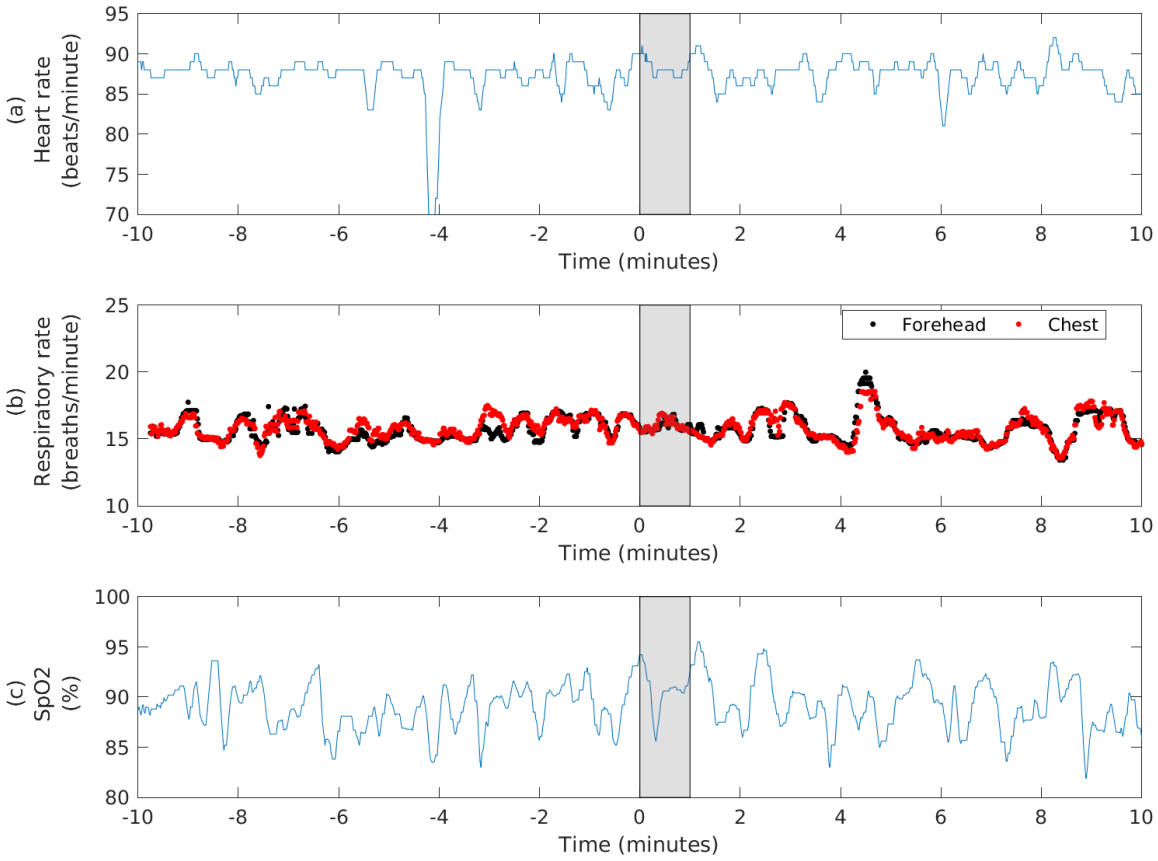


Figure 7.16: Reference vital-sign values for heart rate (a), respiratory rate (b) and SpO_2 (c) for patient NVSM15 10 minutes before and after the time-point S4. There was no reference respiratory rate as the patient's ECG electrodes had been removed (so that there was no impedance pneumography signal), and so the camera-derived values are shown from ROIs on the forehead and the chest (see section 7.2.1). This segment immediately preceded a chest X-ray that confirmed the presence of a right-sided pneumothorax and a diaphragmatic rupture. NVSM15 was a 79 year-old female patient recovering from gynaecological cancer surgery.

7.5.3 Computing respiratory amplitude

For this analysis, we used a similar methodology as described in section 7.4.2 to map the respiratory rate amplitude at each time-point. First, a large rectangular area was

defined as the outline of the chest. All possible 30x30 ROIs were extracted within the rectangular area, with ROIs overlapping as shown in figures 7.10a and 7.10b. For each ROI, the method described in section 7.4.2 was used to calculate the signal amplitude at a given time-point.

To map the signal amplitude back to the position on the patient's chest, we first defined the location of the 30x30 ROI as the location of the 10x10 *super pixel* at its centre. We then chose a colour map from dark blue for amplitude values of 0 and dark red for amplitude values of 15 to allow the amplitude values to be visualised with sufficient resolution within the range of respiratory signal amplitudes found in section 7.4. Each *super pixel* in the video frame was then coloured according to the amplitude of the respiratory signal computed from the 30x30 ROI centred on the *super pixel* (see figures 7.10a and 7.10b). The colour maps were smoothed by convolving a 2D Gaussian filter over the coloured *super pixel* values within the rectangular bounding box. The Gaussian filter had a filter size of 3x3 (*super pixels*) and a kernel width of 0.5 such that the filter only operated on neighbouring *super pixels*.

7.5.4 Results

The respiratory signal amplitude map for time-point S0 is shown in figure 7.17. The maps for time-points S1, S2 and S3 during segment NVSM15-3 are shown in figures 7.18a to 7.18c, respectively. Finally, the respiratory signal amplitude map for time-point S4 is shown in figure 7.19.

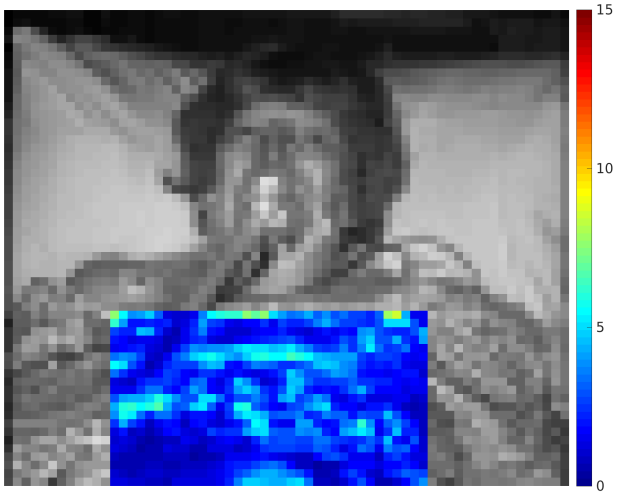
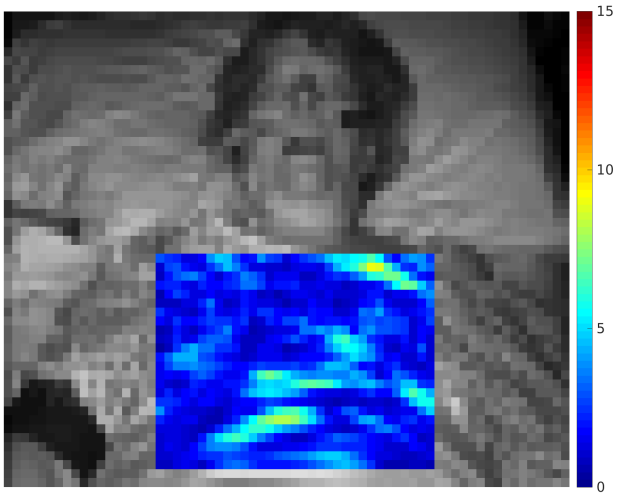
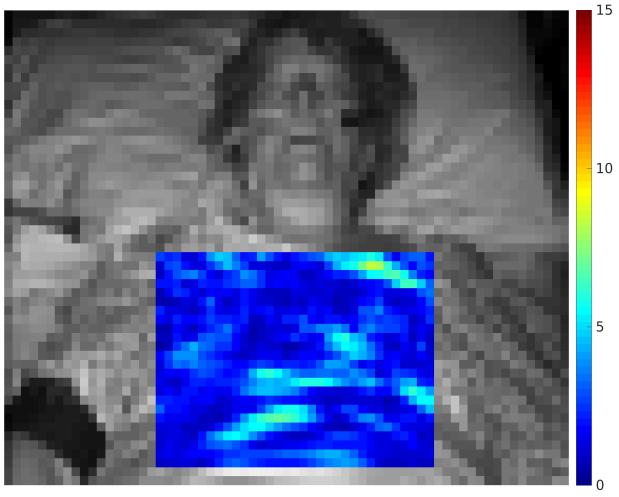


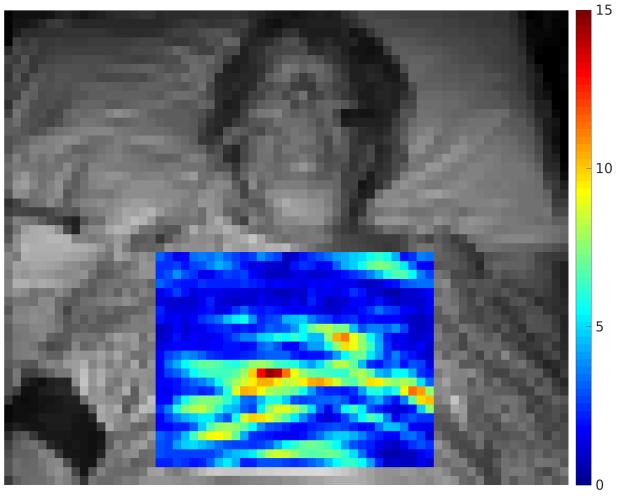
Figure 7.17: Colour map of respiratory signal amplitudes from ROIs on the chest of patient NVSM15 during time-point S0. The patient had a right-sided pneumothorax, but still had an intact diaphragm, as confirmed by chest X-ray.



(a)



(b)



(c)

Figure 7.18: Colour map of respiratory signal amplitudes from ROIs on the chest of patient NVSM15 during time-points S1 (a), S2 (b) and S3 (c). The patient had a right-sided pneumothorax (confirmed by the X-rays around time-points S0 and S4). It is unclear whether the diaphragm had ruptured by the time the data were recorded.

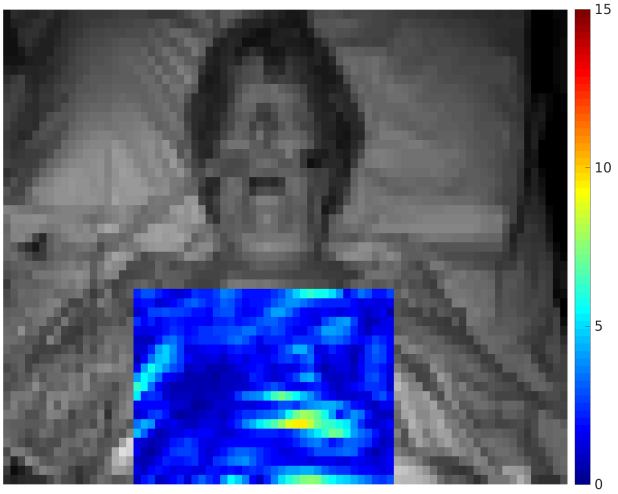


Figure 7.19: Colour map of respiratory signal amplitudes from ROIs on the chest of patient NVSM15 during time-point S4. The patient had a right-sided pneumothorax and a ruptured diaphragm, as confirmed by chest X-ray.

7.5.5 Discussion

The respiratory amplitude map presented for time-point S0 (figure 7.17) shows that the respiratory signal amplitude was relatively uniform throughout the chest. Some small differences in amplitude existed, but these were likely due to edges or areas of high contrast on the patient's clothing. The camera angle at the time of recording was such that the lower part of the chest was out of the video frame, thus restricting ROIs to be chosen in the upper part of the chest.

Figures 7.18a to 7.18c confirm the results presented in section 7.4, showing that there are clear variations in respiratory signal amplitude from different ROIs on the chest during those time intervals. The spatial patterns in respiratory signal amplitude are very similar between time-points S1 (figure 7.18a) and S2 (figure 7.18b), with areas of increased amplitude near the stomach and the edge of clothing near the patient's collarbone. The signal amplitude is relatively low during S1 and S2, which is consistent with earlier findings in this chapter. As expected, the respiratory map in figure 7.18c shows areas of elevated respiratory amplitude around the bottom of the patient's chest, particularly along the line of the lower ribs. The respiratory signal amplitude is up to 5 times as high in these areas compared to the upper part of the chest, which has a much lower range of motion due to respiration.

The respiratory amplitude map at time-point S4 (figure 7.19) showed that most locations on the chest had a low amplitude of motion due to respiration. A small area of increased respiration amplitude occurred around the lower ribs on the left side of

the chest. Other small areas of slightly elevated respiratory signal amplitude occurred on edges of the patient's clothing.

From the respiratory maps presented in this analysis, it is clear that the period of elevated respiration during NVSM15-3 (time-point S3) coincided with an increase in respiratory motion. Physiologically, this may represent an increase in respiratory effort, as the patient increased both her rate of breathing and the amplitude of breaths taken, presumably driven by a need for increased oxygen intake. Continuous respiratory rate monitoring devices do not usually provide information about the total effort of respiration exerted by the patient or the locations of increased motion, so the ability to monitor both chest motion and respiratory rate may be a novel clinical contribution of a future camera-based vital-sign monitoring system. To our knowledge, previous methods of computing spatial maps of respiration have used depth-sensing cameras (Ortmüller et al. [2015]; Tahavori et al. [2014]; Wentz et al. [2012]; Wijenayake and Park [2017]). Here we demonstrate that the same monochromatic video camera used to derive non-contact estimates of heart rate and respiratory rate can also be used to map spatial variations in respiratory motion.

It is unclear what effect the two underlying respiratory pathologies – a right-sided pneumothorax and a ruptured diaphragm – may have had on the spatial patterns of respiratory amplitude. It is possible that a ruptured diaphragm could cause an increase in motion around the ribs. Respiratory motion is usually driven by the diaphragm and the intercostal muscles, which are attached to the ribs. Without the effect of the diaphragm, the movement around the ribs caused by the intercostal muscles may be emphasised. The respiratory motion around the ribs is less pronounced at time-point S4, which is when the ruptured diaphragm had been confirmed. While this may provide evidence against the notion that a ruptured diaphragm would cause a visible increase in respiratory amplitude around the ribs, it may equally be the case that this motion can only be seen during times of increased respiration such as at time-point S3. The respiratory rate at time-point S4 was very steady at 15 breaths/minute, and unfortunately there was no data available for further exploration.

The methods presented in this chapter are limited by the variation in respiratory signal amplitude caused by imaging effects such as the presence of edges, the camera angle, or lighting. In all of the respiratory maps shown in figures 7.17, 7.18a to 7.18c and 7.19, contours of the patient's clothing can be seen as slight variations in respiratory angle. Thus, these methods need further development to account for these sources of variation so that the physiologically relevant spatial differences in chest motion may be modelled accurately. However, significant changes in respiratory signal amplitude

are unlikely to be caused by such imaging effects, so this method may still be valid for detecting large changes in respiratory motion (such as at time-point S3).

