Computer-Aided Analysis of Infant Respiratory Patterns

Carlos Alejandro Robles Rubio

Department of Biomedical Engineering
McGill University
Montreal, Quebec, Canada

April 2016

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy.

Table of Contents

Abstract

Infants recovering from surgery and anesthesia are at risk of life-threatening Postoperative Apnea (POA). There is no way to predict which infants will experience POA, and therefore all infants with postmenstrual age (PMA) ≤ 60 weeks need to be monitored in hospital for at least 12 h postoperatively. Evidence shows a link between abnormal postoperative respiratory patterns and the occurrence of POA. Thus, study of these patterns might be useful to predict an infant's risk of POA, as well as the time when such risk abates.

Comprehensive study of the postoperative respiratory patterns has been limited by two main factors. First, no representative set of respiratory data from infants at risk of POA is publicly available to investigators, and so any POA study involves a data acquisition phase that requires planning, approval, and execution. All these activities consume extensive resources and so the number of infants that can be enrolled is limited by the available budget. Second, there are no appropriate tools for the comprehensive analysis of the respiratory patterns. The most accepted method is conventional manual scoring (CMS), performed by expert scorers following guidelines from the American Academy of Sleep Medicine (AASM). CMS has several limitations: it has low intra- and inter-scorer repeatability, is labor intensive, time-consuming, and expensive. Moreover, CMS does not produce a comprehensive analysis of the respiratory patterns, but rather a list of "clinically relevant" events and the time of their occurrence. Thus, any pattern not considered "clinically relevant" by the AASM guidelines is not scored and cannot be analyzed.

This thesis addresses these limitations by creating a library of infant data and developing several tools for the comprehensive analysis of infant respiratory patterns. We accomplished this in 5 stages. First, we acquired a representative dataset comprising cardiorespiratory signals from infants at risk of POA, and made these data available to the public. Second, we developed a set of tools for the efficient, repeatable, and reliable manual scoring of infant respiratory patterns. We demonstrated that use of these tools produced an analysis with high accuracy and consistency, and improved intra- and inter-scorer repeatability. Third, we developed a method based on Expectation-Maximization (EM) to combine analyses from multiple manual scorers to minimize the effects of intra- and inter-scorer variability and yield "gold standard" results with

very high accuracy and consistency. These two developments improved the accuracy and repeatability of manual analysis but did not address its labor intensive, time-consuming and expensive nature. The fourth stage of this thesis addressed these limitations by automating the analysis. To do this, we developed an Automated Off-line Respiratory Event Detector (AORED) that analyses respiratory patterns by comparing metrics of respiratory behavior to thresholds to determine the presence of patterns. Optimal threshold values were selected using Receiver Operating Characteristics (ROC) analysis using manual analysis results as reference. AORED analysis agreed well with the "gold standard" manual analysis. However, its thresholds were based on the results of manual analysis so its performance will be influenced by the limitations of manual scoring. The fifth stage of the work addressed this through the development of AUREA, an Automated Unsupervised Respiratory Event Analysis system that applies unsupervised, K-means clustering to the metrics of respiratory behavior to classify the respiratory patterns. K-means requires no human intervention to work, so AUREA is completely automated and is not affected by the limitations of manual scoring. The validation results showed that AUREA had substantial accuracy (significantly higher than AORED), and almost perfect consistency.

The contributions from this work will impact the study of infant respiratory patterns and POA in several important ways: (i) the public data library allows investigators to contribute to the study of POA and the analysis of infant cardiorespiratory signals without the need for data acquisition, which is especially relevant for investigators who do not have access or resources to acquire clinical data, or those who focus only on technical aspects such as signal processing; (ii) the methodology establishes a new paradigm for the study of cardiorespiratory data by classifying every sample, which opens the possibility to apply advanced techniques such as time series or time-frequency analyses to the study of respiratory patterns; (iii) these techniques enable the study of the relationship between postoperative respiratory patterns and the occurrence of POA, which could in turn be used to develop a predictor of POA risk; (iv) the manual scoring tools give researchers the ability to study infant respiratory patterns even if they do not have the resources of a sleep laboratory, since these tools are readily available at no cost, and yield a highly accurate and consistent "gold standard" classification of the respiratory patterns; (v)

AUREA enables studies of infant respiration involving large amounts of data (e.g. multi-institutional, longitudinal) because it can analyze the respiratory patterns in an accurate, consistent, objective, fast, and low cost fashion; (vi) AUREA has the potential for real-time implementation, so it could be used to monitor infants at the bedside to provide more detailed, instantaneous information about the respiratory patterns compared to conventional clinical monitors; moreover, (vii) AUREA can be used to analyze adult data and be used in the study of Obstructive Sleep Apnea Syndrome.

Résumé

La récupération post-chirurgicale et anesthésique des nourrissons présente un risque d'Apnée Postopératoire (APO) qui est potentiellement mortelle. La prédiction d'une APO est impossible, par conséquence tous les nourrissons d'âge post-menstruel (APM) ≤ 60 semaines doivent être surveillés à hôpital, pendant au moins 12 heures après la fin de l'opération. Des données ont montré un lien entre des schémas respiratoires postopératoires anormaux et l'occurrence d'une APO. L'étude de ces schémas pourrait ainsi se révéler utile quant à la prédiction d'un risque d'APO chez le nourrisson, mais également à la prédiction du temps nécessaire pour que ce risque disparaisse.

L'étude complète des schémas respiratoires postopératoires a été limitée par deux facteurs principaux. Tout d'abord, aucun échantillon de données représentatives provenant de nourrissons à haut risque d'APO n'est disponible publiquement pour les chercheurs, et donc toute étude sur l'APO doit impérativement comporter une phase d'acquisition de données. Ceci implique planification, approbation et exécution, et ces tâches nécessitent de vastes ressources. Par conséquence, le nombre d'enfants participant à l'étude est limité par le budget disponible. Deuxièmement, aucun outil dédié à l'analyse complète des schémas respiratoires n'est actuellement disponible. La méthode généralement acceptée, le Score Manuel Conventionnel (SMC), est effectuée par des opérateurs experts qui suivent les directives de l'Académie Américaine de la Médecine du Sommeil (AAMS). Le SMC a plusieurs limitations: une faible reproductibilité intra- et inter-opérateur, ainsi que beaucoup d'efforts, de temps et d'argent. De plus, le SMC ne produit pas une analyse complète des schémas respiratoires, mais plutôt une liste d'événements "cliniquement pertinents", accompagnés du moment de leur occurrence. Ainsi, un schéma considéré comme n'étant pas "cliniquement pertinent" d'après les directives de l'AAMS n'est pas évalué et ne peut être analysé.

Cette thèse se penche sur ces limitations en créant une banque de données et en développant des outils pour l'analyse complète de schémas respiratoires chez le nourrisson. Ce travail se découpe en 5 phases. En premier lieu, nous avons acquis une banque de données représentatives comprenant des signaux cardio-respiratoires de nourrissons à risque d'APO, et avons rendu ces

données disponibles au public. Deuxièmement, nous avons développé une série d'outils permettant d'évaluer manuellement les schémas respiratoires des nourrissons de façon efficace, reproductible et fiable. Nous avons démontré que l'utilisation de ces outils produisait une analyse précise et robuste, et améliorait la reproductibilité intra- et inter-opérateur. Troisièmement, nous avons développé une méthode basée sur l'algorithme Espérance-Maximisation (EM) pour combiner les analyses de plusieurs marqueurs manuels afin de minimiser les effets de la variabilité intra- et inter-opérateur, et produire des résultats "gold standard" plus exactes et robustes. Ces deux développements ont amélioré la précision et la reproductibilité de l'analyse manuelle, mais n'ont pas résolu le problème du temps, du coût et des efforts nécessaires. La quatrième étape de cette thèse s'est plus particulièrement penchée sur ces limitations en automatisant l'analyse. Pour cela, nous avons développé un Détecteur d'Évènements Respiratoires Automatisé Hors-ligne (AORED en anglais) qui analyse les schémas respiratoires en comparant des métriques du comportement respiratoire à des seuils afin de déterminer la présence ou non de schémas. Les valeurs optimales des seuils ont été sélectionnées avec l'aide de l'analyse ROC (de l'anglais Receiver Operating Characteristics), en utilisant l'analyse manuelle comme référence. L'analyse AORED a montré une très bonne corrélation avec l'analyse manuelle "gold standard". Cependant, les seuils étaient basés sur les résultats de l'analyse manuelle, donc sa performance était influencée par les limitations du score manuel. La cinquième étape du travail s'est intéressée à ce problème avec le développement d'AUREA (en anglais), un système d'Analyse d'Évènements Respiratoires Automatisée et non-Supervisée, utilisant un clustering non supervisé, basé sur les K-moyennes, et des métriques du comportement respiratoire afin de classifier les schémas respiratoires. L'algorithme des K-moyennes n'a besoin d'aucune intervention humaine pour fonctionner. Ainsi, AUREA est complètement automatisé et n'est pas affecté par les limitations du score manuel. Les résultats de validation ont démontré qu'AUREA avait une très bonne précision (significativement plus élevée qu'AORED) et une très grande robustesse.