This analysis is limited by the inclusion of data from only one patient. To further explore the potential spatial variations in camera-derived respiratory signals, segments of video data need to be extracted from a larger sample of patients, particularly during times of abnormal respiration. This analysis suggests that the extracted segments should come from patients with respiratory pathologies, as periods of normal respiration are unlikely to reveal interesting spatial variations and thus will not be particularly useful for developing this method further.

7.6 Conclusions

Non-contact vital-sign monitoring with a video camera can be used to detect both spatial and temporal variations in respiratory signals. We presented a case study of a patient that experienced significant respiratory distress due to a right-sided pneumothorax and a ruptured diaphragm as diagnosed by clinical experts. Changes in respiratory signal amplitude can be seen in different locations on the chest, which may be symptomatic of the patient's respiratory pathology. This case study demonstrates a potential future use of the video camera for detecting unusual respiratory patterns. Further work is needed to develop methods for separating differences in respiratory motion from other sources related to imaging (e.g. presence of edges, changes in camera angle) from those related to physiology (e.g. breathing asymmetries caused by respiratory pathologies). The ability to map spatial variations in respiratory physiology could be an advantage of camera-based systems over traditional contact vital-sign monitoring techniques as an additional source of clinical information.

Chapter 8

Conclusions and future work

This chapter concludes the thesis, beginning with a review of the main results in section 8.1. We then discuss suggested directions for future work in section 8.2. Finally, we provide concluding remarks in section 8.3.

8.1 Thesis review

Chapter 1 reviewed the background of post-operative monitoring. The physiology of the vital signs were described, as well as manual and continuous methods of measurement. We then introduced the problem of adverse clinical events in post-operative care and current strategies for detecting physiological deterioration. We described how the introduction of continuous vital-sign monitoring technology to post-operative wards may improve patient monitoring and the detection of physiological deterioration, particularly through non-invasive methods such as camera-based vital-sign monitoring. Chapter 1 ended with an overview of the thesis structure.

In chapter 2 we described the CALMS datasets. The CALMS-2 dataset was collected in a clinical trial of ambulatory vital-sign monitoring in an upper-gastrointestinal post-operative ward in the Oxford University Hospitals NHS Trust. The CALMS-2 dataset consisted of vital signs from manual observations, and bedside and ambulatory monitoring equipment from 357 patients who recovered normally and 50 patients who suffered an adverse clinical event. We noted that patients had poor compliance with the continuous monitoring equipment, particularly during the time preceding adverse clinical events. This motivates the development of improved continuous monitoring technology, such as camera-based methods described in chapters 6 and 7. The CALMS-3 dataset was collected retrospectively from the same upper-gastrointestinal post-operative ward as the CALMS-2 dataset. The CALMS-3 dataset consisted of manual vital-sign observations collected retrospectively from 298 patients who recovered normally and 149 patients who suffered an adverse event. The CALMS-3 dataset was designed to have a 2:1 ratio between "normal" and "abnormal" patients, which allows the analysis of longitudinal patterns in vital signs preceding adverse clinical events.

In chapter 3 we assessed the value of features of diurnal vital-sign variability for improving the performance of an EWS for detecting physiological deterioration. We found that diurnal vital-sign variability features were highly predictive of physiological deterioration, particularly respiratory rate and SpO_2 variability. We trained a machine learning based EWS using logistic regression with restricted cubic splines interpolation. Both vital-sign and vital-sign variability features were included. The addition of variability features improved the model performance, achieving a higher AUROC, sensitivity and specificity than NEWS, which is the standard of care in UK hospitals.

Chapter 4 provided a method comparison study of continuous and manual vital-sign data. We tested whether there was a "smoothing effect" whereby manual observations are biased towards more normal values. We also assessed agreement between continuous and manual methods using Bland-Altman analyses, and tested for an "arousal effect" by the vital-sign taker. We found no evidence that a clinically significant "smoothing effect" exists for respiratory rate, heart rate or SpO_2 in acute post-operative care. This result differed from previous studies of the "smoothing effect" in the anaesthesia literature. We also found no evidence of an "arousal effect" caused by the vital-sign recorder. Differences between manually and continuously-recorded vital signs were frequently large, suggesting the methods cannot be used interchangeably.

We derived features of diurnal variability from continuous vital-sign data in chapter 5 for heart rate, respiratory rate and SpO_2. We tested several methods of computing diurnal variability based on percentiles of vital-sign values in the 24 hours before the time of observation. For each method, we computed distributions of diurnal variability values preceding an adverse clinical event (for "abnormal" patients) and preceding discharge (for "normal" patients). We assessed the diurnal variability computation method based on the statistical distance between the distributions of "normal" and "abnormal" data. We found that diurnal variability features from continuous data could be used to distinguish between "normal" and "abnormal" data, but also found that the low quantities of continuous vital-sign data during times of interest in the CALMS-2 dataset limited our analyses. We found that vital-sign variability features computed from manual data were more discriminative than features computed from continuous data, in terms of maximising the statistical distance between distributions of "normal" and "abnormal" data. This result reflects the need for high quality continuous monitoring data, and motivates the development of continuous monitoring methods that are acceptable for long-term monitoring with high compliance.

In chapter 6 we presented a clinical study that was conducted to assess the use

of extended non-contact monitoring in post-operative patients during an elective stay in the ICU. In this study, patients were monitored continuously with an optical video camera, a thermal video camera, and reference continuous vital-sign monitoring equipment for up to 48 hours. We investigated five segments recorded during this study, during which the patients experienced changes in physiology. We presented methodologies to derive heart rate and respiratory rate estimates from video camera data. When compared with reference vital-sign monitoring equipment, the non-contact heart rate estimates had a Mean Absolute Error (MAE) of 0.8 to 2.8 beats/minute, and the non-contact respiratory rate estimates had an MAE of 0.6 to 2.0 breaths/minute. We therefore demonstrated that clinically-relevant changes in vital-sign values can be accurately tracked with camera-based non-contact methods. Consequently, non-contact monitoring technology could be a solution for providing continuous vital-sign data for EWS computation in a post-operative setting.

Finally, in chapter 7 we presented a method for visualising respiratory amplitude in different regions over a patient's chest. We presented a case study of a patient that experienced significant respiratory distress due to a right-sided pneumothorax and a ruptured diaphragm as diagnosed by clinical experts. Changes in respiratory signal amplitude can be seen in different locations on the chest, which may be symptomatic of the patient's respiratory pathology. Changes in respiratory signal amplitude were also independent of changes in respiratory rate, suggesting that respiratory signal amplitude may be an additional source of respiratory information. Overall, respiratory rate and amplitude may be a measure of total respiratory effort. This case study demonstrates a potential future use of the video camera for detecting unusual respiratory patterns.

8.2 Future work

The work presented in this thesis provides many opportunities and important directions for future work to improve the timeliness and precision of detecting physiological deterioration in post-operative care. We split these into three related research themes: predictive modelling with EWSs, the provision of continuous vital-sign monitoring in an ambulatory setting, and non-contact vital-sign monitoring with video cameras. We focus on clinical applications and technical solutions.

8.2.1 Predictive modelling

Improving the sensitivity and specificity of EWSs remains an important area of biomedical engineering research. EWSs with low sensitivity miss too many instances

of deterioration, leaving patients at risk of suffering an adverse event, and EWSs with low specificity require a high workload with too many clinical visits. In chapter 3 we showed that a machine learning based EWS can perform better than traditional scores such as NEWS and ViEWS, in line with previous work such as Pimentel [2015]. Furthermore, we showed that the inclusion of diurnal vital-sign variability features can improve EWS performance. However, the models used in chapter 3 were simplistic, and the performance of machine learning based EWSs may be improved with more powerful modelling strategies.

A disadvantage of traditional EWSs and simple machine learning based EWSs is that they assume that each vital-sign observation is independent and identically distributed. We know that this assumption is incorrect, as the likelihood of a particular vital-sign value, and the interpretation of its importance, should be conditioned on the vital-sign values at previous time-points. The inclusion of vital-sign variability features encodes time-based information in the model and thus partially relaxes the assumption of independence, but this can be further extended with time-series based methods. The use of time-series methods such as hidden Markov models may improve EWS performance and therefore detect physiological deterioration with better sensitivity and specificity. Time-series models have been used for predictive modelling in the ICU (Li-wei et al. [2012]; Saria et al. [2010a]), and further work should be conducted to test their performance with vital-sign data collected during post-operative care.

The logistic regression model used in chapter 3 encoded the non-linear relationship between individual vital signs and the risk of deterioration using restricted cubic splines interpolation. However, the model still assumed that the relationships between the values of each vital-sign were linear. This assumption is incorrect as vital signs have complex non-linear interactions that are influenced by physiological mechanisms. The use of flexible non-linear modelling strategies that encode interactions between different features may therefore improve the performance of machine learning based EWSs. In particular, deep learning models may be able to achieve high performance for detecting deterioration, if there is sufficient data available to effectively train the models. However, as discussed in chapter 1, there is a trade-off between model complexity and model interpretability, and clinicians must be able to understand why model decisions are made so that the physiological abnormality can be appropriately addressed. Future work should investigate deep learning based EWS models and how they may be appropriately used and interpreted in a clinical setting.

Probabilistic Bayesian models should also be investigated for computing EWSs. Bayesian methods such as Gaussian processes allow prior knowledge to be encoded

into the model via distributions over parameters. Posterior distributions can then be computed by updating the prior knowledge with evidence from the available data. This may be used to compute patient-specific EWS models by starting with priors formed from population-level data and providing posterior updates based on a given patient's vital-sign dynamics. Patient-specific models may improve EWS performance by accounting for physiological differences between patients of different demographics, co-morbidities and surgery types. Bayesian methods also provide an estimate of model uncertainty, which could provide additional information to clinicians as they interpret the output of the model. However, probabilistic models such as Gaussian processes do not scale well to large datasets, and some vital signs such as SpO_2 are not easily modelled by common parametric probability distributions, making their implementation difficult. Further work on Bayesian EWS methods should be conducted with a focus on modelling appropriate probability distributions, and scalability.

Finally, improved methods for designing and evaluating EWS models should be developed. The current framework of training and testing EWSs as binary classifiers is clearly limited as vital-sign observations must be labelled as "abnormal" or "normal" based on simple heuristics, such as whether the observation was taken within 24 hours of an adverse event. This methodology is unlikely to reflect the true likelihood of an adverse clinical event, so model performance *in-silico* may not map directly to clinical care. Methods based on survival analysis, such as Cox proportional hazards models (Cox [1972]), may be a more natural framework for EWS model development. Furthermore, binary classification metrics such as the AUROC have known limitations in this setting (Romero-Brufau et al. [2015]). Future work should be conducted to develop robust methodologies for designing and evaluating EWS models such that results from computational experiments reflect the expected impact on clinical care.

8.2.2 Continuous vital-sign monitoring

Even the most high-performing EWS models will still be limited by the frequency of manual observations, as physiological deterioration may occur between clinical visits. In chapter 5 we demonstrated that continuous vital-sign monitoring equipment may be used to detect abnormal vital-sign values between times of manual observations. However, our analysis was limited by the lack of quality continuous vital-sign data. Post-operative patients are difficult to monitor clinically due to issues such as patient discomfort (from attached sensors), and a lack of ambulatory sensors that can monitor all 5 vital signs. Further work must be conducted to develop and test new methods of ambulatory monitoring that can reliably measure patients' vital signs for extended

periods of time.

Reliable continuous vital-sign data may allow the development of complex time-series based EWS models that encode dynamic signatures of patient physiology. In chapter 5 we explored the derivation of diurnal vital-sign variability features. Variability features may be derived over shorter time-windows with continuous data, and more complex methods than using percentiles or extremal values may be used. For example, Saria et al. [2010b] and Li-wei et al. [2015] developed methods for learning short-term dynamic features of deterioration from continuous monitoring data in critical care. Further work should explore methodologies for computing variability features from continuous vital-sign data in post-operative care, and experiments should be conducted to test their value for detecting physiological deterioration. To conduct this work, large datasets of high quality continuous monitoring data collected from post-operative patients will be required.

8.2.3 Non-contact vital-sign monitoring

In chapter 6 we estimated heart rate and respiratory rate from video camera data during clinically important changes in physiology. Non-contact vital-sign monitoring may allow continuous data to be measured in a post-operative setting for extended periods of time. However, there are many technical challenges that must be overcome before video cameras can be implemented as the basis for a feasible stand-alone monitoring system. Non-contact estimation algorithms must be robust to patient motion and changes in body position and pose. Ideally, algorithms will be able to track patient motion in real-time and detect optimal skin segments for extracting physiological signals. Furthermore, camera algorithms must work under different environmental scenarios such as lighting conditions and distances from patients. Future work should develop methodologies that ensure accurate and robust non-contact estimation, even during challenging clinical scenarios.

A major limitation of previous studies of non-contact vital-sign monitoring is the lack of diversity in the clinical cohorts from which data were collected. PPGI signals (see chapter 6) extracted from skin with high pigmentation have been shown to have a lower SNR than those extracted from skin with low pigmentation (Addison et al. [2017]). The consequence of this is that non-contact monitoring algorithms may perform worse for people with dark skin, which clearly presents an ethical issue. It is difficult to develop algorithms that are equally performant for all skin types when non-contact monitoring datasets contain few patients with dark skin. This is demonstrated in chapter 6, with 14 out of 15 patients in the NVSM study having a

skin type of 1 or 2 on the Fitzpatrick scale. It is also possible that factors such as body mass index (BMI) or certain disease types may affect the ability to reliably extract a high quality PPGI waveform. It is important that future research is conducted to collect large datasets of video camera data from diverse patient cohorts, and that non-contact algorithms are developed that perform well for a diverse range of patients.

The clinical benefit of non-contact monitoring systems will be maximised when it is possible to derive multiple physiological parameters from camera data. Currently, heart rate and respiratory rate can be accurately estimated from video camera data, and non-contact estimation of SpO_2 is possible although current methods require calibration (Guazzi et al. [2015]; Verkruysse et al. [2017]). Non-contact estimation of blood pressure may one day be possible via non-contact estimation of PTT. PTT is correlated with blood pressure because the time for the pulse wave to travel between two points of the body is a function of the speed of the waveform, and therefore the arterial pressure. Methodological advances in the estimation of SpO_2, PTT and blood pressure should be explored. Furthermore, thermal video camera data were collected during the NVSM study, from which it may be possible to estimate body temperature. In the future, analysis of the thermal camera data from the NVSM study will allow the feasibility of non-contact temperature estimation to be assessed.