Les contributions de cette thèse auront un impact sur l'étude des schémas respiratoires chez le nourrisson et l'APO de plusieurs façons: (i) la banque de données disponible publiquement permet aux chercheurs de contribuer davantage à l'étude de l'APO et l'analyse de signaux

cardio-respiratoires du nourrisson, sans avoir à acquérir plus de données ; ceci est particulièrement pertinent pour les chercheurs qui n'ont pas accès à, ou les ressources pour acquérir, des données cliniques, ou ceux dont le centre d'intérêt principal traite les aspects techniques du traitement des signaux; (ii) la méthodologie établit un nouveau paradigme pour l'étude des signaux cardio-respiratoires en classifiant chaque donnée, ce qui permet d'appliquer des techniques avancées telles que l'analyse de séries temporelles ou temps-fréquence sur l'étude des schémas respiratoires; (iii) ces techniques rendent possible l'étude de la relation entre les schémas respiratoires postopératoires et l'occurrence de l'APO, qui par la suite pourrait être utilisée pour développer un modèle de prédiction de l'APO; (iv) les chercheurs ont à leur disposition les outils nécessaires pour évaluer manuellement et étudier les schémas respiratoires chez les nourrissons, même s'ils n'ont pas les ressources pour un laboratoire de sommeil; ces outils sont en effet facilement accessibles, gratuitement, et produisent une classification "gold standard" des schémas respiratoires, très précise et robuste; (v) AUREA permet d'étudier la respiration du nourrisson à une plus grande échelle (études multicentriques, longitudinales), car le système peut analyser les schémas respiratoires de façon précise, robuste, objective, rapide et à faible coût; (vi) AUREA a le potentiel d'être exécuté en temps-réel et pourrait ainsi être utilisé au chevet les nourrissons, pour mieux les surveiller et permettant de fournir instantanément des informations détaillées sur les schémas respiratoires, en comparaison à des systèmes de monitoring cliniques conventionnels; (vii) AUREA peut analyser des données provenant d'adultes, et ainsi être utilisé dans l'étude du Syndrome de l'Apnée du Sommeil Obstructive.

Acknowledgements

I would like to thank my supervisors Drs. Robert E. Kearney and Karen A. Brown for the opportunity to carry out my Ph.D. studies under their guidance and support.

I also thank the members of my Ph.D. advisory committee, Drs. Henrietta L. Galiana, Doina Precup, and Gilles Plourde, for their guidance, suggestions, and constructive criticism.

I would like to thank the members of the CardioRespiratory Infant Behavior team for their insightful discussions and useful feedback.

Special thanks to Drs. Cédric Clouchoux, Anne-Marie Lauzon, and Wissam Shalish, and to Pascale Gourdeau and Tania Lozoya Moreno, for their assistance in the translation of the thesis abstract; to Dr. Gianluca Bertolizio for his assistance in the acquisition of data from 2 infants; and to Dr. Ross Wagner, who guided me during the maintenance of the data acquisition system.

I am grateful to all participants in the study and the nurses of the Montreal Children's Hospital for their cooperation for the development of this work.

I also thank Pina Sorrini, Nancy Abate, Daniel Caron, and all other administrative members of the Department of Biomedical Engineering at McGill University, for their assistance and support.

I am truly grateful to my wife Ana Cecilia Puon Diaz, my parents Carlos Salvador and Susana, and my sister Paulina, for their love, affection and support, which led me to the culmination of this work.

Special thanks to the following agencies that provided financial support to pursue my research: Mexican National Council for Science and Technology (CONACYT, www.conacyt.gob.mx), Natural Sciences and Engineering Research Council of Canada (NSERC, www.nserc-crsng.gc.ca), and Queen Elizabeth Hospital of Montreal Foundation Chair in Pediatric Anesthesia, McGill University Faculty of Medicine (www.mcgill.ca/medicine/faculty-medicine).

Preface

Drs. Robert E. Kearney (REK) and Karen A. Brown (KAB) from McGill University have collaborated over the last decade to develop methods to comprehensively analyze infant respiratory patterns. One of the main goals of this research has been to enable the study of respiratory patterns of infants recovering from surgery and anesthesia, since these infants are at increased risk of life-threatening apneic events. To this end, this collaboration has: (i) developed the technology required to acquire cardiorespiratory data from infants in the recovery room following surgery [1, 2]; (ii) developed a series of algorithms for the automated detection of key respiratory patterns including pause, movement artifact, thoraco-abdominal asynchrony, and synchronous breathing [3-6]; and (iii) demonstrated the potential clinical application of these algorithms in the evaluation of infants at risk of postoperative apnea (POA) [7].

Summary of Original Contributions

I expanded this research during my Ph.D. studies by: (i) acquiring representative data from infants recovering from surgery and anesthesia, (ii) developing a formal methodology to comprehensively analyze infant respiratory patterns with high accuracy and consistency, and (iii) developing methods to perform the analysis automatically. This thesis describes my work in detail.

Chapter 3 describes a library of clinical data acquired for the development of this project. These data represent a valuable collection of cardiorespiratory signals because they: (i) are representative of infants at risk of POA; and (ii) were acquired continuously, starting immediately after surgery and lasting for up to 12 h. I made this data fully available to the public, without restriction. There is no other similar POA data available, so in addition to the development of this thesis, this contribution will enable the development of new methods to analyze infant cardiorespiratory data, and will also help advance the clinical understanding of POA.

Chapter 4 presents a set of tools I developed to assist manual scoring of infant respiratory data. Conventional manual scoring (CMS) has been the preferred analysis for respiratory data [8], but

it has been limited by several factors, especially low intra- and inter-scorer repeatability [9]. I developed the tools in Chapter 4 to improve manual scoring to make it a more reliable, "gold standard" analysis. The tools, which I made publicly available [10-12], include: (i) definitions and scoring rules for 6 unique, mutually-exclusive infant respiratory patterns; (ii) RIPScore, a software to apply these rules to infant data; (iii) a curated library of segments representative of the 6 respiratory patterns; (iv) a fully automated training protocol; and (v) a quality control method to monitor scorer performance over time. The tools allow to comprehensively analyze infant respiratory patterns in a continuous, sample-by-sample fashion, while also allow to establish and maintain high intra- and inter-scorer repeatability.

Chapter 5 describes a method I developed to combine the analyses from multiple, manual scorers to further reduce the variability inherent to manual scoring. By using this post-processing method, it is possible to obtain an analysis of the respiratory patterns that has excellent accuracy and consistency, and that significantly improves from results produced by individual, manual scorers. The results obtained by using this method represent a much improved "gold standard" analysis of the respiratory patterns, compared to those obtained with CMS.

Chapter 6 presents a first attempt to automate the analysis of infant respiratory patterns. In this Chapter I introduce AORED, an Automated Off-line Respiratory Event Detector that combines the algorithms previously developed by the group of KAB and REK, to automatically analyze infant respiratory patterns. AORED classifies respiratory patterns by comparing metrics of respiratory behavior to thresholds; the thresholds are selected based on a representative sample of manually analyzed data. AORED classifies the respiratory patterns on a sample-by-sample basis, is robust in high noise conditions, and is amenable for real-time implementation.

Chapter 7 describes AUREA, an Automated Unsupervised Respiratory Event Analysis system. AUREA makes use of clustering, an unsupervised learning technique, to automatically classify respiratory patterns. Because of this, AUREA requires no human intervention to work, contrary to AORED that required a sample of manually scored data to determine classification thresholds. This makes AUREA a fully objective, completely automated method for analysis of infant respiratory patterns. AUREA reduces the analysis time and costs because it eliminates the need

for manual scoring, and also delivers an analysis with near-perfect consistency and substantial accuracy. Similar to AORED, AUREA is also amenable for real-time implementation.

Contribution of Authors

I carried out the work in Chapters 3 to 7 in collaboration with other authors. The author contributions are described next for each Chapter.

Chapter 3

I wrote the Chapter; acquired data from 15 of the 24 infants recruited; scanned and transcribed into text files all handwritten annotations obtained during data acquisition; stored these files together with the acquired physiological data into a central repository; and made the data publicly available after inspecting and anonymizing them by removing all possible identifiers (as indicated in [13]). I also maintained the custom built data acquisition system by calibrating the pulse oximeter and replacing the battery once.

KAB obtained initial approval and annual renewal of the studies; recruited all subjects and obtained parental consent; acquired data from the 24 infants recruited; and provided comments, corrections and suggestions for the written Chapter.

Dr. Gianluca Bertolizio (GB) assisted with data acquisition from 2 infants.

REK provided comments, corrections and suggestions for the written Chapter.

Chapter 4

I conceived and designed the study in conjunction with KAB and REK. I developed RIPScore, compiled it to work as a standalone application, made its source code publicly available, and wrote its manual. I coordinated all manual scorers, stored their results in a central repository, and analyzed all manual analysis results. I created all figures and tables, wrote the Chapter, submitted it for publication to *PLOS ONE*, and was the corresponding author.

KAB and I obtained a second study approval that was necessary to secure funding for recruitment of 2 manual scorers.

KAB provided feedback for the development of RIPScore; manually analyzed the clinical data from Chapter 3; and provided comments, corrections and suggestions for the manuscript.

GB manually analyzed the clinical data from Chapter 3; and provided comments, corrections and suggestions for the manuscript.

REK provided feedback for the development of RIPScore; and provided comments, corrections and suggestions for the manuscript.

Chapter 5

I conceived the algorithm to combine analyses from multiple, manual scorers using Expectation-Maximization, and implemented it in MATLAB; designed the simulations and the study of clinical data; carried out all simulations and evaluations using clinical data; analyzed the results; generated all figures and tables; and wrote the manuscript.