In chapter 7, we developed methods to visualize spatial distributions of respiratory amplitude during periods of abnormal respiratory physiology. Video data can monitor different areas of the body simultaneously, allowing the derivation of novel physiological biomarkers. Spatial maps of other physiological parameters may be clinically valuable, and the development of such methods should be explored further.

Finally, the use of non-contact vital-sign estimates for EWS computation requires exploration. Video data should be collected in post-operative wards for a sufficient number of patients such that physiological deterioration and adverse clinical events are included in the dataset. Future clinical studies should be conducted to compile such datasets, and analyses should be conducted to develop and test EWSs specifically for non-contact monitoring. These EWSs should contain features of vital signs and vital-sign variability, using methods such as those described in chapter 5.

8.3 Concluding remarks

Features of diurnal vital-sign variability are predictive of physiological deterioration, and can be used to improve the performance of EWSs. However, the computation of diurnal variability, and indeed the computation of EWSs, is limited by the low

quantity of data that is collected by routine manual vital-sign observations. Providing post-operative patients with continuous vital-sign monitoring may be beneficial as it could improve their monitoring coverage and sample all important changes in physiology. Furthermore, continuous data could provide a richer source of longitudinal data for deriving patterns that predict physiological deterioration. However, there is poor agreement between continuous and manual vital-sign data, suggesting that EWSs developed for manual vital-sign data may alert differently with continuous vital-sign data. Hence, EWSs should be developed specifically for use with continuous vital-sign monitoring equipment.

Features of diurnal variability may be computed from continuous vital-sign data to discriminate between periods of physiological deterioration and stability. Thus, EWSs with variability features may be developed for continuous data. However, reliably providing post-operative patients with continuous vital-sign monitoring is challenging: patients do not tolerate being tethered to bedside monitors for extended periods and the CALMS-2 study showed that compliance with ambulatory monitors can be poor. Non-contact vital-sign monitoring with a video camera may be a potential method of providing post-operative patients with long-term continuous monitoring. We conducted a clinical trial of non-contact vital-sign monitoring in 15 post-operative patients during elective ICU care, and demonstrated that non-contact methods can be used to accurately monitor important changes in heart rate and respiratory rate. As such, non-contact monitoring methods can detect changes that are relevant to EWS computation. Furthermore, non-contact methods may provide additional value through the derivation of new biomarkers such as respiratory amplitude.

Future work is required before non-contact camera-based vital-sign monitoring is feasible for long-term use in post-operative wards. Algorithms need to be developed to allow robust vital-sign estimation in the presence of motion, variable lighting conditions, and different patient physiology and skin tones. In addition, methods for estimating other vital signs such as SpO_2 and blood pressure should be explored. Multi-centre clinical trials with large patient cohorts should be conducted to test the efficacy of non-contact monitoring with sufficient power. Our results provide a proof-of-concept for a non-contact monitoring system that uses machine learning to detect physiological deterioration in real-time; the realisation of this vision may one day prevent adverse events and reduce the length of hospital stays.

Appendix A

The Bland-Altman Phenomenon

A.1 Background

Bland-Altman plots (Bland and Altman [1986]) are commonly used in measurement method comparison studies to assess agreement. In a Bland-Altman plot, the difference between measurement values taken by the two methods are plotted against the average of the measurement values, allowing the bias between measurement methods to be visualised (the average of the differences) as well as the limits of agreement, which bound 95% of the differences. The Bland-Altman method is commonly misused, for example by plotting the differences between measurement methods against the values from one of the methods, which Bland and Altman [1995] showed to be misleading. This introduces a spurious correlation making it erroneously appear that the differences between methods are related to the measurement values. This was demonstrated theoretically by Bland and Altman [1995], and reproduced here below.

Consider the values measured by two methods, M_1 and M_2, which are correlated by ρ and have variances σ_1^2 and σ_2^2. Unless the two methods are very different, ρ should be large and σ_1^2 and σ_2^2 should be similar. If the Bland-Altman method is correctly followed, the correlation coefficient between the difference and average is:

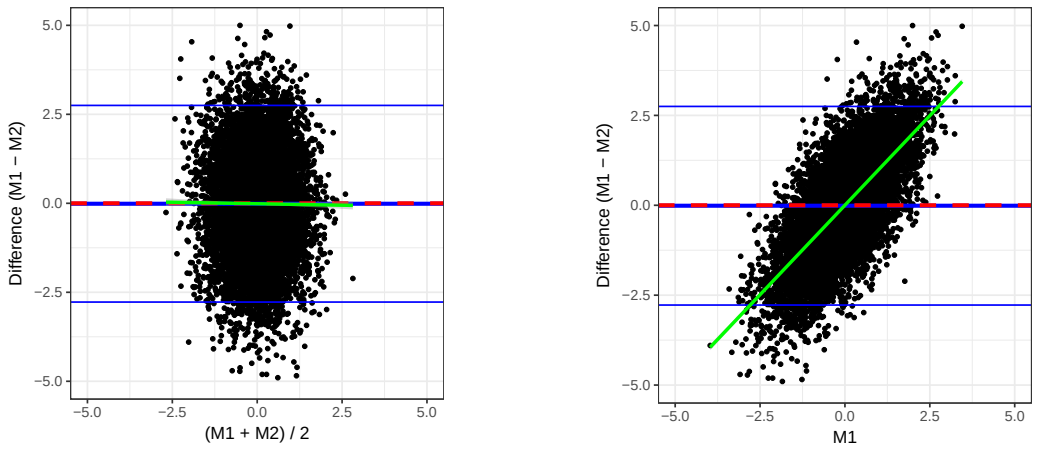
$$Corr(M_1 - M_2, \frac{M_1 + M_2}{2}) = \frac{\sigma_1^2 - \sigma_2^2}{\sqrt{(\sigma_1^2 + \sigma_2^2)^2 - 4\rho\sigma_1^2\sigma_2^2}} \tag{A.1}$$

This correlation will be zero if the variances are equal and will be small unless the differences are truly related to the measurement value. If the differences between the measurements are plotted against the values of one of the measurement methods, the correlation between the differences and the measurement values is given by:

$$Corr(M_1 - M_2, M_2) = \frac{\rho\sigma_1^2 - \sigma_2^2}{\sqrt{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}} \tag{A.2}$$

This correlation will usually be negative, and even if σ_1^2 and σ_2^2 are equal, equation (A.2) will be reduced to $-\sqrt{\frac{1-\rho}{2}}$. Thus, the practice of plotting the differences against the values of one of the methods is guaranteed to show a relationship between

the differences and the measurement values. We demonstrate this effect in figures A.1a and A.1b, which show Bland-Altman plots for two synthetic measurement sources M_1 and M_2 simulated as values drawn from a Gaussian distribution with mean of 0 and standard deviation of 1 (thus the difference and value are uncorrelated). In figure A.1a a correct Bland-Altman plot is shown and in figure A.1b the methodological error is shown, which demonstrates the spurious correlation.



(a) Correct Bland-Altman **(b)** Incorrect Bland-Altman

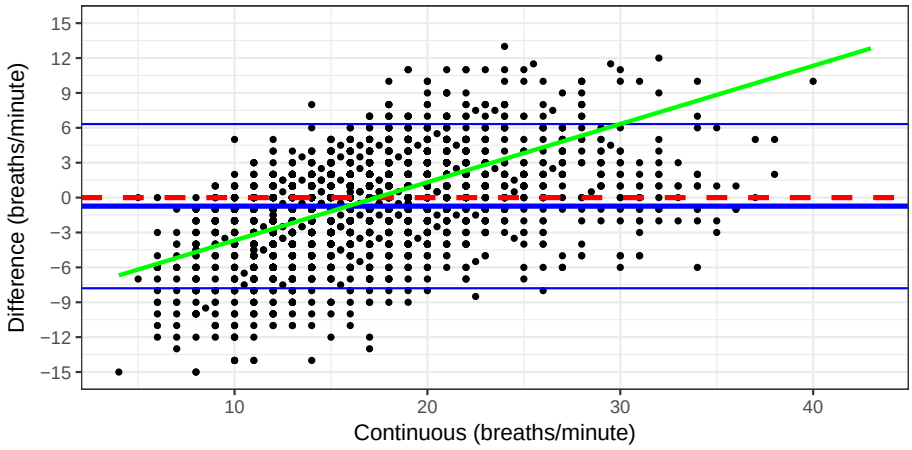
Figure A.1: Bland-Altman plots showing the differences between measurements from two methods against (correctly) the averages of the measurements (a) and against (incorrectly) the measurement values from one of the methods (b).

A.2 The "smoothing effect"

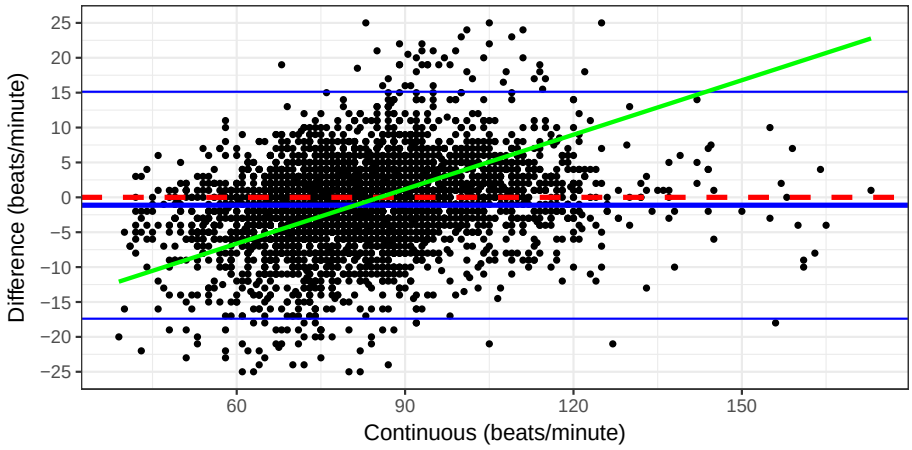
To demonstrate the influence of this spurious correlation on real data, we replicate the "smoothing effect" and agreement analysis (chapter 4) between manual and continuous vital-sign measurements shown in figures 4.4a to 4.4c. In this instance, however, we introduce the error described in appendix A.1.

We subtracted the manual vital-sign from the continuous vital-sign in each measurement pair. We then computed incorrect Bland-Altman plots by plotting the continuous-manual difference against the continuous vital-sign value (rather than against the continuous-manual average). We again (see section 4.4) show lines for the bias, LoAs and the regression from the linear mixed effects model (see equation (4.1)).

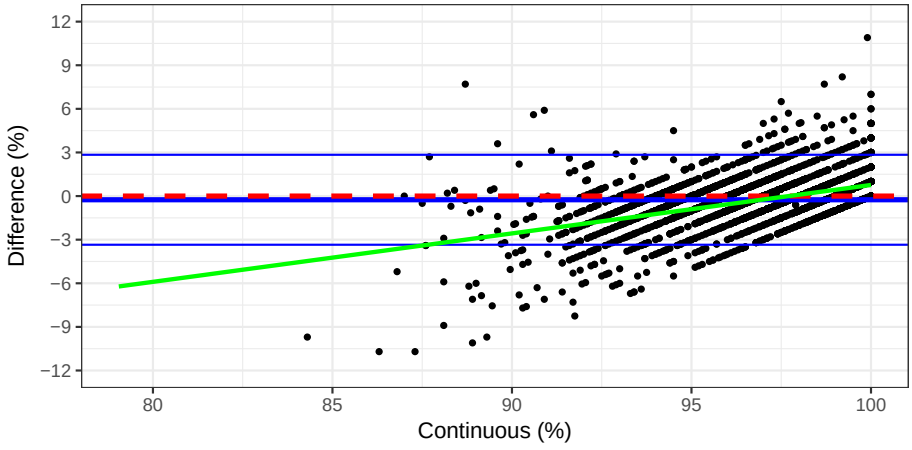
The results are shown in figures A.2a to A.2c. A clear spurious correlation can be seen when compared with figures 4.4a to 4.4c, with a much larger regression slope. Notably, Taenzer et al. [2014] used an equivalent (incorrect) method in their analysis, so the reported results were likely to be affected by this spurious correlation.



(a) Respiratory rate



(b) Heart rate



(c) SpO_2

Figure A.2: Incorrect Bland-Altman plots for the comparison of continuous and manual measurements of respiratory rate (a), heart rate (b) and SpO_2 (c). The (continuous - manual) differences are plotted against the continuous measurements. Measurement values are black dots, bias and LoAs are in blue (thick middle line is bias) and the zero y-intercept is in red. The predicted difference at continuous measurement values from the mixed effects model (equation (4.1)) is shown in green.

References

Sheila Adam and Mandy Odell. An acute problem? A report of the National Confidential Enquiry into Patient Outcome and Death. *Nursing in Critical Care*, 10(5):225–227, 2005.

Paul S Addison, Dominique Jacquel, David MH Foo, and Ulf R Borg. Video-based heart rate monitoring across a range of skin pigmentations during an acute hypoxic challenge. *Journal of Clinical Monitoring and Computing*, 0:1–10, 2017.

Thomas Ahrens. The most important vital signs are not being measured. *Australian Critical Care*, 21(1):3–5, 2008.

Douglas G Altman and J Martin Bland. Diagnostic tests 1: Sensitivity and specificity. *The BMJ*, 308(6943):1552, 1994.

Carlos A Alvarez, Christopher A Clark, Song Zhang, Ethan A Halm, John J Shannon, Carlos E Girod, Lauren Cooper, and Ruben Amarasingham. Predicting out of intensive care unit cardiopulmonary arrest or death using electronic medical record data. *BMC Medical Informatics and Decision Making*, 13(1):28, 2013.

Lars W Andersen, Won Young Kim, Maureen Chase, Sharri J Mortensen, Ari Moskowitz, Victor Novack, Michael N Cocchi, Michael W Donnino, et al. The prevalence and significance of abnormal vital signs prior to in-hospital cardiac arrest. *Resuscitation*, 98:112–117, 2016.

Mary Armitage, Jane Eddleston, and Tim Stokes. NICE guidelines: Recognising and responding to acute illness in adults in hospital: Summary of NICE guidance. *The BMJ*, 335(7613):258–259, 2007.

Diana E Ayala, Ana Moyá, Juan J Crespo, Carmen Castiñeira, Manuel Domínguez-Sardiña, Sonia Gomara, Elvira Sineiro, Artemio Mojón, María J Fontao, and Ramón C Hermida. Circadian pattern of ambulatory blood pressure in hypertensive patients with and without type 2 diabetes. *Chronobiology International*, 30(1-2): 99–115, 2013.

Jack Badawy, Oanh Kieu Nguyen, Christopher Clark, Ethan A Halm, and Anil N Makam. Is everyone really breathing 20 times a minute? Assessing epidemiology and variation in recorded respiratory rate in hospitalised adults. *BMJ Quality & Safety*, 26(10):832–836, 2017.

Tessy Badriyah, James S Briggs, Paul Meredith, Stuart W Jarvis, Paul E Schmidt, Peter I Featherstone, David R Prytherch, and Gary B Smith. Decision-tree early warning score (DTEWS) validates the design of the National Early Warning Score (NEWS). *Resuscitation*, 85(3):418–423, 2014.

Mirza Mansoor Baig, Hamid Gholamhosseini, and Martin J Connolly. A comprehensive survey of wearable and wireless ECG monitoring systems for older adults. *Medical & Biological Engineering & Computing*, 51(5):485–495, 2013.

Thomas C Bailey, Yixin Chen, Yi Mao, Chenyang Lu, Gregory Hackmann, Scott T Micek, Kevin M Heard, Kelly M Faulkner, and Marin H Kollef. A trial of a real-time alert for clinical deterioration in patients hospitalized on general medical wards. *Journal of Hospital Medicine*, 8(5):236–242, 2013.

Jane E Ball, Trevor Murrells, Anne Marie Rafferty, Elizabeth Morrow, and Peter Griffiths. 'Care left undone' during nursing shifts: Associations with workload and perceived quality of care. *BMJ Quality & Safety*, 23:116–125, 2014.

Paul Barach and Stephen D Small. Reporting and preventing medical mishaps: lessons from non-medical near miss reporting systems. *The BMJ*, 320(7237):759–763, 2000.

Charlotte Barfod, Marlene Mauson Pankoke Lauritzen, Jakob Klim Danker, György Sölétormos, Jakob Lundager Forberg, Peter Anthony Berlac, Freddy Lippert, Lars Hyldborg Lundstrøm, Kristian Antonsen, and Kai Henrik Wiborg Lange. Abnormal vital signs are strong predictors for intensive care unit admission and in-hospital mortality in adults triaged in the emergency department - a prospective cohort study. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, 20(1):28, 2012.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

Brett K Beaulieu-Jones, Patryk Orzechowski, and Jason H Moore. Mapping patient trajectories using longitudinal extraction and deep learning in the MIMIC-III critical care database. *Pacific Symposium on Biocomputing*, 23:123–132, 2018.

Kate Beaumont, Dagmar Luettel, and Richard Thomson. Deterioration in hospital patients: Early signs and appropriate actions. *Nursing Standard*, 23(1):43–48, 2008.

Rinaldo Bellomo, Donna Goldsmith, Shigehiko Uchino, Jonathan Buckmaster, Graeme Hart, Helen Opdam, William Silvester, Laurie Doolan, and Geoffrey Gutteridge.

Prospective controlled trial of effect of medical emergency team on postoperative morbidity and mortality rates. *Critical Care Medicine*, 32(4):916–921, 2004.

Rinaldo Bellomo, Michael Ackerman, Michael Bailey, Richard Beale, Greg Clancy, Valerie Danesh, Andreas Hvarfner, Edgar Jimenez, David Konrad, Michele Lecardo, et al. A controlled trial of electronic automated advisory vital signs monitoring in general hospital wards. *Critical Care Medicine*, 40(8):2349–2361, 2012.

J HM Bergmann and AH McGregor. Body-worn sensor design: What do patients and clinicians want? *Annals of Biomedical Engineering*, 39(9):2299–2312, 2011.

Edgar A Bernal, Eribaweimon Shilla, Himanshu Madhu, Lalit K Mestha, and Robert P Loce. Non-contact monitoring of spatio-temporal respiratory mechanics via depth sensing, September 24 2015. US Patent App. 14/223,402.

Donald M Berwick, David R Calkins, C Joseph McCannon, and Andrew D Hackbarth. The 100,000 lives campaign: Setting a goal and a deadline for improving health care quality. *JAMA: The Journal of the American Medical Association*, 295(3):324–327, 2006.

Anil Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhyā: The Indian Journal of Statistics*, 7(4):401–406, 1946.

Deepak Bhonagiri, David V Pilcher, Michael J Bailey, et al. Increased mortality associated with after-hours and weekend admission to the intensive care unit: A retrospective analysis. *Medical Journal of Australia*, 194(6):287–292, 2011.

Gordon Bingham, Irma Bilgrami, Mandy Sandford, Sarah Larwill, Judit Orosz, Carl Luckhoff, and Tony Kambourakis. Avoiding adult in-hospital cardiac arrest: A retrospective cohort study to determine preventability. *Australian Critical Care*, 31 (4):219–225, 2018.

Philip F Binkley. Predicting the potential of wearable technology. *IEEE Engineering in Medicine and Biology Magazine*, 22(3):23–27, 2003.

Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

J Martin Bland and Douglas G Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327(8476):307–310, 1986.

J Martin Bland and Douglas G Altman. Comparing methods of measurement: Why plotting difference against standard method is misleading. *The Lancet*, 346(8982): 1085–1087, 1995.

J Martin Bland and Douglas G Altman. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, 8(2):135–160, 1999.

Frank E Block. Normal fluctuation of physiologic cardiovascular variables during anesthesia and the phenomenon of "smoothing". *Journal of Clinical Monitoring*, 7 (2):141–145, 1991.

Kathleen L Bobay, Karen L Fiorelli, and Alfred J Anderson. Failure to rescue: A preliminary study of patient-level factors. *Journal of Nursing Care Quality*, 23(3): 211–215, 2008.

Giles Bond-Smith, Ajay P Belgaumkar, Brian R Davidson, and Kurinchi Selvan Gurusamy. Enhanced recovery protocols for major upper gastrointestinal, liver and pancreatic surgery. *Cochrane Database of Systematic Reviews*, (2), 2016.

Márcio M Boniatti, Neusa Azzolini, Marina V Viana, Berenice SP Ribeiro, Renata S Coelho, Rodrigo K Castilho, Márcio R Guimarães, Lia Zorzi, Luis F Schulz, Edison M Rodrigues Filho, et al. Delayed medical emergency team calls and associated outcomes. *Critical Care Medicine*, 42(1):26–30, 2014.

Timothy Bonnici, Christina Orphanidou, David Vallance, Alexander Darrell, and Lionel Tarassenko. Testing of wearable monitors in a real-world hospital environment: What lessons can be learnt? In *Ninth International Conference on Wearable and Implantable Body Sensor Networks*, pages 79–84. IEEE, 2012.

Timothy Bonnici, Lionel Tarassenko, David A Clifton, and Peter Watkinson. The digital patient. *Clinical Medicine*, 13(3):252–257, 2013.

Kendrick Boyd, Kevin H Eng, and C David Page. Area under the precision-recall curve: Point estimates and confidence intervals. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 451–466. Springer, 2013.

Andrea Bravi, André Longtin, and Andrew JE Seely. Review and classification of variability analysis techniques with clinical applications. *Biomedical Engineering Online*, 10(1):90, 2011.

Troyen A Brennan, Lucian L Leape, Nan M Laird, Liesi Hebert, A Russell Localio, Ann G Lawthers, Joseph P Newhouse, Paul C Weiler, and Howard H Hiatt. Incidence of adverse events and negligence in hospitalized patients: Results of the Harvard Medical Practice Study I. *New England Journal of Medicine*, 324(6):370–376, 1991.

Richard P Brent. *Algorithms for minimization without derivatives*. Courier Corporation, 2013.

Thomas F Budinger. Biomonitoring with wireless communications. *Annual Review of Biomedical Engineering*, 5(1):383–412, 2003.

Michael D Buist, Gaye E Moore, Stephen A Bernard, Bruce P Waxman, Jeremy N Anderson, and Tuan V Nguyen. Effects of a medical emergency team on reduction of incidence of and mortality from unexpected cardiac arrests in hospital: Preliminary study. *The BMJ*, 324(7334):387–390, 2002.

L Cabré, J Mancebo, JF Solsona, P Saura, I Gich, L Blanch, G Carrasco, MC Martin, Bioethics Working Group of the SEMICYUC, et al. Multicenter study of the multiple organ dysfunction syndrome in intensive care units: The usefulness of Sequential Organ Failure Assessment scores in decision making. *Intensive Care Medicine*, 31(7):927–933, 2005.

A John Camm, Marek Malik, J Thomas Bigger, Günter Breithardt, Sergio Cerutti, Richard J Cohen, Philippe Coumel, Ernest L Fallen, Harold L Kennedy, RE Kleiger, et al. Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *European Heart Journal*, 17(3):354–381, 1996.

Hanqing Cao, Larry Eshelman, Nicolas Chbat, Larry Nielsen, Brian Gross, and Mohammed Saeed. Predicting ICU hemodynamic instability using continuous multiparameter trends. In *30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3803–3806. IEEE, 2008.

M Cardona-Morrell, M Prgomet, R Lake, M Nicholson, R Harrison, J Long, J Westbrook, J Braithwaite, and K Hillman. Vital signs monitoring and nurse-patient interaction: A qualitative observational study of hospital practice. *International Journal of Nursing Studies*, 56:9–16, 2016a.

M Cardona-Morrell, M Prgomet, RM Turner, M Nicholson, and K Hillman. Effectiveness of continuous or intermittent vital signs monitoring in preventing adverse

events on general wards: A systematic review and meta-analysis. *International Journal of Clinical Practice*, 70(10):806–824, 2016b.

Lucienne TQ Cardoso, Cintia MC Grion, Tiemi Matsuo, Elza HT Anami, Ivanil AM Kauss, Ludmila Seko, and Ana M Bonametti. Impact of delayed admission to intensive care units on mortality of critically ill patients: A cohort study. *Critical Care*, 15(1):R28, 2011.

B Carstensen, L Gurrin, C Claus Ekstrom, and M Figurski. MethComp: Functions for analysis of agreement in method comparison studies. R package version 1.22.2, 2016.

Bendix Carstensen, Julie Simpson, and Lyle C Gurrin. Statistical models for assessing agreement in method comparison studies with replicate measurements. *The International Journal of Biostatistics*, 4(1):16, 2008.

D Mi Catley, C Thornton, C Jordan, JR Lehane, D Royston, and JG Jones. Pronounced, episodic oxygen desaturation in the postoperative period: Its association with ventilatory pattern and analgesic regimen. *Anesthesiology*, 63(1):20–28, 1985.

Sitthichok Chaichulee, Mauricio Villarroel, Joao Jorge, Carlos Arteta, Gabrielle Green, Kenny McCormick, Andrew Zisserman, and Lionel Tarassenko. Multi-task convolutional neural network for patient detection and skin segmentation in continuous non-contact vital sign monitoring. In *12th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 266–272. IEEE, 2017.

Indra M Chakravarty, JD Roy, and Radha Govind Laha. *Handbook of methods of applied statistics*. McGraw-Hill, 1967.

Donald B Chalfin, Stephen Trzeciak, Antonios Likourezos, Brigitte M Baumann, R Phillip Dellinger, DELAY-ED study group, et al. Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. *Critical Care Medicine*, 35(6):1477–1483, 2007.

Paul S Chan, Renuka Jain, Brahmajee K Nallmothu, Robert A Berg, and Comilla Sasson. Rapid response teams: A systematic review and meta-analysis. *Archives of Internal Medicine*, 170(1):18–26, 2010.

Peter Charlton, Drew A Birrenkott, Timothy Bonnici, Marco AF Pimentel, Alistair EW Johnson, Jordi Alastruey, Lionel Tarassenko, Peter J Watkinson, Richard Beale,

and David A Clifton. Breathing rate estimation from the electrocardiogram and photoplethysmogram: A review. *IEEE Reviews in Biomedical Engineering*, 2017.

Arthur Charpentier and Emmanuel Flachaire. Log-transform kernel density estimation of income distribution. *L'Actualité Économique*, 91(1-2):141–159, 2015.

Liangyou Chen, Andrew T Reisner, Andrei Gribok, and Jaques Reifman. Exploration of prehospital vital sign trends for the prediction of trauma outcomes. *Prehospital Emergency Care*, 13(3):286–294, 2009.

Octav Chipara, Chenyang Lu, Thomas C Bailey, and Gruia-Catalin Roman. Reliable clinical monitoring using wireless sensor networks: Experiences in a step-down hospital unit. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*, pages 155–168. ACM, 2010.

Euisun Choi and Chulhee Lee. Feature extraction based on the Bhattacharyya distance. *Pattern Recognition*, 36(8):1703–1709, 2003.

George Chrysochoou and Scott R Gunn. Demonstrating the benefit of medical emergency teams (MET) proves more difficult than anticipated. *Critical Care*, 10 (2):306, 2006.

Matthew M Churpek, Trevor C Yuen, Christopher Winslow, Ari A Robicsek, David O Meltzer, Robert D Gibbons, and Dana P Edelson. Multicenter development and validation of a risk stratification tool for ward patients. *American Journal of Respiratory and Critical Care Medicine*, 190(6):649–655, 2014.

Matthew M Churpek, Richa Adhikari, and Dana P Edelson. The value of vital sign trends for detecting clinical deterioration on the wards. *Resuscitation*, 102:1–5, 2016.

Matthew Michael Churpek, Dana P Edelson, Ji Yeon Lee, Kyle Carey, and Ashley Snyder. Association between survival and time of day for rapid response team calls in a national registry. *Critical Care Medicine*, 45(10):1677–1682, 2017.

Gari D Clifford and David Clifton. Wireless technology in disease management and medicine. *Annual Review of Medicine*, 63:479–492, 2012.

David A Clifton, Lei Clifton, Dona-Maria Sandu, GB Smith, Lionel Tarassenko, Sarah A Vollam, and Peter J Watkinson. 'Errors' and omissions in paper-based

early warning scores: The association with changes in vital signs - a database analysis. *BMJ Open*, 5(7):e007376, 2015.

Lei Clifton, David A Clifton, Marco AF Pimentel, Peter J Watkinson, and Lionel Tarassenko. Predictive monitoring of mobile patients by combining clinical observations with data from wearable sensors. *IEEE Journal of Biomedical and Health Informatics*, 18(3):722–730, 2014.

Nancy R Cook. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*, 115(7):928–935, 2007.

Richard I Cook, John S McDonald, and Enrico Nunziata. Differences between handwritten and automatic blood pressure records. *Anesthesiology*, 71(3):385–390, 1989.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

Michelle Cretikos, Jack Chen, Ken Hillman, Rinaldo Bellomo, Simon Finfer, Arthas Flabouris, Merit Study Investigators, et al. The objective medical emergency team activation criteria: A case–control study. *Resuscitation*, 73(1):62–72, 2007.

Michelle A Cretikos, Rinaldo Bellomo, Ken Hillman, Jack Chen, Simon Finfer, and Arthas Flabouris. Respiratory rate: The neglected vital sign. *Medical Journal of Australia*, 188(11):657–659, 2008.

J Paul Curry and Carla R Jungquist. A critical assessment of monitoring practices, patient deterioration, and alarm fatigue on inpatient wards: A review. *Patient Safety in Surgery*, 8(1):29, 2014.