KAB and REK collaborated in the design of the simulations and the study of clinical data, and provided comments and suggestions for the manuscript.

Chapter 6

Initial work in this study was performed by Ahmed A. Aoude (AAA). He conceived the algorithm that combines detectors to classify the respiratory patterns, and implemented the individual, automated detectors in MATLAB. KAB manually analyzed the clinical data, and AAA used this manual analysis to carry out a preliminary validation of the detectors. AAA also wrote a draft of the manuscript in collaboration with KAB, REK, and Dr. Henrietta L. Galiana (HLG), which incorporated preliminary versions of the figures and Tables 6.1 and 6.2.

Following the completion of AAA master's degree, it was determined that substantial additional algorithmic development, simulations and analysis were required to bring level of the manuscript

to that required for publication. I was asked to undertake this as part of my Ph.D. research. To this end I:

(i) standardized the filter bank from Table 6.1 so that all filters were the same order, and had the same peak-to-peak ripple in the pass band and minimum attenuation in the stop band;
(ii) redesigned the simulation study to make the movement artifact simulation consistent with previous work [4], make the simulation of asynchronous-breathing more realistic by adding a transition window during the phase shift, and increase the number of realizations;
(iii) improved the analysis of clinical data by incorporating the area under the Receiver Operating Characteristics (ROC) curve as a metric of detector performance, and devised and implemented a method to obtain optimum detector thresholds based on ROC analysis;
(iv) implemented the full, combined AORED classification algorithm in MATLAB;
(v) organized all MATLAB scripts, clinical data and manual scoring results in a central repository;
(vi) carried out the final analyses, including all simulations, evaluation of individual detector performance, and assessment of agreement between AORED and the manual scorer;
(vii) designed and generated all figures (except Figs. 6.1 and 6.3 prepared by AAA which were slightly modified), as well as Table 6.3 to reflect all final results;
(viii) rewrote the manuscript with input from KAB and REK to describe the updated methods, results, and interpretation of the findings; and
(ix) acted as the corresponding author when the paper was submitted to *IEEE Trans Biomed Eng*.

In recognition of the fact that both AAA and I made major contributions to the paper, all authors agreed that AAA would be listed as the first author and I as the senior author.

Chapter 7

I conceived the metrics of respiratory behavior and all algorithms underlying AUREA, and implemented them in MATLAB. I designed and carried out the study; designed and generated all figures and tables; analyzed and interpreted the results; and wrote the manuscript.

KAB and REK provided comments, suggestions, and corrections for the manuscript, figures and tables.

Publications

I have prepared several publications and conference presentations throughout the course of my Ph.D. studies, including:

Journal Articles

(i) A. A. Aoude, R. E. Kearney, K. A. Brown, H. L. Galiana, and **C. A. Robles-Rubio**, "Automated Off-Line Respiratory Event Detection for the Study of Postoperative Apnea in Infants," *IEEE Trans Biomed Eng*, vol. 58, pp. 1724-1733, 2011.

(ii) **C. A. Robles-Rubio**, J. Kaczmarek, S. Chawla, L. Kovacs, K. A. Brown, R. E. Kearney, and G. M. Sant Anna, "Automated analysis of respiratory behavior in extremely preterm infants and extubation readiness," *Pediatr Pulmonol*, vol. 50, pp. 479-486, 2015.

(iii) **C. A. Robles-Rubio**, G. Bertolizio, K. A. Brown, and R. E. Kearney, "Scoring Tools for the Analysis of Infant Respiratory Inductive Plethysmography Signals," *PLoS ONE*, vol. 10, p. e0134182, 2015.

Conference Articles

(i) **C. A. Robles-Rubio**, K. A. Brown, and R. E. Kearney, "Automated Unsupervised Respiratory Event Analysis," in Conf Proc 33rd IEEE Eng Med Biol Soc, Boston, USA, 2011, pp. 3201-3204.

(ii) **C. A. Robles-Rubio**, K. A. Brown, and R. E. Kearney, "Detection of Breathing Segments in Respiratory Signals," in Conf Proc 34th IEEE Eng Med Biol Soc, San Diego, USA, 2012, pp. 6333-6336.

(iii) **C. A. Robles-Rubio**, K. A. Brown, and R. E. Kearney, "A New Movement Artifact Detector for Photoplethysmographic Signals," in Conf Proc 35th IEEE Eng Med Biol Soc, Osaka, Japan, 2013, pp. 2295 - 2299.

(iv) **C. A. Robles-Rubio**, K. A. Brown, G. Bertolizio, and R. E. Kearney, "Automated Analysis of Respiratory Behavior for the Prediction of Apnea in Infants following General Anesthesia," in Conf Proc 36th IEEE Eng Med Biol Soc, Chicago IL, USA, 2014, pp. 262-265.

Conference Abstracts

(i) **C. A. Robles-Rubio**, R. E. Kearney, and K. A. Brown, "Automated pause frequency estimation to assess the risk of Postoperative Apnea in infants," presented at the 12th Int Symp Sleep Breath, Barcelona, Spain, 2011.

(ii) **C. A. Robles-Rubio**, J. Kaczmarek, K. A. Brown, R. E. Kearney, and G. M. Sant'Anna, "Prediction of Extubation Readiness in Extreme Preterm Infants using Automated Analysis of the Respiratory Pattern," presented at the Pediatr Acad Soc Annu Meet, Boston, USA, 2012.

(iii) **C. A. Robles-Rubio**, R. E. Kearney, and K. A. Brown, "Automated Classification of Pauses, Breathing and Movement Artifacts in Infant Respiratory Data," presented at the Soc Anesth Sleep Med Annu Conf, Washington DC, USA, 2012.

(iv) **C. A. Robles-Rubio**, R. E. Kearney, and K. A. Brown, "Inclusion of Lissajous Plot on Scoring Software Improves Classification of Thoracoabdominal Asynchrony," presented at the 13th Int Symp Sleep Breath, Montreal, Canada, 2013.

(v) **C. A. Robles-Rubio**, K. A. Brown, and R. E. Kearney, "Improving the Accuracy of Manual Analysis of Respiratory Behavior by using Expectation-Maximization to Combine Results from Multiple Scorers," presented at the 37th Annu Int Conf IEEE Eng Med Biol Soc, Milan, Italy, 2015.

Manuscripts in preparation

(i) **C. A. Robles-Rubio**, K. A. Brown, and R. E. Kearney, "Improving Manual Scoring of Respiratory Patterns using Expectation-Maximization," to be submitted to *IEEE Trans Biomed Eng*.

(ii) **C. A. Robles-Rubio**, K. A. Brown, and R. E. Kearney, "Automated Unsupervised Analysis of Infant Respiratory Patterns," to be submitted to *IEEE Trans Biomed Eng*.

List of Acronyms

Acronym	Definition
AAP	American Academy of Pediatrics
AASM	American Academy of Sleep Medicine
ABD	Abdomen
AHI	Apnea-hypopnea index
AI	Apnea index
ANN	Artificial Neural Network
AOP	Apnea of prematurity
AORED	Automated Off-line Respiratory Event Detector
ASB	Asynchronous-breathing
AUC	Area under the ROC curve
AUREA	Automated Unsupervised Respiratory Event Analysis
BAD	Bad Data
CHIME	Collaborative Home Infant Monitoring Evaluation
CMS	Conventional manual scoring
ECG	Electrocardiogram
EDR	ECG derived respiration
EM	Expectation-Maximization
IMC	Isovolume Maneuver Calibration
IP	Impedance Pneumography
IRB	Institutional Review Board
IS	Individual Scorer
IQR	Interquartile range
LDA	Linear discriminant analysis
McCRIBS	McGill CardioRespiratory Infant Behavior Software
MCH	Montreal Children's Hospital

Acronym	Definition
MV	Majority vote
MVT	Movement artifact
NAF	Nasal airflow
OSAS	Obstructive Sleep Apnea Syndrome
PACU	Postanesthesia Care Unit
PAU	Respiratory pause
PDF	Probability density function
PMA	Postmenstrual age
POA	Postoperative Apnea
PPG	Photoplethysmography
PSG	Polysomnography
QB	Quiet breathing
QDC	Qualitative Diagnostic Calibration
QDA	Quadratic discriminant analysis
RCG	Ribcage
REF	Reference scorer
REM	Rapid eye movement
RIP	Respiratory Inductive Plethysmography
RMS	Root mean square
ROC	Receiver Operating Characteristics
SAT	Blood oxygen saturation
SC1	Scorer 1
SC2	Scorer 2
SC3	Scorer 3
SIDS	Sudden Infant Death Syndrome
SIH	Sigh
SNR	Signal-to-noise ratio

Acronym	Definition
STFT	Short-Time Fourier Transform
SYB	Synchronous-breathing
TAA	Thoraco-abdominal asynchrony
UNK	Unknown
XOR	Exclusive OR

List of Symbols

Note that symbol usage may be different in different chapters due to publication issues.

Chapter	Symbol	Description
2	ϕ	Degree of thoraco-abdominal asynchrony
2	μ	Air viscosity
2	σ	Standard deviation operator
2	F	Flow of air
2	M_{RCG} , M_{ABD}	Magnetometer signals from ribcage and abdomen
2	P	Pressure drop occurring along a length l on an infinitely long tube of diameter d
2	P_D , P_{FA}	Probabilities of detection and false alarm
2	SUM , DIF	Signals representing the sum and difference of rcg and abd
2	$\overline{\lvert SUM \rvert}$, $\overline{\lvert DIF \rvert}$	Rectified averages of SUM and DIF
2	V_{RCG} , V_{ABD}	Respiratory volume of ribcage and abdomen
2	V_{TOT}	Respiratory volume
2	a , b , c , c_{RCG} , c_{ABD}	Linear model parameters for V_{RCG} and V_{ABD}
2	d	Normalized distance of any point on the ROC curve to the chance line (scaled to the range 0 to 1)
2	k , m	Proportionality and scale parameters for RIP calibration
2	rcg , abd	Raw RIP signals from ribcage and abdomen
4	Θ	Set including the 6 unique respiratory patterns (SYB, ASB, SIH, PAU, MTV, UNK)
4	κ	Fleiss' statistic for assessment of inter-scorer agreement
4	Cn	Consensus RIP pattern function
4	N_T	Length of the transition window T
4	N_i	Number of samples with consensus pattern i
4	N_j	Number of times the N_i samples had been assigned to pattern j
4	$\mathbf{P}$	Confusion matrix of manual scorers

Chapter	Symbol	Description
4	$P_{i,j}$	Element of $\mathbf{P}$ with consensus pattern i and scored pattern j
4	T	Transition window for concatenated segments
4	x_k	A data sample
5	α , β	Parameters of the beta distribution used in the simulations
5	ε	Convergence error
5	κ	Fleiss' statistic for assessment of inter-scorer agreement
5	C	Number of unique, mutually exclusive RIP patterns
5	I	Indicator function
5	N	Number of samples of RIP data
5	P_c	Marginal probability of pattern c within the whole dataset
5	$\mathbf{Q}^{real}$	Combined confusion matrix from 3 manual scorers
5	$\mathbf{Q}^{s}$	Confusion matrix for scoring sequence s
5	$\mathbf{Q}^{s,sim}$	Simulated confusion matrix for simulated scoring sequence s
5	R	Number of scorers
5	S	Number of individual scoring sequences
5	$T[n]$	Most likely pattern of sample n
5	$\mathbf{T}^{sim}$	Simulated true pattern vector
5	$\hat{T}_{EM}$	Expectation-Maximization estimate
5	$\hat{\mathbf{T}}_{EM}^{sim}$	Expectation-Maximization estimate in simulation study
5	$\hat{T}_{IS}^{s}[n]$	Pattern assigned by scoring sequence s to sample n
5	$\hat{\mathbf{T}}_{IS}^{s,sim}$	Simulated scoring sequence
5	$\hat{T}_{MV}$	Majority vote estimate
5	$\hat{\mathbf{T}}_{MV}^{sim}$	Majority vote estimate in simulation study
5	$W_c[n]$	Probability that sample n has the pattern c
6	$\wp^{rc}$, $\wp^{ab}$	Median power of all segments of length N_P in rc_f and ab_f
6	$\wp_i^{rc}$, $\wp_i^{ab}$	Power of a segment of length N_M in rc_i and ab_i
6	ΔT_P , ΔT_M , ΔT_A	Detection delay for pause, movement and asynchrony

Chapter	Symbol	Description
6	ϕ	Asynchrony test statistic
6	ϕ_E	Asynchrony estimation error
6	ϕ_{true}	Actual simulated asynchrony
6	γ_A	Threshold used with A
6	$\gamma_{A_{opt}}$	Optimum threshold for A
6	γ_M^{rc} , γ_M^{ab}	Thresholds used with M^{rc} and M^{ab}
6	$\gamma_{M_{opt}}^{rc}$, $\gamma_{M_{opt}}^{ab}$	Optimum thresholds for M^{rc} and M^{ab}
6	γ_P^{rc} , γ_P^{ab}	Thresholds used with P^{rc} and P^{ab}
6	$\gamma_{P_{opt}}^{rc}$, $\gamma_{P_{opt}}^{ab}$	Optimum thresholds for P^{rc} and P^{ab}
6	ρ_{11} , ρ_{12} , ρ_{13}	Scaling parameters used to construct RC
6	ρ_{21} , ρ_{22} , ρ_{23}	Scaling parameters used to construct AB
6	F_s	Sampling Frequency
6	M^{rc} , M^{ab}	Movement detectors used with rc and ab
6	N_P , N_M , N_A	Window lengths used with p^{rc} , p^{ab} , m^{rc} , m^{ab} , and ϕ
6	P , M , A	Overall detectors for Pause, Movement and Asynchrony
6	P_D , P_{FA}	Probabilities of detection and false alarm
6	P^{rc} , P^{ab}	Pause detectors used with rc and ab
6	RC , AB	Simulated RIP signals for Ribcage and Abdomen
6	T_P , T_M , T_A	Start time for simulated pause, movement and asynchrony
6	T_{PD} , T_{MD} , T_{AD}	Actual detection time of pause, movement and asynchrony
6	T_n	Simulated signals period
6	W_L	Length parameter for w
6	$f_{\max}$	Breathing frequency estimate
6	g_1 , g_2	Simulated electronic noise in RC and AB
6	$i_{\max}$	Filter from the bank with the output with highest power
6	k	Degree of simulated asynchrony
6	m^{rc} , m^{ab}	Movement test statistics for rc and ab

Chapter	Symbol	Description
6	$\tilde{m}_1$, $\tilde{m}_2$	Simulated movement process in RC and AB
6	n_0	Sample where simulation of asynchrony begins
6	p^{rc} , p^{ab}	Pause test statistics for rc and ab
6	rc , ab	Raw RIP signals from ribcage and abdomen
6	rc_S , ab_S	Selectively filtered version of rc and ab
6	rc_{bp} , ab_{bp}	Band-pass filtered rc and ab
6	rc_i , ab_i	Output from the ith filter in the bank for rc and ab
6	w	Transition window for simulation of asynchrony
7	ΔT_S , ΔT_E	Detection start and end delays
7	$\alpha_{cluster}$, α_{metric}	Cluster and metric outlier detection parameters
7	$\boldsymbol{\gamma}_{jm}$	Point in the decision boundary produced by K-means
7	κ	Fleiss' statistic for assessment of inter-scorer agreement
7	$\mathbf{v}_{jm}$	Vector normal to the K-means decision boundary
7	C_j , C_m	Clusters produced by K-means
7	FP^{PAU}	Number of samples incorrectly classified as PAU
7	L	K-means class-assigning function
7	N_B	Length of the window used to estimate the power of SUM_{HP} and DIF_{HP}
7	N_{DT}	Length of the trend removal window
7	N_{MA}	Length of the moving-average filter
7	N^{PAU}	Number of samples with "gold standard" score not equal to PAU
7	N_{QV} , N_{QRMS}	Length of the quantile estimation windows
7	N_{RMS}	Length of the root mean square estimation window
7	N_{SMO}	Length of the smoothing window
7	N_V	Length of the variance estimation window
7	P	Number of input metrics for K-means
7	P_D , P_{FA}	Probabilities of detection and false alarm

Chapter	Symbol	Description
7	P_D^{ASB} , P_{FA}^{ASB}	Probabilities of detection and false alarm for the ASB pattern
7	P_D^{BAD} , P_{FA}^{BAD}	Probabilities of detection and false alarm for the BAD pattern
7	P_D^{PAU} , P_{FA}^{PAU}	Probabilities of detection and false alarm for the PAU pattern
7	P_D^{SYB} , P_{FA}^{SYB}	Probabilities of detection and false alarm for the SYB pattern
7	P^{PAU}	Total number of samples "gold standard" scored as PAU
7	RCG , ABD	De-trended ribcage and abdominal RIP signals
7	RCG_B , ABD_B	Binary ribcage and abdominal signals
7	RCG_{MA} , ABD_{MA}	Moving-average filtered ABD
7	RCG_{SMO}	Smoothed RCG
7	RCG_T	Low frequency trend of RCG_{raw}
7	RCG_{raw} , ABD_{raw}	Raw ribcage and abdominal RIP signals
7	SUM , DIF	Sum and difference of the binary RCG and ABD signals
7	SUM_{HP} , DIF_{HP}	High-pass filtered SUM and DIF
7	TP^{PAU}	Number of samples correctly classified by the automated method as PAU
7	T_S , T_E	Segment start and end times
7	$\hat{T}_S$, $\hat{T}_E$	Detection start and end times
7	b^+ , b^-	Synchronous- and asynchronous-breathing metrics
7	c_j , c_m	Centroids of C_j and C_m
7	d	Normalized distance of any point on the ROC curve to the chance line (scaled to the range 0 to 1)
7	f_s	Sampling frequency
7	npp_{RCG} , npp_{ABD}	Non-periodic power of RCG and ABD
7	nv_{RCG} , nv_{ABD}	Normalized variance of RCG and ABD
7	rms_{RCG} , rms_{ABD}	Root mean square of RCG_{MA} and ABD_{MA}
7	$rms_{RCG}^{(q)}$, $rms_{ABD}^{(q)}$	Q^{th} quantile of rms_{RCG} and rms_{ABD}
7	v_{RCG}	Variance of RCG

Chapter	Symbol	Description
7	$v_{RCG}^{(q)}$	Q^{th} quantile of v_{RCG}
7	w_{jm}	K-means decision boundary weighting factor
7	$\mathbf{x}[n]$	A data sample

1. Introduction

Newborn infants, especially those born prematurely, may be affected by a number of respiratory conditions due to an underdeveloped respiratory system. One of the main respiratory problems faced by infants is apnea, defined as a period during which there is no respiratory airflow. If prolonged, apneic episodes are life threatening. In infants, four conditions are associated with apnea: Sudden Infant Death Syndrome (SIDS), neonatal Obstructive Sleep Apnea Syndrome (OSAS), Apnea of Prematurity (AOP), and Postoperative Apnea (POA). The focus of this thesis is on the development of general methods for the analysis of respiratory signals that are applicable to the study and management of POA.