RJ Cusack and JF Coutts. Do not be alarmed, the patient is monitored. *Critical Care*, 7(5):349–350, 2003.

Brian H Cuthbertson, Massoud Boroujerdi, Laurin McKie, Lorna Aucott, and Gordon Prescott. Can physiological variables and early warning scoring systems allow early recognition of the deteriorating surgical patient? *Critical Care Medicine*, 35(2): 402–409, 2007.

Maria Cvach. Monitor alarm fatigue: An integrative review. *Biomedical Instrumentation & Technology*, 46(4):268–277, 2012.

Mohamed Daabiss. American Society of Anaesthesiologists physical status classification. *Indian Journal of Anaesthesia*, 55(2):111–115, 2011.

Albert Dahan, Leon Aarts, and Terry W Smith. Incidence, reversal, and prevention of opioid-induced respiratory depression. *The Journal of the American Society of Anesthesiologists*, 112(1):226–238, 2010.

Jonathan Daly. *Video camera monitoring to detect changes in haemodynamics*. PhD thesis, University of Oxford, 2016.

Daniel Ldecke. *ggeffects: Create Tidy Data Frames of Marginal Effects for 'ggplot' from Model Outputs*, 2018. URL `https://CRAN.R-project.org/package=ggeffects`. R package version 0.3.2.

Gerard de Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rPPG. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013.

MA DeVita, RS Braithwaite, R Mahidhara, S Stuart, M Foraida, and RL Simmons. Use of medical emergency team responses to reduce hospital cardiopulmonary arrests. *Quality and Safety in Health Care*, 13(4):251–254, 2004.

Michael DeVita. Medical emergency teams: Deciphering clues to crises in hospitals. *Critical Care*, 9(4):325–326, 2005.

Michael A DeVita, Rinaldo Bellomo, Kenneth Hillman, John Kellum, Armando Rotondi, Dan Teres, Andrew Auerbach, Wen-Jon Chen, Kathy Duncan, Gary Kenward, et al. Findings of the first consensus conference on medical emergency teams. *Critical Care Medicine*, 34(9):2463–2478, 2006.

Michael A DeVita, Gary B Smith, Sheila K Adam, Inga Adams-Pizarro, Michael Buist, Rinaldo Bellomo, Robert Bonello, Erga Cerchiari, Barbara Farlow, Donna Goldsmith, et al. "Identifying the hospitalised patient in crisis" - a consensus conference on the afferent limb of rapid response systems. *Resuscitation*, 81(4): 375–382, 2010.

J Hugh Devitt, Theodore Rapanos, Matt Kurrek, Marsha M Cohen, and Melissa Shaw. The anesthetic record: Accuracy and completeness. *Canadian Journal of Anesthesia*, 46(2):122–128, 1999.

Dimitri Dorcaratto, Luís Grande, and Manuel Pera. Enhanced recovery in gastrointestinal surgery: Upper gastrointestinal surgery. *Digestive Surgery*, 30(1):70–78, 2013.

J Ede, S Vollam, O Gibson, J Darbyshire, L Tarassenko, and PJ Watkinson. Non-contact vital sign monitoring in critical care; a human factors analysis of staff perceptions. In preparation, 2018.

Malcolm Elliott and Alysia Coventry. Critical care: The eight vital signs of patient monitoring. *British Journal of Nursing*, 21(10):621–625, 2012.

Dominik Maria Endres and Johannes E Schindelin. A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7):1858–1860, 2003.

Gabriel J Escobar, Juan Carlos LaGuardia, Benjamin J Turk, Arona Ragins, Patricia Kipnis, and David Draper. Early detection of impending physiologic deterioration among patients who are not in intensive care: Development of predictive models using data from an automated electronic medical record. *Journal of Hospital Medicine*, 7(5):388–395, 2012.

David Evans, Brent Hodgkinson, and Judith Berry. Vital signs in hospital patients: A systematic review. *International Journal of Nursing Studies*, 38(6):643–650, 2001.

Sibylle Fallet, Virginie Moser, Fabian Braun, and Jean-Marc Vesin. Imaging photoplethysmography: What are the best locations on the face to estimate heart rate? In *Computing in Cardiology Conference*, pages 341–344. IEEE, 2016.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

KCH Fearon, Olle Ljungqvist, M Von Meyenfeldt, A Revhaug, CHC Dejong, K Lassen, J Nygren, J Hausel, M Soop, Jens Andersen, et al. Enhanced recovery after surgery: A consensus review of clinical care for patients undergoing colonic resection. *Clinical Nutrition*, 24(3):466–477, 2005.

Litong Feng, Lai-Man Po, Xuyuan Xu, and Yuming Li. Motion artifacts suppression for remote imaging photoplethysmography. In *19th International Conference on Digital Signal Processing*, pages 18–23. IEEE, 2014.

R Fensli, PE Pedersen, T Gundersen, and O Hejlesen. Sensor acceptance model – measuring patient acceptance of wearable sensors. *Methods of Information in Medicine*, 47(1):89–95, 2008.

JR Fernández, RC Hermida, and A Mojón. Chronobiological analysis techniques. Application to blood pressure. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 367(1887):431–445, 2009.

Flavio Lopes Ferreira, Daliana Peres Bota, Annette Bross, Christian Mélot, and Jean-Louis Vincent. Serial evaluation of the SOFA score to predict outcome in critically ill patients. *JAMA: The Journal of the American Medical Association*, 286(14):1754–1758, 2001.

GP Findlay, H Shotton, K Kelly, and M Mason. Time to intervene? A review of patients who underwent cardiopulmonary resuscitation as a result of an in-hospital cardiorespiratory arrest. *A Report by the National Confidential Enquiry into Patient Outcome and Death*, 2012.

Susannah G Fleming and Lionel Tarassenko. A comparison of signal processing techniques for the extraction of breathing rate from the photoplethysmogram. *International Journal Of Biological and Medical Research*, 2(4):232–236, 2007.

Emily Fox, Erik B Sudderth, Michael I Jordan, and Alan S Willsky. Nonparametric Bayesian learning of switching linear dynamical systems. In *Advances in Neural Information Processing Systems*, pages 457–464, 2009.

Cory Franklin and James Mathew. Developing strategies to prevent inhospital cardiac arrest: Analyzing responses of physicians and nurses in the hours before the event. *Critical Care Medicine*, 22(2):244–247, 1994.

Bradley A Fritz, Yixin Chen, Teresa M Murray-Torres, Stephen Gregory, Arbi Ben Abdallah, Alex Kronzer, Sherry Lynn McKinnon, Thaddeus Budelier, Daniel L Helsten, Troy S Wildes, Anshuman Sharma, and Michael Simon Avidan. Using machine learning techniques to develop forecasting algorithms for postoperative complications: Protocol for a retrospective study. *BMJ Open*, 8(4):e020124, 2018.

Eugene S Fu, John B Downs, John W Schweiger, Rafael V Miguel, and Robert A Smith. Supplemental oxygen impairs detection of hypoventilation by pulse oximetry. *Chest*, 126(5):1552–1558, 2004.

Shinsuke Fujiwara, Tomotaka Koike, Megumi Moriyasu, Masashi Nakagawa, Kazuaki Atagi, Alan K Lefor, Shigeki Fujitani, IHCA study group, Takeshi Ikeda, Yuka

Takamatsu, et al. A retrospective study of in-hospital cardiac arrest. *Acute Medicine & Surgery*, 3(4):320–325, 2016.

Sanjay Galhotra, Michael A DeVita, Richard L Simmons, Andrea Schmid, Medical Emergency Response Improvement Team (MERIT) Committee, et al. Impact of patient monitoring on the diurnal pattern of medical emergency team activation. *Critical Care Medicine*, 34(6):1700–1706, 2006.

Sanjay Galhotra, Michael A DeVita, Richard L Simmons, and Mary Amanda Dew. Mature rapid response system and potentially avoidable cardiopulmonary arrests in hospital. *Quality and Safety in Health Care*, 16(4):260–265, 2007.

Haiyan Gao, Ann McDonnell, David A Harrison, Tracey Moore, Sheila Adam, Kathleen Daly, Lisa Esmonde, David R Goldhill, Gareth J Parry, Arash 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*, 33(4):667–679, 2007.

Tia Gao, Dan Greenspan, Matt Welsh, Radford R Juang, and Alex Alm. Vital signs monitoring and patient tracking over a wireless network. In *27th Annual International Conference of the Engineering in Medicine and Biology Society*, pages 102–105. IEEE, 2006.

J Gardner-Thorpe, N Love, J Wrightson, S Walsh, and N Keeling. The value of Modified Early Warning Score (MEWS) in surgical in-patients: A prospective observational study. *The Annals of The Royal College of Surgeons of England*, 88 (6):571–575, 2006.

LA Geddes, MH Voelz, CF Babbs, JD Bourland, and WA Tacker. Pulse transit time as an indicator of arterial blood pressure. *Psychophysiology*, 18(1):71–74, 1981.

Erina Ghosh, Larry Eshelman, Lin Yang, Eric Carlson, and Bill Lord. Description of vital signs data measurement frequency in a medical/surgical unit at a community hospital in United States. *Data in Brief*, 16:612–616, 2018.

Jay A Gold and Kay Simmons. The 100,000 Lives Campaign: Saving lives in Wisconsin and the nation. *Policy*, 105(8):89–90, 2006.

Ary L Goldberger. Heartbeats, hormones, and health: Is variability the spice of life? *American Journal of Respiratory and Critical Care Medicine*, 163(6):1289–1290, 2001.

David R Goldhill and Anne Sumner. Outcome of intensive care patients in a group of British intensive care units. *Critical Care Medicine*, 26(8):1337–1345, 1998.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. The MIT Press, 2016.

Marian Grade, Michael Quintel, and B Michael Ghadimi. Standard perioperative management in gastrointestinal surgery. *Langenbeck's Archives of Surgery*, 396(5): 591–606, 2011.

Ake Grenvik, Stanley Ballou, Edward McGinley, J Eugene Millen, Wils L Cooley, and Peter Safar. Impedance pneumography: Comparison between chest impedance changes and respiratory volumes in 11 healthy volunteers. *Chest*, 62(4):439–443, 1972.

Alessandro R Guazzi, Mauricio Villarroel, Joao Jorge, Jonathan Daly, Matthew C Frise, Peter A Robbins, and Lionel Tarassenko. Non-contact measurement of oxygen saturation with an RGB camera. *Biomedical Optics Express*, 6(9):3320–3338, 2015.

Gregory Hackmann, Minmin Chen, Octav Chipara, Chenyang Lu, Yixin Chen, Marin Kollef, and Thomas C Bailey. Toward a two-tier clinical warning system for hospitalized patients. In *AMIA Annual Symposium Proceedings*, pages 511–519. American Medical Informatics Association, 2011.

Chris Hands, Eleanor Reid, Paul Meredith, Gary B Smith, David R Prytherch, Paul E Schmidt, and Peter I Featherstone. Patterns in the recording of vital signs and early warning scores: Compliance with a clinical escalation protocol. *BMJ Quality & Safety*, 22(9):719–726, 2013.

James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.

Alistair Hann. *Multi-parameter monitoring for early warning of patient deterioration*. PhD thesis, University of Oxford, 2008.

Frank Harrell. *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *arXiv preprint arXiv:1703.07771*, 2017.

Miguel Hernandez-Silveira, Kamran Ahmed, Su-Shin Ang, Fahriya Zandari, Tinaz Mehta, Rebecca Weir, Alison Burdett, Chris Toumazou, and Stephen J Brett. Assessment of the feasibility of an ultra-low power, wireless digital patch for the continuous ambulatory monitoring of vital signs. *BMJ Open*, 5(5):e006606, 2015.

K Hillman, M Parr, A Flabouris, G Bishop, and A Stewart. Redefining in-hospital resuscitation: The concept of the medical emergency team. *Resuscitation*, 48(2): 105–110, 2001a.

Ken Hillman, Jack Chen, Michelle Cretikos, Rinaldo Bellomo, Daniel Brown, Gordon Doig, Simon Finfer, Arthas Flabouris, MERIT Study Investigators, et al. Introduction of the medical emergency team (MET) system: A cluster-randomised controlled trial. *Lancet*, 365(9477):2091–2097, 2005.

Ken M Hillman, Peter J Bristow, Tien Chey, Kathy Daffurn, Theresa Jacques, Sandra L Norman, Gillian F Bishop, and Grant Simmons. Duration of life-threatening antecedents prior to intensive care admission. *Intensive Care Medicine*, 28(11): 1629–1634, 2002.

KM Hillman, PJ Bristow, T Chey, K Daffurn, T Jacques, SL Norman, GF Bishop, and G Simmons. Antecedents to hospital deaths. *Internal Medicine Journal*, 31(6): 343–348, 2001b.

Timothy J Hodgetts, Gary Kenward, Ioannis G Vlachonikolis, Susan Payne, and Nicolas Castle. The identification of risk factors for cardiac arrest and formulation of activation criteria to alert a medical emergency team. *Resuscitation*, 54(2): 125–131, 2002a.

Timothy J Hodgetts, Gary Kenward, Ioannis Vlackonikolis, Susan Payne, Nicolas Castle, Robert Crouch, Neil Ineson, and Loua Shaikh. Incidence, location and reasons for avoidable in-hospital cardiac arrest in a district general hospital. *Resuscitation*, 54(2):115–123, 2002b.

Edmond A Hooker, Daniel J O'Brien, Daniel F Danzl, Jennifer AC Barefoot, and James E Brown. Respiratory rates in emergency department patients. *The Journal of Emergency Medicine*, 7(2):129–132, 1989.

Marilyn Hravnak, Leslie Edwards, Amy Clontz, Cynthia Valenta, Michael A DeVita, and Michael R Pinsky. Defining the incidence of cardiorespiratory instability

in patients in step-down units using an electronic integrated monitoring system. *Archives of Internal Medicine*, 168(12):1300–1308, 2008.

Marilyn Hravnak, Michael A DeVita, Amy Clontz, Leslie Edwards, Cynthia Valenta, and Michael R Pinsky. Cardiorespiratory instability before and after implementing an integrated monitoring system. *Critical Care Medicine*, 39(1):65–72, 2011.

Marilyn Hravnak, Lujie Chen, Artur Dubrawski, Eliezer Bose, and Michael R Pinsky. Temporal distribution of instability events in continuously monitored step-down unit patients: Implications for rapid response systems. *Resuscitation*, 89:99–105, 2015.

Michael Imhoff and Silvia Kuhls. Alarm algorithms in critical care monitoring. *Anesthesia & Analgesia*, 102(5):1525–1537, 2006.

Triin Jakobson, Juri Karjagin, Liisa Vipp, Martin Padar, Ants-Hendrik Parik, Liis Starkopf, Hartmut Kern, Olavi Tammik, and Joel Starkopf. Postoperative complications and mortality after major gastrointestinal surgery. *Medicina*, 50(2):111–117, 2014.

John T James. A new, evidence-based estimate of patient harms associated with hospital care. *Journal of Patient Safety*, 9(3):122–128, 2013.

Emma Jeffs, Sarah Vollam, J Duncan Young, Lois Horsington, Breda Lynch, and Peter J Watkinson. Wearable monitors for patients following discharge from an intensive care unit: Practical lessons learnt from an observational study. *Journal of Advanced Nursing*, 72(8):1851–1862, 2016.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016.

Maximilian J Johnston, Sonal Arora, Dominic King, George Bouras, Alex M Almoudaris, Rachel Davis, and Ara Darzi. A systematic review to identify the factors that affect failure to rescue and escalation of care in surgery. *Surgery*, 157(4): 752–763, 2015.

Daryl Jones, Samantha Bates, Stephen Warrillow, Helen Opdam, Donna Goldsmith, Geoff Gutteridge, and Rinaldo Bellomo. Circadian pattern of activation of the

medical emergency team in a teaching hospital. *Critical Care*, 9(4):R303–R306, 2005.

Daryl Jones, Rinaldo Bellomo, Samantha Bates, Stephen Warrillow, Donna Goldsmith, Graeme Hart, and Helen Opdam. Patient monitoring and the timing of cardiac arrests and medical emergency team calls in a teaching hospital. *Intensive Care Medicine*, 32(9):1352–1356, 2006.

Daryl Jones, Rinaldo Bellomo, Graeme K Hart, Ambica Parma, RT Gibney, Sean M Bagshaw, Gaurav Bhatia, Tim Leong, Glenn Eastwood, Leah Peck, et al. The timing of Rapid-Response Team activations: A multicentre international study. *Critical Care and Resuscitation*, 15(1):15–20, 2013.

Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001. URL `http://www.scipy.org/`.

Alexei A Kamshilin, Serguei Miridonov, Victor Teplov, Riku Saarenheimo, and Ervin Nippolainen. Photoplethysmographic imaging of high spatial resolution. *Biomedical Optics Express*, 2(4):996–1006, 2011.

Alexei A Kamshilin, Victor Teplov, Ervin Nippolainen, Serguei Miridonov, and Rashid Giniatullin. Variability of microcirculation detected by blood pulsation imaging. *PLOS ONE*, 8(2):e57117, 2013.

Alexei A Kamshilin, Ervin Nippolainen, Igor S Sidorov, Petr V Vasilev, Nikolai P Erofeev, Natalia P Podolian, and Roman V Romashko. A new look at the essence of the imaging photoplethysmography. *Scientific Reports*, 5:10494, 2015.

Alexei A Kamshilin, Igor S Sidorov, Laura Babayan, Maxim A Volynsky, Rashid Giniatullin, and Oleg V Mamontov. Accurate measurement of the pulse wave delay with imaging photoplethysmography. *Biomedical Optics Express*, 7(12):5138–5147, 2016.

Walter Karlen, Srinivas Raman, J Mark Ansermino, and Guy A Dumont. Multiparameter respiratory rate estimation from the photoplethysmogram. *IEEE Transactions on Biomedical Engineering*, 60(7):1946–1953, 2013.

Dennis L Kasper and Tinsley Randolph Harrison. *Harrison's Principles of Internal Medicine*. McGraw-Hill, Medical Publishing Division, New York, 2005.

Juliane Kause, Gary Smith, David Prytherch, Michael Parr, Arthas Flabouris, and Ken Hillman. A comparison of antecedents to cardiac arrests, deaths and emergency intensive care admissions in Australia and New Zealand, and the United Kingdom - the ACADEMIA study. *Resuscitation*, 62(3):275–282, 2004.

Michael G Kenward and James H Roger. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53(3):983–997, 1997.

Sara Khalid. *Data Fusion Models for Detection of Vital-sign Deterioration in Acutely Ill Patients*. PhD thesis, University of Oxford, 2014.

Sara Khalid, David A Clifton, and Lionel Tarassenko. A Bayesian patient-based model for detecting deterioration in vital signs using manual observations. In *International Symposium on Foundations of Health Informatics Engineering and Systems*, pages 146–158, 2014.

Aditya Khosla, Yu Cao, Cliff Chiung-Yu Lin, Hsu-Kuang Chiu, Junling Hu, and Honglak Lee. An integrated machine learning approach to stroke prediction. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 183–192. ACM, 2010.

Won Young Kim, Yu Jung Shin, Jin Mi Lee, Jin Won Huh, Younsuck Koh, Chae-Man Lim, and Sang Bum Hong. Modified early warning score changes prior to cardiac arrest in general wards. *PLOS ONE*, 10(6):e0130523, 2015.

Robert E Kleiger, J Philip Miller, J Thomas Bigger, and Arthur J Moss. Decreased heart rate variability and its association with increased mortality after acute myocardial infarction. *American Journal of Cardiology*, 59(4):256–262, 1987.

T Klingeberg and Marcel Schilling. Mobile wearable device for long term monitoring of vital signs. *Computer Methods and Programs in Biomedicine*, 106(2):89–96, 2012.

William A Knaus, Elizabeth A Draper, Douglas P Wagner, and Jack E Zimmerman. APACHE II: A severity of disease classification system. *Critical Care Medicine*, 13 (10):818–829, 1985.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

Una Kyriacos, J Jelsma, and S Jordan. Monitoring vital signs using early warning scoring systems: A review of the literature. *Journal of Nursing Management*, 19(3): 311–330, 2011.

Stephen T Lawless. Crying wolf: False alarms in a pediatric intensive care unit. *Critical Care Medicine*, 22(6):981–985, 1994.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553): 436–444, 2015.

A Lee, G Bishop, KM Hillman, and K Daffurn. The medical emergency team. *Anaesthesia and Intensive Care*, 23(2):183–186, 1995.

Jos GC Lerou, Ris Dirksen, Marc van Daele, Gerard MM Nijhuis, and Jan F Crul. Automated charting of physiological variables in anesthesia: A quantitative comparison of automated versus handwritten anesthesia records. *Journal of Clinical Monitoring*, 4(1):37–47, 1988.

Q Li and GD Clifford. Dynamic time warping and machine learning for signal quality assessment of pulsatile signals. *Physiological Measurement*, 33(9):1491, 2012.

H Lehman Li-wei, Shamim Nemati, Ryan P Adams, and Roger G Mark. Discovering shared dynamics in physiological signals: Application to patient monitoring in ICU. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5939–5942. IEEE, 2012.

H Lehman Li-wei, Ryan P Adams, Louis Mayaud, George B Moody, Atul Malhotra, Roger G Mark, and Shamim Nemati. A physiological time series dynamics-based approach to patient monitoring and outcome prediction. *IEEE Journal of Biomedical and Health Informatics*, 19(3):1068–1076, 2015.

Chih-Jen Lin, Ruby C Weng, and S Sathiya Keerthi. Trust region Newton methods for large-scale logistic regression. In *Proceedings of the 24th International Conference on Machine Learning*, pages 561–568. ACM, 2007.

S Suave Lobodzinski. ECG patch monitors for assessment of cardiac rhythm abnormalities. *Progress in Cardiovascular Diseases*, 56(2):224–229, 2013.

Paris B Lovett, Jason M Buchwald, Kai Stürmann, and Polly 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*, 45(1):68–76, 2005.

Hua Ma, Andriy I Bandos, Howard E Rockette, and David Gur. On use of partial area under the ROC curve for evaluation of diagnostic performance. *Statistics in Medicine*, 32(20):3449–3458, 2013.

CA Mackay, DP Burke, JA Burke, KM Porter, D Bowden, and D Gorman. Association between the assessment of conscious level using the AVPU system and the Glasgow coma scale. *Prehospital Immediate Care*, 4:17–19, 2000.

Ray R Maddox, Harold Oglesby, Carolyn K Williams, Marianne Fields, and Sherry Danello. Continuous respiratory monitoring and a "smart" infusion system improve safety of patient-controlled analgesia in the postoperative period. In Kerm Henriksen, James B Battles, Margaret A Keyes, and Mary L Grady, editors, *Advances in Patient Safety: New Directions and Alternative Approaches (Vol. 4: Technology and Medication Safety)*. Agency for Healthcare Research and Quality (US), 2008.

Yuka Maeda, Masaki Sekine, and Toshiyo Tamura. Relationship between measurement site and motion artifacts in wearable reflected photoplethysmography. *Journal of Medical Systems*, 35(5):969–976, 2011.

Ritesh Maharaj, Ivan Raffaele, and Julia Wendon. Rapid response systems: A systematic review and meta-analysis. *Critical Care*, 19(1):254, 2015.

Yi Mao, Yixin Chen, Gregory Hackmann, Minmin Chen, Chenyang Lu, Marin Kollef, and Thomas C Bailey. Medical data mining for early deterioration warning in general hospital wards. In *11th International Conference on Data Mining Workshops*, pages 1042–1049. IEEE, 2011.

Markos Markou and Sameer Singh. Novelty detection: A review part 1: Statistical approaches. *Signal Processing*, 83(12):2481–2497, 2003.

Catherine A Martin and Barry P McGrath. White-coat hypertension. *Clinical and Experimental Pharmacology and Physiology*, 41(1):22–29, 2014.

Frank Daniel Martos-Benítez, Anarelys Gutiérrez-Noyola, and Adisbel Echevarría-Víctores. Postoperative complications and clinical outcomes among patients undergoing thoracic and gastrointestinal cancer surgery: A prospective cohort study. *Revista Brasileira de Terapia Intensiva*, 28(1):40–48, 2016.

MATLAB. *Version 9.0.0 (R2016a)*. The MathWorks Inc., Natick, Massachusetts, 2016.

Jackie McBride, Debbie Knight, Jo Piper, and Gary B Smith. Long-term effect of introducing an early warning score on respiratory rate charting on general wards. *Resuscitation*, 65(1):41–44, 2005.

Louise McCallum and Dan Higgins. Measuring body temperature. *Nursing Times*, 108(45):20–22, 2012.

Donna Katzman McClish. Analyzing a portion of the ROC curve. *Medical Decision Making*, 9(3):190–195, 1989.

Daniel McDuff, Sarah Gontarek, and Rosalind W Picard. Improvements in remote cardiopulmonary measurement using a five band digital camera. *IEEE Transactions on Biomedical Engineering*, 61(10):2593–2601, 2014a.

Daniel McDuff, Sarah Gontarek, and Rosalind W Picard. Remote detection of photoplethysmographic systolic and diastolic peaks using a digital camera. *IEEE Transactions on Biomedical Engineering*, 61(12):2948–2954, 2014b.

Jennifer McGaughey, Fiona Alderdice, Robert Fowler, Atul Kapila, Alain Mayhew, and Marianne Moutray. Outreach and Early Warning Systems (EWS) for the prevention of Intensive Care admission and death of critically ill adult patients on general hospital wards. *Cochrane Database of Systematic Reviews*, (3), 2007.

Beate H McGhee and Elizabeth J Bridges. Monitoring arterial blood pressure: What you may not know. *Critical Care Nurse*, 22(2):60–79, 2002.

Ann McGinley and Rupert M Pearse. A national early warning score for acutely ill patients. *The BMJ*, 345:e5310, 2012.

Calum McGregor, Sanjiv Chohan, and Jonathon OReilly. Collaborative approach to reducing cardiac arrests in an acute medical unit. *BMJ Open Quality*, 6(2):e000026, 2017.

G McNeill and D Bryden. Do either early warning systems or emergency response teams improve hospital patient survival? A systematic review. *Resuscitation*, 84 (12):1652–1667, 2013.

Peter McQuillan, Sally Pilkington, Alison Allan, Bruce Taylor, Alasdair Short, Giles Morgan, Mick Nielsen, David Barrett, and Gary Smith. Confidential inquiry into quality of care before admission to intensive care. *The BMJ*, 316(7148):1853–1858, 1998.

DJ Meredith, D Clifton, P Charlton, J Brooks, CW Pugh, and L Tarassenko. Photoplethysmographic derivation of respiratory rate: A review of relevant physiology. *Journal of Medical Engineering & Technology*, 36(1):1–7, 2012.

Jeffrey A Meyerhardt, Joel E Tepper, and Alan P Venook. Special series: Advances in GI cancer. *Journal of Clinical Oncology*, 33(16):1717–1720, 2015.

Michael W Millar-Craig, Charles N Bishop, and EB Raftery. Circadian variation of blood-pressure. *The Lancet*, 311(8068):795–797, 1978.

QJW Milner and GR Mathews. An assessment of the accuracy of pulse oximeters. *Anaesthesia*, 67(4):396–401, 2012.

Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: Review, opportunities and challenges. *Briefings in Bioinformatics*, pages 1–11, 2017.

Hamed Monkaresi, Rafael A Calvo, and Hong Yan. A machine learning approach to improve contactless heart rate monitoring using a webcam. *IEEE Journal of Biomedical and Health Informatics*, 18(4):1153–1160, 2014.

Joseph Randall Moorman, Waldemar A Carlo, John Kattwinkel, Robert L Schelonka, Peter J Porcelli, Christina T Navarrete, Eduardo Bancalari, Judy L Aschner, Marshall Whit Walker, Jose A Perez, et al. Mortality reduction by heart rate characteristic monitoring in very low birth weight neonates: A randomized trial. *The Journal of Pediatrics*, 159(6):900–906, 2011.

RJM Morgan, F Williams, and MM Wright. An early warning scoring system for detecting developing critical illness. *Clinical Intensive Care*, 8(2):100, 1997.

CF Mullins and A Psirides. Activities of a Medical Emergency Team: A prospective observational study of 795 calls. *Anaesthesia and Intensive Care*, 44(1):34–43, 2016.

Vishal Nangalia, David R Prytherch, and Gary B Smith. Health technology assessment review: Remote monitoring of vital signs - current status and future challenges. *Critical Care*, 14(5):233, 2010.

Jack Needleman, Peter Buerhaus, Soeren Mattke, Maureen Stewart, and Katya Zelevinsky. Nurse-staffing levels and the quality of care in hospitals. *New England Journal of Medicine*, 346(22):1715–1722, 2002.

Jack Needleman, Peter Buerhaus, V Shane Pankratz, Cynthia L Leibson, Susanna R Stevens, and Marcelline Harris. Nurse staffing and inpatient hospital mortality. *New England Journal of Medicine*, 364(11):1037–1045, 2011.

Su Q Nguyen, Edwin Mwakalindile, James S Booth, Vicki Hogan, Jordan Morgan, Charles T Prickett, John P Donnelly, and Henry E Wang. Automated electronic medical record sepsis detection in the emergency department. *PeerJ*, 2:e343, 2014.

Centre for Clinical Practice at NICE. Acutely ill patients in hospital: Recognition of and response to acute illness in adults in hospital. Technical Report 50, National Institute for Health and Clinical Excellence, 2007.