Infants recovering from surgery and anesthesia are at increased risk of life threatening POA [14, 15]. Reports indicate that the prevalence of POA in these infants ranges from 3 % to 49 % [16-18], although this varies depending on the definition of apnea, monitoring technology, anesthetic technique, and method of analysis [17]. Many risk factors have been evaluated to date: age, weight, prematurity, surgical procedure, history of apnea, anemia, and the use of perioperative medications, including anesthetics, analgesics and opioids. Of these, a postmenstrual age (PMA) less than 60 weeks is the most important [14, 17, 19-23]. Thus, clinical guidelines recommend that all infants with PMA $\leq$ 60 weeks be continuously monitored postoperatively in a hospital setting [14].

However, POA occurs only in a minority of infants, so many at-risk infants are kept in hospital unnecessarily, which adds stress to families, increases costs, and consumes hospital resources reducing their availability. This is because there is as yet no way of predicting which infants will develop apnea, nor is possible to determine at what time following surgery the risk of apnea abates [17]. However, there is evidence that POA events are associated with abnormal postoperative respiratory patterns [24-26]. This suggests that an analysis of the underlying postoperative respiratory patterns could result in a better prediction of POA.

Some studies involving sleep laboratories have applied conventional manual scoring (CMS) [8, 27] to study POA and the postoperative respiratory patterns [24-26, 28-32]. However, CMS is labor intensive, subjective, very costly, and suffers from low intra-, and inter-scorer repeatability

[9]. Full automation of the analysis has not yet been achieved, and therefore the mainstay analysis continues to be CMS performed by sleep laboratory technicians [8].

Since POA events are rare, long postoperative records from many infants will be required to establish their relation to abnormal respiratory patterns. CMS is an inappropriate method to analyze such long records due to its limitations as results would be obscured by the inherent low intra- and inter-scorer repeatability, and the costs to analyze such long records would be very high since it is a very labor intensive, specialized task. Consequently, the study of POA and the postoperative respiratory patterns has not advanced appreciably recently.

1.1. Objective

The overall objective of this thesis was to improve the analysis of infant respiratory patterns by developing an automated, comprehensive, reliable (i.e., high accuracy), repeatable (i.e., high consistency), fast, and low-cost methodology to classify the respiratory patterns as a function of time. The intention of this methodology is to address the limitations associated with CMS, to enable the analysis of long infant data records, and support the comprehensive study of how POA, and other disorders of infant respiration, relate to abnormal respiratory patterns.

1.2. Thesis Overview

Chapter 2 provides a review of topics relevant to this thesis including: (i) postoperative apnea, (ii) respiratory monitoring technologies, (iii) current methods for respiratory pattern analysis, and (iv) existing cardiorespiratory datasets from infants. Chapter 3 describes the representative infant data acquired for the development of the thesis. The next two Chapters describe the development of a comprehensive methodology to manually analyze infant respiratory patterns with much higher repeatability than CMS, with the objective to provide a “gold standard” reference for evaluation of automated methods. Thus, Chapter 4 presents and validates a set of tools for manual scoring of infant respiratory inductive plethysmography (RIP) data, which improve upon CMS by making the analysis comprehensive, reliable, and repeatable. Chapter 5 shows how to enhance the accuracy and consistency of manual scoring results obtained using the tools developed in Chapter 4, by combining the analyses from multiple scorers using machine learning

methods. Then, the two following Chapters describe methods that automate the analysis of respiratory patterns using supervised and unsupervised machine learning. Chapter 6 describes and validates AORED, an Automated Off-line Respiratory Event Detector that analyses infant respiratory patterns by combining detectors of respiratory pauses, movement artifacts, and asynchronous-breathing. Chapter 7 presents and evaluates AUREA, an Automated Unsupervised Respiratory Event Analysis system that improves the classification of infant respiratory patterns from that of AORED, and requires no human intervention. Lastly, Chapter 8 discusses the original contributions of this thesis, implications of the findings, and avenues for future work, and provides a summarizing conclusion of the work.

2. Review of Relevant Literature

This Chapter reviews topics important to this thesis. Section 2.1 describes postoperative apnea (POA), provides a review of previous POA studies, and discusses the main limitations found in these studies. Section 2.2 reviews the technologies available for respiratory monitoring. Section 2.3 provides a survey of existing methods for the analysis of respiratory patterns. Section 2.4 describes an existing infant cardiorespiratory dataset and discusses its suitability for the study of respiratory patterns and POA. Section 2.5 summarizes the findings and provides the rationale for this thesis. Special attention is given to aspects relevant to newborn infants.

2.1. Postoperative Apnea

Apnea is a clinical event that is defined as an extended cessation of airflow during which there is no supply of fresh oxygen or elimination of carbon dioxide. If the pause in respiration is prolonged, it may cause a life-threatening fall in blood oxygen saturation and cardiac arrest. There are 3 types of apnea: (i) central, which originates from an interruption of the respiratory rhythm generator in the respiratory control centre resulting in no central drive to the respiratory muscles; (ii) obstructive, which is produced by a physical obstruction of the airway at the level of the pharynx or larynx [33]; and (iii) mixed, that is a combination of central and obstructive apneas [34].

Conditions that may predispose a young infant to apnea include: (i) drug administration, including anesthesia, (ii) gastro-esophageal reflux, (iii) central nervous system lesions, (iv) infection, (v) fluctuations in ambient temperature, (vi) cardiac abnormalities, (vii) immunization, (viii) metabolic derangements, (ix) anemia, (x) upper airway obstructions, (xi) abdominal distention, and (xii) chronic lung disease of prematurity [15, 34].

In infants there are 4 clinical conditions where apnea plays an important role: (i) postoperative apnea (POA), where life threatening apnea may occur in the immediate period following surgery and anesthesia; (ii) apnea of prematurity (AOP), which occurs in premature infants due to underdeveloped respiratory systems; (iii) Sudden Infant Death Syndrome (SIDS), that is the sudden, unexplained death of an infant, and generally occurs during sleep; and (iv) neonatal

Obstructive Sleep Apnea Syndrome (OSAS), which consists on repetitive episodes of obstructive apnea during sleep.

Postoperative apnea (POA) is a subset of Apnea of Infancy, a category defined by both the American Academy of Pediatrics (AAP), and the American Academy of Sleep Medicine (AASM) in its classification of sleep disorders [35]. POA was first described in 1982 by Steward [15] who hypothesized that immaturity represents a potential complicating factor that may persist as the child grows older. In Steward's study, infants who experienced apnea were younger and weighed less than those who did not. Based on these results, Steward recommended that all surgeries in young, ex-premature infants be performed as inpatients, so that they could be monitored postoperatively for at least 24 hours.

At present investigators do not understand the etiology of POA, but hypothesize that it likely arises from the stress of surgery and the influence of anesthetic drugs on the immature respiratory control system of the infant [14, 15, 36-38].

2.1.1. Occurrence Time of Apnea

In his retrospective study [15], Steward found that POA events occurred immediately after surgery, and up to 12 h postoperatively. Kurth et al. [14] observed that the majority (72 %) of infants who experienced prolonged POA had their first episode within 2 h after surgery; the remainder had their first POA episode from 2 h to 12 h postoperatively. Infants with postmenstrual age (PMA) between 32 to 40 weeks continued to have prolonged apneas from 12 h to 48 h postoperatively [14]. In contrast, infants with PMA from 48 to 55 weeks had prolonged apnea from admission to recovery room only up to 10 h postoperatively [14]. Later, Bell et al. [39] reported that POA events may occur up to 72 h after surgery. Therefore, it is generally recommended that infants recovering from surgery and anesthesia be continuously monitored in hospital for at least the first 12 h postoperatively, and those infants who exhibit POA episodes within the first 12 h should be monitored for an additional 12 h, and up to 72 h.

2.1.2. Type of Apnea

Initial studies of POA focused on the incidence of central apnea [14, 15, 20-22, 40-42]. However, airway obstruction is a frequent cause of apnea in preterm infants [43, 44]. Thus, Kurth and LeBard [19] prospectively studied 74 former preterm infants with PMA < 50 weeks, and found that more than 70 % of the POA episodes were central, 21 % to 24 % were mixed, and less than 10 % were obstructive in nature. Central and mixed apnea occurred in most infants with POA, while obstructive apnea occurred only in one third. All mixed apneas started with a central respiratory pause which was followed by airway obstruction. Additionally, blood oxygen desaturation was significantly larger and more frequent during mixed and obstructive apneas, compared to central apneas. The conclusion from this work was that airway obstruction is a frequent component of POA that leads to larger decreases in blood oxygen saturation (SAT) than apneas without obstruction.

Thus, the majority of POA events in infants can be expected to be central in nature; however, a significant proportion of apneas will involve an obstructive component. Since POA episodes with an obstructive component lead to larger drops in SAT, infants should be continuously monitored to detect these periods of airway obstruction.