Lena Nilsson, Anders Johansson, and Sigga Kalman. Monitoring of respiratory rate in postoperative care using a new photoplethysmographic technique. *Journal of Clinical Monitoring and Computing*, 16(4):309–315, 2000.

Daniel J Niven, Jonathan E Gaudet, Kevin B Laupland, Kelly J Mrklas, Derek J Roberts, and Henry Thomas Stelfox. Accuracy of peripheral thermometers for estimating temperature: A systematic review and meta-analysis. *Annals of Internal Medicine*, 163(10):768–777, 2015.

Jerry P Nolan, Jasmeet Soar, Gary B Smith, Carl Gwinnutt, Francesca Parrott, Sarah Power, David A Harrison, Edel Nixon, Kathryn Rowan, et al. Incidence and outcome of in-hospital cardiac arrest in the United Kingdom National Cardiac Arrest Audit. *Resuscitation*, 85(8):987–992, 2014.

David JP O'Callaghan, Parveen Jayia, Eyston Vaughan-Huxley, Michael Gribbon, Maie Templeton, James RA Skipworth, and Anthony C Gordon. An observational study to determine the effect of delayed admission to the intensive care unit on patient outcome. *Critical Care*, 16(5):R173, 2012.

E Andrew Ochroch, Michael W Russell, William C Hanson III, Gayle A Devine, Andrew J Cucchiara, Mark G Weiner, and Sanford J Schwartz. The impact of continuous pulse oximetry monitoring on intensive care unit admissions from a postsurgical care floor. *Anesthesia & Analgesia*, 102(3):868–875, 2006.

M Odell, IJ Rechner, A Kapila, T Even, D Oliver, CWH Davies, L Milsom, A Forster, and K Rudman. The effect of a critical care outreach service and an early warning scoring system on respiratory rate recording on the general wards. *Resuscitation*, 74(3):470–475, 2007.

Christina Orphanidou, David Clifton, Shahab Khan, Marc Smith, Jacques Feldmar, and Lionel Tarassenko. Telemetry-based vital sign monitoring for ambulatory

hospital patients. In *31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4650–4653. IEEE, 2009.

Christina Orphanidou, Timothy Bonnici, Peter Charlton, David Clifton, David Vallance, and Lionel Tarassenko. Signal-quality indices for the electrocardiogram and photoplethysmogram: Derivation and applications to wireless monitoring. *IEEE Journal of Biomedical and Health Informatics*, 19(3):832–838, 2015.

J Ortmüller, T Gauer, M Wilms, H Handels, and R Werner. Respiratory surface motion measurement by Microsoft Kinect: Implementation and evaluation of a clinical setup. *Current Directions in Biomedical Engineering*, 1(1):270–273, 2015.

Frank J Overdyk, Rickey Carter, Ray R Maddox, Jarred Callura, Amy E Herrin, and Craig Henriquez. Continuous oximetry/capnometry monitoring reveals frequent desaturation and bradypnea during patient-controlled analgesia. *Anesthesia & Analgesia*, 105(2):412–418, 2007.

Alexandros Pantelopoulos and Nikolaos G Bourbakis. A survey on wearable sensor-based systems for health monitoring and prognosis. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(1):1–12, 2010.

S Parissopoulos and S Kotzabassaki. Critical care outreach and the use of early warning scoring systems; a literature review. *ICUs and Nursing Web Journal*, 21: 1–11, 2005.

Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

Shyamal Patel, Hyung Park, Paolo Bonato, Leighton Chan, and Mary Rodgers. A review of wearable sensors and systems with application in rehabilitation. *Journal of NeuroEngineering and Rehabilitation*, 9(1):21, 2012.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

John Asger Petersen, Kristian Antonsen, and Lars S Rasmussen. Frequency of early warning score assessment and clinical deterioration in hospitalized patients: A randomized trial. *Resuscitation*, 101:91–96, 2016.

Clemence Petit, Rick Bezemer, and Louis Atallah. A review of recent advances in data analytics for post-operative patient deterioration detection. *Journal of Clinical Monitoring and Computing*, 32(3):391–402, 2018.

Jason Phua, Wang Jee Ngerng, and Tow Keang Lim. The impact of a delay in intensive care unit admission for community-acquired pneumonia. *European Respiratory Journal*, 36(4):826–833, 2010.

Marco AF Pimentel. *Modelling of vital-sign data from post-operative patients*. PhD thesis, University of Oxford, 2015.

Marco AF Pimentel, David A Clifton, Lei Clifton, and Lionel Tarassenko. Probabilistic estimation of respiratory rate using Gaussian processes. In *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2902–2905. IEEE, 2013a.

Marco AF Pimentel, David A Clifton, Lei Clifton, Peter J Watkinson, and Lionel Tarassenko. Modelling physiological deterioration in post-operative patient vital-sign data. *Medical & Biological Engineering & Computing*, 51(8):869–877, 2013b.

Marco AF Pimentel, David A Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014.

Marco AF Pimentel, Peter H Charlton, and David A Clifton. Probabilistic estimation of respiratory rate from wearable sensors. In *Wearable Electronics Sensors*, pages 241–262. Springer, 2015.

Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics Express*, 18(10):10762–10774, 2010.

Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE Transactions on Biomedical Engineering*, 58(1):7–11, 2011.

K Povšič, M Jezeršek, and J Možina. Real-time 3D visualization of the thoracoabdominal surface during breathing with body movement and deformation extraction. *Physiological Measurement*, 36(7):1497, 2015.

Mirela Prgomet, Magnolia Cardona-Morrell, Margaret Nicholson, Rebecca Lake, Janet Long, Johanna Westbrook, Jeffrey Braithwaite, and Ken Hillman. Vital signs

monitoring on general wards: Clinical staff perceptions of current practices and the planned introduction of continuous monitoring technology. *International Journal for Quality in Health Care*, 28(4):515–521, 2016.

George Priestley, Wendy Watson, Arash Rashidian, Caroline Mozley, Daphne Russell, Jonathan Wilson, Judith Cope, Dianne Hart, Diana Kay, Karen Cowley, et al. Introducing Critical Care Outreach: A ward-randomised trial of phased introduction in a general hospital. *Intensive Care Medicine*, 30(7):1398–1404, 2004.

Aleš Procházka, Martin Schätz, Oldřich Vyšata, and Martin Vališ. Microsoft Kinect visual and depth sensors for breathing and heart rate analysis. *Sensors*, 16(7):996, 2016a.

Aleš Procházka, Oldřich Vyšata, Martin Schätz, Hana Charvátová, Carmen Paz Suárez Araujo, Oana Geman, and Vladimír Mařík. Video processing and 3D modelling of chest movement using MS Kinect depth sensor. In *International Workshop on Computational Intelligence for Multimedia Understanding*, pages 1–5. IEEE, 2016b.

David R Prytherch, Gary B Smith, Paul Schmidt, Peter I Featherstone, Kate Stewart, Debbie Knight, and Bernie Higgins. Calculating early warning scores - a classroom comparison of pen and paper and hand-held computer methods. *Resuscitation*, 70 (2):173–178, 2006.

David R Prytherch, Gary B Smith, Paul E Schmidt, and Peter I Featherstone. ViEWS - towards a national early warning score for detecting adult inpatient deterioration. *Resuscitation*, 81(8):932–937, 2010.

Richard Pullinger, Sarah Wilson, Rob Way, Mauro Santos, David Wong, David Clifton, Jacqueline Birks, and Lionel Tarassenko. Implementing an electronic observation and early warning score chart in the emergency department: A feasibility study. *European Journal of Emergency Medicine*, 24(6):e11–e16, 2017.

Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. Benchmarking deep learning models on large healthcare datasets. *Journal of Biomedical Informatics*, 83:112–134, 2018.

Joshua Pyke, Andreas H Taenzer, Christian E Renaud, and Susan P McGrath. Developing a continuous monitoring infrastructure for detection of inpatient deterioration. *Joint Commission Journal on Quality and Patient Safety*, 38(9):428–431, 2012.

John A Quinn and Christopher KI Williams. Known unknowns: Novelty detection in condition monitoring. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 1–6. Springer, 2007.

John A Quinn, Christopher KI Williams, and Neil McIntosh. Factorial switching linear dynamical systems applied to physiological condition monitoring. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1537–1551, 2009.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2013. URL `http://www.R-project.org/`.

Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

Roberto Refinetti and Michael Menaker. The circadian rhythm of body temperature. *Physiology & Behavior*, 51(3):613–637, 1992.

David L Reich, Richard K Wood Jr, Raafat Mattar, Marina Krol, David C Adams, Sabera Hossain, and Carol A Bodian. Arterial blood pressure and heart rate discrepancies between handwritten and computerized anesthesia records. *Anesthesia & Analgesia*, 91(3):612–616, 2000.

Joshua S Richman and J Randall Moorman. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology: Heart and Circulatory Physiology*, 278(6):H2039–H2049, 2000.

Santiago Romero-Brufau, Jeanne M Huddleston, James M Naessens, Matthew G Johnson, Joel Hickman, Bruce W Morlan, Jeffrey B Jensen, Sean M Caples, Jennifer L Elmer, Julie A Schmidt, et al. Widely used track and trigger scores: Are they ready for automation in practice? *Resuscitation*, 85(4):549–552, 2014.

Santiago Romero-Brufau, Jeanne M Huddleston, Gabriel J Escobar, and Mark Liebow. Why the C-statistic is not informative to evaluate early warning scores and what metrics to use. *Critical Care*, 19(1):285, 2015.

Jason C Rubenstein, Mark Freher, Alan Kadish, and Jeffrey J Goldberger. Diurnal heart rate patterns in inappropriate sinus tachycardia. *Pacing and Clinical Electrophysiology*, 33(8):911–919, 2010.

Mohammed Saeed, Mauricio Villarroel, Andrew T Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H Kyaw, Benjamin Moody, and Roger G Mark. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database. *Critical Care Medicine*, 39(5): 952–960, 2011.

Nawal Salahuddin, Azam Shafquat, Qussay Marashly, Khaled Juan Zaza, Moh Sharshir, Moazzum Khurshid, Zeeshan Ali, Melissa Malgapo, Mouhamad Ghyath Jamil, Mohamed Shoukri, et al. Increases in heart rate variability signal improved outcomes in rapid response team consultations: A cohort study. *Cardiology Research and Practice*, 2018, 2018.

Seyed Salehizadeh, Duy Dao, Jeffrey Bolkhovsky, Chae Cho, Yitzhak Mendelson, and Ki H Chon. A novel time-varying spectral filtering algorithm for reconstruction of motion artifact corrupted heart rate signals during intense physical activities using a wearable photoplethysmogram sensor. *Sensors*, 16(1):10, 2016.

Monica Sapo, Shaozhi Wu, Shadnaz Asgari, Norma McNair, Farzad Buxey, Neil Martin, and Xiao Hu. A comparison of vital signs charted by nurses with automated acquired values using waveform quality indices. *Journal of Clinical Monitoring and Computing*, 23(5):263–271, 2009.

Suchi Saria, Daphne Koller, and Anna Penn. Discovering shared and individual latent structure in multiple time series. Technical report, 2010a. URL `http://arxiv.org/abs/1008.2028`.

Suchi Saria, Daphne Koller, and Anna Penn. Learning individual and population level traits from clinical temporal data. In *Neural Information Processing Systems (NIPS) Workshop on Predictive Models in Personalized Medicine*, 2010b.

Roland MH Schein, Nelson Hazday, Maria Pena, Bradley H Ruben, and Charles L Sprung. Clinical antecedents to in-hospital cardiopulmonary arrest. *Chest*, 98(6): 1388–1392, 1990.

Andrea Schmid, Leslie Hoffman, Mary Beth Happ, Gail A Wolf, and Michael DeVita. Failure to rescue: A literature review. *Journal of Nursing Administration*, 37(4): 188–198, 2007.

Paul E Schmidt, Paul Meredith, David R Prytherch, Duncan Watson, Valerie Watson, Roger M Killen, Peter Greengross, Mohammed A Mohammed, and Gary B Smith.

Impact of introducing an electronic physiological surveillance system on hospital mortality. *BMJ Quality & Safety*, 24(1):10–20, 2015.

Michael J Schull and Donald A Redelmeier. Continuous electrocardiographic monitoring and cardiac arrest outcomes in 8,932 telemetry ward patients. *Academic Emergency Medicine*, 7(6):647–652, 2000.

Andrew JE Seely and Peter T Macklem. Complex systems and the technology of variability analysis. *Critical Care*, 8(6):R367–R384, 2004.

Sue Sendelbach and Marjorie Funk. Alarm fatigue: A patient safety concern. *AACN Advanced Critical Care*, 24(4):378–386, 2013.

M Michael Shabot. The HP CareVue clinical information system. *International Journal of Clinical Monitoring and Computing*, 14(3):177–184, 1997.

Manish A Shah and Robert C Kurtz. Upper gastrointestinal cancer predisposition syndromes. *Hematology/Oncology Clinics of North America*, 24(5):815–835, 2010.

Dangdang Shao, Chenbin Liu, Francis Tsow, Yuting Yang, Zijian Du, Rafael Iriya, Hui Yu, and Nongjian Tao. Noncontact monitoring of blood oxygen saturation using camera and dual-wavelength imaging system. *IEEE Transactions on Biomedical Engineering*, 63(6):1091–1098, 2016.

Julie T Sharpley and Janet C Holden. Introducing an early warning scoring system in a district general hospital. *Nursing in Critical Care*, 9(3):98–103, 2004.

William C Shoemaker, Paul Appel, and Richard Bland. Use of physiologic monitoring to predict outcome and to assist in clinical decisions in critically ill postoperative patients. *The American Journal of Surgery*, 146(1):43–50, 1983.

Jeffrey H Silber, Sankey V Williams, Henry Krakauer, and J Sanford Schwartz. Hospital and patient characteristics associated with death after surgery: A study of adverse occurrence and failure to rescue. *Medical Care*, 30(7):615–629, 1992.

James E Sinex. Pulse oximetry: Principles and limitations. *The American Journal of Emergency Medicine*, 17(1):59–66, 1999.

Sarah P Slight, Calvin Franz, Michael Olugbile, Harvey V Brown, David W Bates, and Eyal Zimlichman. The return on investment of implementing a continuous monitoring system in general medical-surgical units. *Critical Care Medicine*, 42(8): 1862–1868, 2014.

Andrew F Smith and Jeremy Wood. Can some in-hospital cardio-respiratory arrests be prevented? A prospective survey. *Resuscitation*, 37(3):133–137, 1998.

Gary B Smith, David R Prytherch, Paul Schmidt, Peter I Featherstone, Debbie Knight, Gill Clements, and Mohammed A Mohammed. Hospital-wide physiological surveillance - a new approach to the early identification and management of the sick patient. *Resuscitation*, 71(1):19–28, 2006.