2.1.3. Risk Factors

Several studies have investigated potential risk factors that predispose infants to POA, including: demographics, prematurity, anesthetic management, and the pre- and postoperative respiratory patterns. Four risk factors have been identified to date. The first and most important is age, specifically a PMA ≤ 60 weeks [14, 17, 37, 39, 41, 45, 46]. The second is prematurity [15, 37, 45], although full-term infants may also develop POA [24, 39, 47-50]. The third is anemia [41], particularly for infants with PMA > 43 weeks [17]. The fourth may be weight at surgery [51], however, only a single retrospective study identified this as an independent risk factor.

Due to these findings, it is generally recommended that all infants with PMA ≤ 60 weeks undergoing any surgery be placed in extended cardiorespiratory monitoring postoperatively.

No significant differences in the incidence of POA have been found in the following variables: (i) birth weight [14]; (ii) body temperature in the recovery room [14]; (iii) history of apnea [14, 19, 41]; (iv) use of perioperative drugs [14, 17]; and (v) surgical procedures, including inguinal hernia repair, orchiopexy, ventriculo-peritoneal shunt, bronchoscopy, esophagoscopy, tracheo-esophageal fistula repair, colostomy, circumcision, laparotomy, central line placement, and other miscellaneous minor procedures [14, 37]. However, these variables should be prospectively assessed using larger sample sizes to verify that they do not constitute a risk factor for POA [17].

2.1.4. Respiratory Patterns

Welborn et al. [20] found that preterm infants with PMA < 45 weeks had a higher incidence of postoperative periodic breathing. Later, Kurth et al. [14] studied the perioperative respiratory patterns of preterm infants (PMA ≤ 60 weeks) before, and after surgery with general anesthesia. They found that the patterns of infants who experienced POA were different pre- and postoperatively; the latter being characterized by breathing interspersed with respiratory pauses in the recovery room. Conversely, the respiratory patterns of infants without POA were similar pre- and postoperatively. The pre-operative respiratory patterns did not predict well the occurrence of POA. Similar results have been observed elsewhere in preterm [19, 39] and full-term infants [24].

These findings suggest that a comprehensive analysis of the postoperative respiratory patterns may provide insight about POA. We believe this analysis should aim to describe all respiratory patterns that an infant may display while recovering from surgery. Moreover, the analysis should describe the occurrence of these respiratory patterns as a sequence in time, as well as the properties of each occurrence including the pattern type and length. By having these properties it would be possible to study the relation between different patterns, their time of occurrence, and POA; which could help to determine which infants are at risk, and at what time after surgery the risk of POA abates.

2.1.5. Limitations of Studies of Postoperative Apnea

The study of POA has been limited by two main factors [17]: (i) non-standardized monitoring periods, and (ii) different methods to detect apneic events. This section describes these factors in detail.

2.1.5.1. Monitoring Period

The monitoring period in previous studies of POA has not been standardized. Most studies have monitored infants only postoperatively, but some have done it preoperatively as well. Preoperative monitoring has been done for at least 2 h, with a minimum of 30 min of sleep [14], or immediately before surgery [50]; and for at least 12 h in hospital [20] or in the infant's home [52]. The postoperative monitoring period has varied more, lasting 2 h [19], 4 h [20, 53], 12 h [20-22, 40, 41, 50, 52], or up to 24 h [45, 53]. Also, the start of the monitoring has varied from the immediate postoperative period [45, 50, 52] up to 30 min after surgery [14].

Preoperative monitoring has been found to have no predictive ability for POA [14, 19, 39], and so it may not be necessary. However, appropriate postoperative monitoring is essential to detect apneic events, and study their relation to postoperative respiratory patterns. Based on the literature [14, 39], the monitoring period should include at least the 12 h right after surgery. Studies that monitored for less than 12 h, or that started monitoring some time after surgery, may have missed important events, limiting the validity of their conclusions.

2.1.5.2. Monitoring Technology and Method to Detect Apneas

Another limitation of POA studies is the technology used to monitor infant respiration and detect apneic episodes. Retrospective studies retrieved the information from the annotations on patient charts [15, 45]. Other studies used alarm monitors, and events were recorded by nursing staff [39, 53]. Most studies monitored respiration continuously using the impedance pneumography signal [14, 19-22, 39-41, 50, 52], but only a few measured airflow [19, 50], or blood oxygen saturation [19, 50, 52]. In those studies where cardiorespiratory signals were acquired, the data was manually analyzed by one of the authors [14, 19, 50], pulmonologists [20-22, 40, 41], or a trained scorer [52].

The selection of methods to monitor respiration and detect apnea is important because of their different sensitivities. There is considerable variation in the number of apneas detected among POA studies, and this difference is related to the combination of monitoring technology and the method used to detect apnea [17]. The incidence of apnea is greater when recordings of cardiorespiratory signals are manually scored, than when estimated using annotations from nurses that describe the clinical state of the patient when the monitor alarmed. In fact, Bell et al. [39] reported that 80 % of infants who experienced POA were incorrectly classified as non-POA based on nurse observations. Thus, we think that for proper detection of POA, infants should be comprehensively monitored using continuous cardiorespiratory signals.

Moreover, infants with POA frequently have mixed apneas, and the obstructive component is difficult to detect using impedance pneumography without a measure of airflow [19]. Thus, previous studies [14, 15, 20, 37] likely overlooked the obstructive component of POA. The respiratory inductive plethysmograph (RIP) is an alternative sensor that measures respiratory movements of the ribcage and the abdomen, and does detect both central and obstructive apnea [8, 54-56]. Thus, we believe that POA studies should use RIP to monitor respiration, detect apneas, and distinguish between central and obstructive components.

Finally, there is a need to monitor SAT to determine the clinical significance of apneic events [52], and many early studies did not take this into account.

2.2. Respiratory Monitoring Technology

There are several sensors available to monitor infant respiration, and so it is important to review them to determine the best option for POA studies. These sensors measure airflow, concentration of carbon dioxide, or respiratory volume. Airflow is the rate of change (i.e., first derivative with respect to time) of volume, and so only one of these variables needs to be measured to obtain the other. The following sections describe the main sensors for each variable.

2.2.1. Airflow Sensors

In general, airflow sensors measure the flow coming from, or entering the airway opening [57]. The most common devices are the pneumotachograph and thermal sensors.

2.2.1.1. Pneumotachograph

The pneumotachograph is connected to the airway opening, such that inspired and expired air flows through the device. The internal structure of the device is designed to maintain the flow nearly laminar. Laminar airflow along a tube produces a pressure drop given by the Poiseuille equation

$$P = \frac{128 \mu l F}{\pi d^4}, \tag{2.1}$$

where P is the pressure drop occurring along a length l on an infinitely long tube of diameter d, and F is the flow of air with viscosity of μ [57]. Thus, measuring this pressure drop with a differential pressure transducer gives a measurement of airflow.

The pneumotachograph is considered the "gold standard" sensor for monitoring respiration because it is connected directly to the airway and measures airflow [58]. However, it has some important limitations. First, the tubes and connectors attached to the pneumotachograph greatly affect the relationship between differential pressure and airflow [57]. Thus, pneumotachographs must be calibrated with the same tubes and connectors that are used during measurement, and calibration must be repeated every time the system is reassembled [57]. Second, during prolonged use, secretions or condensation may collect inside the pneumotachograph and change the calibration. Consequently, pneumotachographs must be recalibrated after extended use [57]. Third, the space inside the pneumotachograph, a.k.a. dead space, can cause rebreathing due to the accumulation of expired carbon dioxide [57]. Consequently, It is necessary to apply an external, bias flow to clear expired gases from the transducer and respiratory circuit, and this bias flow must be removed from the measured airflow [57]. The magnitude of the bias flow that can be applied is limited by the measurement range of the pneumotachograph. Fourth, pneumotachographs must be attached directly to the airway, and so continuous, prolonged monitoring is uncomfortable and impractical, especially for infants in the recovery room whom are continuously handled by parents and nursing staff. For all these reasons, the pneumotachograph is not a viable sensor for long term studies of infant respiratory patterns.

2.2.1.2. Thermal Sensor

Thermal sensors are small, heated elements (e.g., thermistor, thermocouple) that are placed close to an airway opening to measure changes in temperature experienced due to contact with moving air. These temperature variations are approximately proportional to the square root of the air velocity [57], so they can be used to measure airflow.

Since they are smaller than a pneumotachograph, thermal sensors interfere less with respiration, and so represent a better option for monitoring airflow in infants for extended periods of time. The main disadvantages of thermal sensors are that they: (i) have a nonlinear relationship to airflow, (ii) require careful control of environmental temperature, (iii) cannot determine the direction of flow and so cannot distinguish between inspiration and expiration, and (vi) measure local air velocity, contrasted to pneumotachographs which measure volumetric flow [57-59]. For these reasons, thermal sensors are mostly used for qualitative determination of airflow [58].

2.2.2. Carbon Dioxide Sensor: Capnograph

The capnograph measures the concentration of carbon dioxide during respiration [60]. To do this, the capnograph emits infrared light that passes through a sample of breathed air and is detected by a photo-detector, and analyzes the frequency spectrum of the received light [60]. Carbon dioxide strongly absorbs infrared light while oxygen does not [61]. Consequently, an increase in the concentration of carbon dioxide during expiration will result in a drop of infrared power in the received signal. Conversely, a drop in carbon dioxide concentration during inspiration will result in increased infrared signal power.