Gary B Smith, David R Prytherch, Paul E Schmidt, and Peter I Featherstone. Review and performance evaluation of aggregate weighted 'track and trigger' systems. *Resuscitation*, 77(2):170–179, 2008a.

Gary B Smith, David R Prytherch, Paul E Schmidt, Peter I Featherstone, and Bernie Higgins. A review, and performance evaluation, of single-parameter track and trigger systems. *Resuscitation*, 79(1):11–21, 2008b.

Gary B Smith, David R Prytherch, Paul Meredith, Paul E Schmidt, and Peter I Featherstone. The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation*, 84(4):465–470, 2013.

LN Smith, ML Smith, ME Fletcher, and AJ Henderson. A 3D machine vision method for non-invasive assessment of respiratory function. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 12(2):179–188, 2016.

Stephanie K Sprogis, Judy Currey, Julie Considine, Ian Baldwin, and Daryl Jones. Physiological antecedents and ward clinician responses before medical emergency team activation. *Critical Care and Resuscitation*, 19(1):50–56, 2017.

Ioan Stanculescu, Christopher KI Williams, and Yvonne Freer. A hierarchical switching linear dynamical system applied to the detection of sepsis in neonatal condition monitoring. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, pages 752–761, 2014.

Robert K Stoelting. Continuous postoperative electronic monitoring and the will to require it. *Anesthesia and Analgesia*, 121(3):579–581, 2015.

Christian P Subbe, Haiyan Gao, and David A Harrison. Reproducibility of physiological track-and-trigger warning systems for identifying at-risk patients on the ward. *Intensive Care Medicine*, 33(4):619–624, 2007.

Christian P Subbe, Bernd Duller, and Rinaldo Bellomo. Effect of an automated notification system for deteriorating ward patients on clinical outcomes. *Critical Care*, 21(1):52, 2017.

Olaf Such. Near infrared CCD imaging of hemodynamics. In *18th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 5, pages 2105–2106. IEEE, 1996.

Olaf Such, Sabine Acker, and Valdimir Blazek. Mapped hemodynamic data acquisition by near infrared CCD imaging. In *Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 2, pages 637–639. IEEE, 1997.

Shyam Sukumar, Florian Roghmann, Vincent Q Trinh, Jesse D Sammon, Mai-Kim Gervais, Hung-Jui Tan, Praful Ravi, Simon P Kim, Jim C Hu, Pierre I Karakiewicz, et al. National trends in hospital-acquired preventable adverse events after major cancer surgery in the USA. *BMJ Open*, 3(6):e002843, 2013.

Zhuo Sun, Daniel I Sessler, Jarrod E Dalton, PJ Devereaux, Aram Shahinyan, Amanda J Naylor, Matthew T Hutcherson, Patrick S Finnegan, Vikas Tandon, Saeed Darvish-Kazem, et al. Postoperative hypoxemia is common and persistent: A prospective blinded observational study. *Anesthesia and Analgesia*, 121(3):709–715, 2015.

Krishnaswamy Sundararajan, Arthas Flabouris, and Campbell Thompson. Diurnal variation in the performance of rapid response systems: The role of critical care services - a review article. *Journal of Intensive Care*, 4(1):15, 2016.

Bonni Syeda, Michael Gottsauner-Wolf, Stefan Denk, Phillip Pichler, Aliasghar Khorsand, and Dietmar Glogar. Arterial compliance: A diagnostic marker for atherosclerotic plaque burden? *American Journal of Hypertension*, 16(5):356–362, 2003.

AH Taenzer, GT Blike, S McGrath, J Pyke, M Herrick, C Renaud, and J Morgan. Postoperative monitoring - the Dartmouth experience. *Anesthesia Patient Safety Foundation Newsletter*, 27(1):1–21, 2012.

Andreas H Taenzer, Joshua B Pyke, Susan P McGrath, and George T Blike. Impact of pulse oximetry surveillance on rescue events and intensive care unit transfers: A before-and-after concurrence study. *Anesthesiology*, 112(2):282–287, 2010.

Andreas H Taenzer, Joshua B Pyke, and Susan P McGrath. A review of current and emerging approaches to address failure-to-rescue. *The Journal of the American Society of Anesthesiologists*, 115(2):421–431, 2011.

Andreas H Taenzer, Joshua Pyke, Michael D Herrick, Thomas M Dodds, and Susan P McGrath. A comparison of oxygen saturation data in inpatients with low oxygen saturation using automated continuous monitoring and intermittent manual data charting. *Anesthesia & Analgesia*, 118(2):326–331, 2014.

F Tahavori, M Alnowami, and K Wells. Marker-less respiratory motion modeling using the Microsoft Kinect for Windows. In *Medical Imaging 2014: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 9036, page 90360K. International Society for Optics and Photonics, 2014.

Chihiro Takano and Yuji Ohta. Heart rate measurement based on a time-lapse image. *Medical Engineering & Physics*, 29(8):853–857, 2007.

Toshiyo Tamura, Yuka Maeda, Masaki Sekine, and Masaki Yoshida. Wearable photoplethysmographic sensors - past and present. *Electronics*, 3(2):282–302, 2014.

L Tarassenko, A Hann, A Patterson, E Braithwaite, K Davidson, V Barber, and D Young. Biosign: Multi-parameter monitoring for early warning of patient deterioration. In *Proceedings of the 3rd IEE International Seminar on Medical Applications of Signal Processing*, pages 71–76. IET, 2005.

L Tarassenko, A Hann, and D Young. Integrated monitoring and analysis for early warning of patient deterioration. *British Journal of Anaesthesia*, 97(1):64–68, 2006.

L Tarassenko, M Villarroel, A Guazzi, J Jorge, DA Clifton, and C Pugh. Non-contact video-based vital sign monitoring using ambient light and auto-regressive models. *Physiological Measurement*, 35(5):807–831, 2014.

Lionel Tarassenko, David A Clifton, Michael R Pinsky, Marilyn T Hravnak, John R Woods, and Peter J Watkinson. Centile-based early warning scores derived from statistical distributions of vital signs. *Resuscitation*, 82(8):1013–1018, 2011.

Graham Teasdale and Bryan Jennett. Assessment of coma and impaired consciousness: A practical scale. *The Lancet*, 304(7872):81–84, 1974.

David N Thrush. Are automated anesthesia records better? *Journal of Clinical Anesthesia*, 4(5):386–389, 1992.

Joonas Tirkkonen, Tero Tamminen, and Markus B Skrifvars. Outcome of adult patients attended by rapid response teams: A systematic review of the literature. *Resuscitation*, 112:43–52, 2017.

Rebecca M Trinkle and Arthas Flabouris. Documenting rapid response system afferent limb failure and associated patient outcomes. *Resuscitation*, 82(7):810–814, 2011.

Chris H Van Leuvan and Imogen Mitchell. Missed opportunities? An observational study of vital sign measurements. *Critical Care and Resuscitation*, 10(2):111–115, 2008.

Kim van Loon, Bas van Zaane, Els J Bosch, Cor J Kalkman, and Linda M Peelen. Non-invasive continuous respiratory monitoring on general hospital wards: A systematic review. *PLOS ONE*, 10(12):e0144626, 2015.

Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics Express*, 16(26):21434–21445, 2008.

Wim Verkruysse, Marek Bartula, Erik Bresch, Mukul Rocque, Mohammed Meftah, and Ihor Kirenko. Calibration of contactless pulse oximetry. *Anesthesia and Analgesia*, 124(1):136–145, 2017.

Mauricio Villarroel, Alessandro Guazzi, João Jorge, Sara Davis, Peter Watkinson, Gabrielle Green, Asha Shenvi, Kenny McCormick, and Lionel Tarassenko. Continuous non-contact vital sign monitoring in neonatal intensive care unit. *Healthcare Technology Letters*, 1(3):87–91, 2014.

Mauricio Villarroel, Joao Jorge, Chris Pugh, and Lionel Tarassenko. Non-contact vital sign monitoring in the clinic. In *12th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 278–285. IEEE, 2017.

Todd R Vogel, Viktor Y Dombrovskiy, Jeffrey L Carson, Alan M Graham, and Stephen F Lowry. Postoperative sepsis in the United States. *Annals of Surgery*, 252 (6):1065–1071, 2010.

Terri Watkins, Lynn Whisman, and Pamela Booker. Nursing assessment of continuous vital sign surveillance to improve patient safety on the medical/surgical unit. *Journal of Clinical Nursing*, 25(1-2):278–281, 2016.

PJ Watkinson, VS Barber, JD Price, A Hann, L Tarassenko, and JD Young. A randomised controlled trial of the effect of continuous electronic physiological

monitoring on the adverse event rate in high risk medical and surgical patients. *Anaesthesia*, 61(11):1031–1039, 2006.

PJ Watkinson, Marco AF Pimentel, Lei Clifton, David A Clifton, JD Vollam, S Young, and L Tarassenko. Early detection of physiological deterioration in post-surgical patients using wearable technology combined with an integrated monitoring system: A pre- and post-interventional study. In preparation 2018.

Mariska Weenk, Harry van Goor, Bas Frietman, Lucien JLPG Engelen, Cornelis JHM van Laarhoven, Jan Smit, Sebastian JH Bredie, and Tom H van de Belt. Continuous monitoring of vital signs using wearable devices on the general ward: Pilot study. *JMIR mHealth and uHealth*, 5(7):e91, 2017.

Matthew B Weinger. Dangers of postoperative opioids. *Anesthesia Patient Safety Foundation Newsletter*, 21(4):61–67, 2006.

Matthew B Weinger and Lorri A Lee. No patient shall be harmed by opioid-induced respiratory depression. *Anesthesia Patient Safety Foundation Newsletter*, 26(21): 26–28, 2011.

Thomas Wentz, Hadi Fayad, Julien Bert, Olivier Pradier, Jean-François Clément, Sarah Vourch, Nicolas Boussion, and Dimitris Visvikis. Accuracy of dynamic patient surface monitoring using a time-of-flight camera and b-spline modeling for respiratory motion characterization. *Physics in Medicine & Biology*, 57(13): 4175–4193, 2012.

K White, IA Scott, A Vaux, and CM Sullivan. Rapid response teams in adult hospitals: Time for another look? *Internal Medicine Journal*, 45(12):1211–1220, 2015.

Fokko P Wieringa, Frits Mastik, and Antonius FW van der Steen. Contactless multiple wavelength photoplethysmographic imaging: A first step toward "SpO2 Camera" technology. *Annals of Biomedical Engineering*, 33(8):1034–1041, 2005.

Udaya Wijenayake and Soon-Yong Park. Real-time external respiratory motion measuring technique using an RGB-D camera and principal component analysis. *Sensors*, 17(8):1840, 2017.

JG Williams, SE Roberts, MF Ali, Wai-Yee Cheung, DR Cohen, Gaynor Demery, Adrian Edwards, Margot Greer, MD Hellier, HA Hutchings, et al. Gastroenterology services in the UK. The burden of disease, and the organisation and delivery of

services for gastrointestinal and liver disorders: A review of the evidence. *Gut*, 56 (Supplement 1):1–113, 2007.

Ross M Wilson, William B Runciman, Robert W Gibberd, Bernadette T Harrison, Liza Newby, John D Hamilton, et al. The quality in Australian health care study. *Medical Journal of Australia*, 163(9):458–471, 1995.

Bradford D Winters, Sallie J Weaver, Elizabeth R Pfoh, Ting Yang, Julius Cuong Pham, and Sydney M Dy. Rapid-response systems as a patient safety strategy: A systematic review. *Annals of Internal Medicine*, 158:417–425, 2013.

David Wong, Timothy Bonnici, Julia Knight, Lauren Morgan, Paul Coombes, and Peter Watkinson. SEND: A system for electronic notification and documentation of vital sign observations. *BMC Medical Informatics and Decision Making*, 15(1):68, 2015.

David Wong, Timothy Bonnici, Julia Knight, Stephen Gerry, James Turton, and Peter Watkinson. A ward-based time study of paper and electronic documentation for recording vital sign observations. *Journal of the American Medical Informatics Association*, 24(4):717–721, 2017.

Jenna Wong, Monica Taljaard, Alan J Forster, and Carl van Walraven. Does adding risk-trends to survival models improve in-hospital mortality predictions? A cohort study. *BMC Health Services Research*, 11(1):171, 2011.

T Wu. PPGI: New development in noninvasive and contactless diagnosis of dermal perfusion using near infrared light. *J. GCPD eV*, 7(1):17–24, 2003.

Ting Wu, Vladimir Blazek, and Hans Juergen Schmitt. Photoplethysmography imaging: A new noninvasive and noncontact method for mapping of the dermal perfusion changes. In *Optical Techniques and Instrumentation for the Measurement of Blood Composition, Structure, and Dynamics*, volume 4163, pages 62–71. International Society for Optics and Photonics, 2000.

Yuping Xiao, M Pamela Griffin, Douglas E Lake, and J Randall Moorman. Nearest-neighbor and logistic regression analyses of clinical and heart rate characteristics in the early diagnosis of neonatal sepsis. *Medical Decision Making*, 30(2):258–266, 2010.

Lonny Yarmus and David Feller-Kopman. Pneumothorax in the critically ill patient. *Chest*, 141(4):1098–1105, 2012.

Tuba Yilmaz, Robert Foster, and Yang Hao. Detecting vital signs with wearable wireless sensors. *Sensors*, 10(12):10837–10862, 2010.

Jordan C Yoder, Trevor C Yuen, Matthew M Churpek, Vineet M Arora, and Dana P Edelson. A prospective study of nighttime vital sign monitoring frequency and risk of clinical deterioration. *JAMA Internal Medicine*, 173(16):1554–1555, 2013.

William J Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

Michael P Young, Valerie J Gooder, Karen Bride, Brent James, and Elliott S Fisher. Inpatient transfers to the intensive care unit. *Journal of General Internal Medicine*, 18(2):77–83, 2003.

Nina Zaproudina, Victor Teplov, Ervin Nippolainen, Jukka A Lipponen, Alexei A Kamshilin, Matti Närhi, Pasi A Karjalainen, and Rashid Giniatullin. Asynchronicity of facial blood perfusion in migraine. *PLOS ONE*, 8(12):e80189, 2013.

Sebastian Zaunseder, Alexander Trumpp, Hannes Ernst, Michael Förster, and Hagen Malberg. Spatio-temporal analysis of blood perfusion by imaging photoplethysmography. In *Optical Diagnostics and Sensing XVIII: Toward Point-of-Care Diagnostics*, volume 10501, page 105010X. International Society for Optics and Photonics, 2018.

Robert M Zollinger, JF Kreul, and AJL Schneider. Man-made versus computer-generated anesthesia records. *Journal of Surgical Research*, 22(4):419–424, 1977.