Several factors limit the use of capnographs for monitoring infant respiration for extended periods of time: (i) if the tubing is not heated properly condensation may occur, changing the optical properties of the medium and thus affecting the measurements; (ii) tubing may get obstructed by moisture and/or sputum especially during extended monitoring sessions [62], and (iii) leaks in the system can lead to distorted measurements.

2.2.3. Respiratory Volume Sensors

Devices that measure the respiratory volume do so in one of two ways: (i) spirometers measure volume changes at the airway opening, and (ii) impedance pneumography and respiratory inductive plethysmography (RIP) measure changes in volume in the body surface (i.e., at the torso) [57]. These devices are described next.

2.2.3.1. Spirometer

A spirometer connects the airway opening to a sealed chamber. Inhalation and exhalation produce changes in the volume of air inside this chamber, and these changes are transduced to a volume signal [57]. The spirometer most commonly used with infants connects the airway to a hollow cylinder, open at the bottom, which floats inside a water-filled reservoir [57]. Thus, with each breath, the volume of air inside the cylinder varies displacing the cylinder which is transduced to an electrical volume signal using a potentiometer [57].

Spirometers have a number of limitations, including: (i) volume measurements are inaccurate if there are leaks in the system; (ii) the amount of water in the reservoir constrains the measurement range; (iii) ideally, the spirometer volume should be similar to the volume being measured to maximize the measurement accuracy, however commercially available spirometers are designed mostly for adults, children, and older infants, and so have much larger volumes than the low respiratory volumes of newborns; (iv) during prolonged monitoring excess of carbon dioxide must be eliminated after expiration to maintain a normal air composition and minimize rebreathing; and (v) similarly to pneumotachographs, spirometers are fully connected to the airway and may interfere and alter the natural respiratory patterns [57].

2.2.3.2. Impedance Pneumography

Impedance pneumography [63] uses surface electrodes similar to those used to record an electrocardiogram, to measure the changes in thoracic electrical impedance produced by respiratory movements [58, 64]. The underlying principle is that during inspiration, an increase in air volume reduces the electric conductivity of the chest by: (i) changing the electrical properties of the medium, and (ii) elongating the conductive paths [65]. This results in increased impedance. The opposite effect occurs during expiration.

The most common type of impedance pneumograph applies a small, constant, alternating-current, carrier signal to the subject [63, 66]. In this setup, the carrier peak to peak voltage is proportional to the magnitude of the thoracic impedance, and so impedance variations produced by respiration modulate the carrier amplitude. A demodulator is used to extract these impedance variations, yielding the respiratory signal [64]. This signal can then be scaled to match the output of a spirometer to represent respiratory volume [65].

This technique is non-invasive, and does not interfere with the natural breathing patterns, so it can be used for prolonged monitoring of infants [67]. However, it is prone to cardiogenic oscillations that may often be misinterpreted as breathing [30], leading to missed apneas. Moreover, impedance pneumography cannot detect obstructive apneas because it cannot distinguish between normal and paradoxical chest movements [30, 59]. Consequently, impedance pneumography is not recommended for research of POA.

2.2.3.3. Strain Gauges

Strain gauges are belts that are placed around the torso to monitor respiratory movements [58]. Changes in volume due to respiration produce a stretching of the belts, which is transduced to an electrical signal. There are three types of strain gauge [58]: (i) wire, (ii) mercury in Silastic, and (iii) piezoelectric crystal. The output of the wire and mercury strain gauges is directly related to the stretching force [58], and according to Adams et al. [68] they can reflect volume amplitudes relatively accurately. The output of the piezoelectric strain gauge, on the other hand, is related to instantaneous changes in strain, and exhibits exponential decays when the strain is kept constant (e.g., during a breath-hold). For this reason, it is recommended that piezoelectric strain gauges be used only for qualitative assessments [58].

2.2.3.4. Respiratory Inductive Plethysmography

An alternative, non-invasive sensor of respiration is the Respiratory Inductive Plethysmograph (RIP) [69]. RIP uses two elastic bands, a.k.a. respibands, placed around the ribcage (RCG) and the abdomen (ABD) to measure cross-sectional changes produced by respiratory movements. These respiratory movements are then transduced into electrical signals, scaled, and summed to

yield a signal proportional to respiratory volume. RIP has been frequently used for monitoring of respiration in infants, especially in research applications [7, 30, 70-72].

Each respiband consists of an elastic cloth with an insulated, sewn-on wire that follows a quasi-sinusoidal pattern [73]. This wire is connected to an oscillator that generates a low-amplitude, sinusoidal voltage at around 300 kHz [59]. Changes in the cross-sectional area enclosed by the respiband produce variations in the self-inductance of the wire, which modulates the frequency of the signal [59]. Thus, during respiration, these frequency variations can be demodulated to yield a signal representing respiratory movements [59]. Experiments have shown that changes in the cross-sectional area of physiologically-shaped phantoms enclosed by a respiband have a linear relationship with the demodulated, output voltage [73].

2.2.3.4.1. Relationship between RIP Measurements and Respiratory Volume

The theory relating displacements in RCG and ABD to respiratory volume is based on the work by Konno and Mead [54]. They proposed that the respiratory volume could be modeled with two degrees of freedom, one corresponding to the volume in RCG, and the other to the volume in ABD. They found that variations in RCG volume were linearly related to antero-posterior displacements of RCG. A similar, linear relationship was found between ABD volume and ABD displacement. These displacements were measured from reference points situated at the level of the nipples for RCG, and the umbilicus for ABD.

Stagg et al. [74] used this relationship to estimate respiratory volume using magnetometers to measure antero-posterior displacements of RCG and ABD. In this study, respiratory volume (V_{TOT}) was equal to the sum of the volumes of RCG (V_{RCG}) and ABD (V_{ABD}). V_{RCG} and V_{ABD} were estimated from the Konno and Mead model as

$$\begin{aligned} V_{RCG} &= aM_{RCG} + c_{RCG} \\ V_{ABD} &= bM_{ABD} + c_{ABD} \end{aligned}, \tag{2.2}$$

where M_{RCG} and M_{ABD} were the magnetometer signals, and a, b, c_{RCG}, and c_{ABD} were linear model parameters. The total volume was determined as the linear combination of the RCG and ABD magnetometer signals as

$$V_{TOT} = aM_{RCG} + bM_{ABD} + c\,. \tag{2.3}$$

The values of a, b, and c were determined by calibrating the signal V_{TOT} against a volume signal obtained by integrating the output of a pneumotachograph [74].

Similarly, Sackner et al. [75] later evaluated the ability of cross-sectional area measurements by RIP to estimate the respiratory volume, and compared it to the volume measured by an spirometer. Thus, respiratory volume was modeled as

$$V_{TOT} = V_{RCG} + V_{ABD} = a \cdot rcg + b \cdot abd \tag{2.4}$$

where rcg and abd were the RIP signals, and a and b the model parameters. Their results showed that this relation could estimate the spirometer volumes accurately.

2.2.3.4.2. Calibration

Several studies sought to determine the best method to estimate the parameters of the model in equation (2.4) [75-78], a process referred to as RIP calibration. Calibration methods often rewrite equation (2.4) as

$$V_{TOT} = m\left(k \cdot rcg + abd\right) \tag{2.5}$$

where the parameter k determines the proportional relationship between rcg and abd, and m scales the calibrated sum to make V_{TOT} match the volume measured by a spirometer or an integrated pneumotachograph signal.

Two methods have been widely used for RIP calibration: (i) isovolume maneuver calibration (IMC), and (ii) qualitative diagnostic calibration (QDC). The isovolume maneuver was first described by Konno and Mead [54], who found that volumes on RCG and ABD are negatively correlated when the airway is occluded. During IMC [76], the subject performs an isovolume maneuver in which they voluntary occlude the airway, and shift volume from RCG to ABD and vice versa. This makes $V_{TOT} = 0$ and the volume changes in RCG equal to those in ABD but with opposite sign [77], yielding

$$k = \frac{-abd}{rcg}. \tag{2.6}$$

Thus, for IMC the RIP signals are recorded and plotted against each other. If the signals are scaled correctly, the plot should be a line with an angle of –45°. If this is not the case, k must be adjusted to achieve this proportion. To complete calibration, the subject then breathes into a spirometer or pneumotachograph that measures the actual volume, and m is determined such that V_{TOT} equals the measured volume.

IMC requires subjects to voluntarily perform an isovolume maneuver; this is not possible in uncooperative subjects like infants. QDC was introduced by Sackner et al. [77] to estimate the calibration parameters without requiring either an isovolume maneuver, or measurements from spirometers or pneumotachographs.

They claimed that a relation similar to equation (2.6) could be established during natural, unobstructed breathing [77]. This was based on the observation that relative volumes of RCG and ABD vary from breath to breath, even when breaths have a similar V_{TOT} [74]. Thus, QDC assumed that if a subject was breathing with a constant V_{TOT}, taking the standard deviation of both sides of equation (2.5) would yield $\sigma(V_{TOT}) = 0$, and k could be approximated as

$$k \approx \frac{-\sigma(abd)}{\sigma(rcg)}, \tag{2.7}$$

where σ is the standard deviation operator. Since in practice V_{TOT} will not be constant, QDC made the additional assumption that an approximately constant V_{TOT} could be obtained by collecting a large number of breaths, and discarding those whose amplitudes were very different from the mean. Sackner et al. also concluded that it was not necessary to determine the value of m in equation (2.5) to monitor apnea, since relative changes in volume or flow would be sufficient [55, 56, 77].

The assumptions of QDC are not well-founded, which translates into poor estimates of V_{TOT} [56, 79]. This is especially true when there are postural changes after calibration [28, 56, 79]. This

severely restricts the use of QDC in studies involving long respiratory records. Indeed, Weese-Mayer et al. [70] reported that inaccurate RIP calibration was responsible for more than 30 % of the obstructive apnea events missed in infants using the Collaborative Home Infant Monitoring Evaluation (CHIME) monitor.

Since neither IMC nor QDC represent a suitable option for calibration of RIP in infants, the focus of recent studies in infants has moved to the use of uncalibrated RIP [4, 55, 56, 80].

2.2.3.4.3. Advantages

RIP has several advantages for the study of infant postoperative respiratory patterns, compared to other sensors of respiration. First, unlike pneumotachographs, thermal sensors, capnographs, and spirometers, RIP does not require any equipment to be attached to the airway, which would interfere with natural breathing and might distort the breathing patterns. Additionally, infants recovering from surgery are continuously handled by nursing staff and parents, and require feeding, so it is not practical to have a sensor blocking the airway for extended periods of time. Second, in contrast to impedance pneumography, RIP is not affected by cardiogenic artifacts that produce missed detections of central apneic events.

2.3. Analysis of Respiratory Patterns

Sensors of respiration generally yield signals following a quasi-sinusoidal pattern representing regular breathing. This pattern is the most common and arises when a person breathes normally. However, a person may display a number of different respiratory events over time, which will be manifested as different patterns in the respiratory signals. For example, a central apnea will produce a “flat” signal pattern, while a sigh will produce an oscillation with larger amplitude and length compared to that of regular breaths. Analysis of “respiratory patterns” comprises the identification of the different patterns in respiratory signals, and the temporal relation among them.

Initial attempts to study infant respiratory patterns and POA relied on nurse’s annotations of apneic events detected by standard clinical respiratory monitors using apnea alarms. More recent studies have used a more formal, manual analysis where expert, trained scorers scroll through

epochs of respiratory data to detect relevant events [8]. This is currently the most accepted analysis method, and is considered the "gold standard". However, this conventional manual scoring (CMS) is very labor intensive, expensive, and suffers from low intra- and inter-scorer repeatability [9, 81]. Consequently there have been a number of attempts to automate the respiratory pattern analysis. These manual and automated approaches are reviewed in the following sections.

2.3.1. Manual Analysis

In 2007 the American Academy of Sleep Medicine (AASM) published guidelines for the manual analysis of respiratory data from both pediatric and adult subjects [27]; these were revised in 2012 [8]. These guidelines were designed to detect sleep apnea and analyze a subject's sleep.

The AASM guidelines describe the sensors that should be used to detect apnea as either: recommended, which should be routinely used; or alternative, which may be used if the recommended sensor fails or produces an unreliable signal. The recommended airflow sensor for detection of apnea is an oronasal thermal sensor. The sum of RCG and ABD RIP signals, calibrated or uncalibrated, is listed as an alternative sensor for detection of apnea (volume estimate). Another option is the time-derivative of this sum (airflow estimate). A sensor of respiratory effort is required to distinguish between central and obstructive apnea. The recommended sensors for respiratory effort include esophageal manometers and RIP from both RCG and ABD. Pulse oximetry is the recommended sensor for blood oxygen saturation (SAT).

Additionally, the guidelines list a number of scoring rules classified as: (i) recommended rules, which should be routinely used; (ii) acceptable rules, alternative to the recommended rules that may be used at the discretion of the investigator; and (iii) optional rules, which do not need to be followed but are available. Separate rules are provided for pediatric and adult subjects [8]. However, the pediatric rules do not distinguish between infants and older children, but are indicated for any child less than 18 years of age.

In pediatric patients, a pattern is scored as apnea if it meets all the following criteria [8]:

(i) The peak respiratory signal excursion drops by ≥ 90 % from the pre-event baseline.

(ii) The duration of the ≥ 90 % drop lasts at least the minimum duration as specified by central, obstructive, or mixed apnea duration criteria.

(iii) The event meets respiratory effort criteria for central, obstructive, or mixed apnea, as described next.

The guidelines classify pediatric apnea according to the following rule [8]:

(i) Central apnea: The event meets the apnea criteria, is associated with no inspiratory effort throughout the entire duration of the event, and meets at least one of the following conditions:
- The event lasts 20 s or more
- The event lasts at least the duration of two breaths during baseline breathing, and is associated with an arousal, or a SAT drop of 3 % or more.
- For infants aged < 1 year, the event lasts at least the duration of two breaths during baseline breathing, and is associated with a decrease in heart rate to less than 50 beats per min for at least 5 s, or less than 60 beats per min for 15 s.

(ii) Obstructive apnea: The event meets the apnea criteria for at least 2 breaths during baseline breathing, and is associated with respiratory effort throughout the entire period of no airflow.

(iii) Mixed apnea: The event meets the apnea criteria for at least the duration of 2 breaths during baseline breathing, and is associated with absent respiratory effort during one portion of the event, and the presence of inspiratory effort in another portion, regardless of which portion comes first.

These conventional manual scoring (CMS) guidelines have helped to standardize the analysis of infant respiratory patterns. Use of these guidelines became widespread by the development of commercial software that facilitated their application to infant cardiorespiratory data. Some examples of these software are: Crystal PSG by Cleveland Medical Devices Inc.; Embla Sandman Elite, REMbrandt PSG, and Embla RemLogic by Natus Medical Inc.

However, CMS analysis is limited by 5 important factors:

(i) Only expert, trained scorers may perform the analysis, which limits the availability of scorers and increases the costs.

(ii) Even when the guidelines are applied by expert scorers, results have very low intra- and inter-scorer repeatability [9]. This variability makes it difficult to compare the manual analyses from different scorers, or from different patients. This limits the statistical power of any test, since high variability in the analysis method would obscure important findings. Moreover, if the objective was to minimize the effect of scorer variability, only one scorer should perform the analysis. However, this approach is inadequate because the results would be biased towards the subjective judgments of a single scorer, making any findings less universal.

(iii) Results are not available in real-time nor close to it, since the analysis is very time consuming and labor intensive.

(iv) Rules only define "clinically relevant" events (e.g., apnea), but do not require the comprehensive scoring of the complete, continuous respiratory patterns, consequently large sections of the respiratory data are not scored. This is only useful to describe the occurrence of events, but it does not allow establishing any relation between different respiratory patterns and the occurrence of apneas. Thus, it is not possible to estimate the risk of future apnea based on previously observed respiratory patterns.

(v) Analyses based on the AASM guidelines generally report results as an index of disordered breathing known as the apnea index (AI), defined as the number of apneas per hour. Studies may also report a similar apnea-hypopnea index (AHI), where hypopneas are defined as partial decreases in ventilation. Both AI and AHI represent a summary of relevant respiratory events; however, these indices ignore all the relationships, and time correlations between different respiratory patterns. This means that an analysis using AASM guidelines reporting AI or AHI cannot determine if a particular respiratory pattern, or a sequence of patterns, is predictive of apnea. Sighs [7], and short respiratory pauses [82] are examples of respiratory patterns not defined by the AASM guidelines that have been linked to apnea in infants. Thus, at the end of all the effort required to apply the AASM guidelines, the result is only a log of respiratory events rather than a sample-by-sample, continuous sequence of respiratory

patterns that can be used to determine the temporal correlations between respiratory patterns and apneas using signals and systems analysis.

Investigators have aimed to address some of the limitations of manual scoring by automating the analysis of respiratory patterns. The following section reviews these efforts.

2.3.2. Engineering and Machine Learning Applied to Respiratory Pattern Analysis

Automated methods for analysis of respiratory patterns have focused on 3 aspects: (i) apnea detection; (ii) estimation of thoraco-abdominal asynchrony and its relationship to obstructive apnea; and (iii) detection of artifacts in respiratory signals. These methods use detectors and/or classifiers that evaluate a portion of data and decide whether a pattern is present or not. This section provides a brief description of how detectors and classifiers are evaluated, and then uses this to review and compare different methods developed for the analysis of respiratory patterns.

2.3.2.1. Evaluation of Detectors and Classifiers

Before reviewing automated methods for analysis of respiratory patterns, it is important to understand how to assess the performance of detectors and classifiers. The standard approach to characterize a detector is in terms of its probabilities of detection P_D and false alarm P_{FA} [83]. P_D, also known as sensitivity, describes the probability of marking a candidate event as apnea when it is in fact a true apnea (i.e., true positive). Conversely, P_{FA} indicates the probability of marking a candidate event as apnea when it is not an apnea (i.e., false positive). Specificity, another term commonly used, is equal to $1 - P_{FA}$. Thus, the performance of a detector or classifier can be summarized as a single pair of P_D and P_{FA}, with the ideal detector having values of $P_D = 1$ and $P_{FA} = 0$.

In many instances it is possible to vary the values of a detector's parameters to obtain various combinations of P_D and P_{FA}. The relation between P_D and P_{FA} as a function of the parameter defines the receiver operating characteristics (ROC) curve. Fig. 2.1 shows an example of an ROC curve. The area under the ROC curve (AUC) is a measure of performance [84].

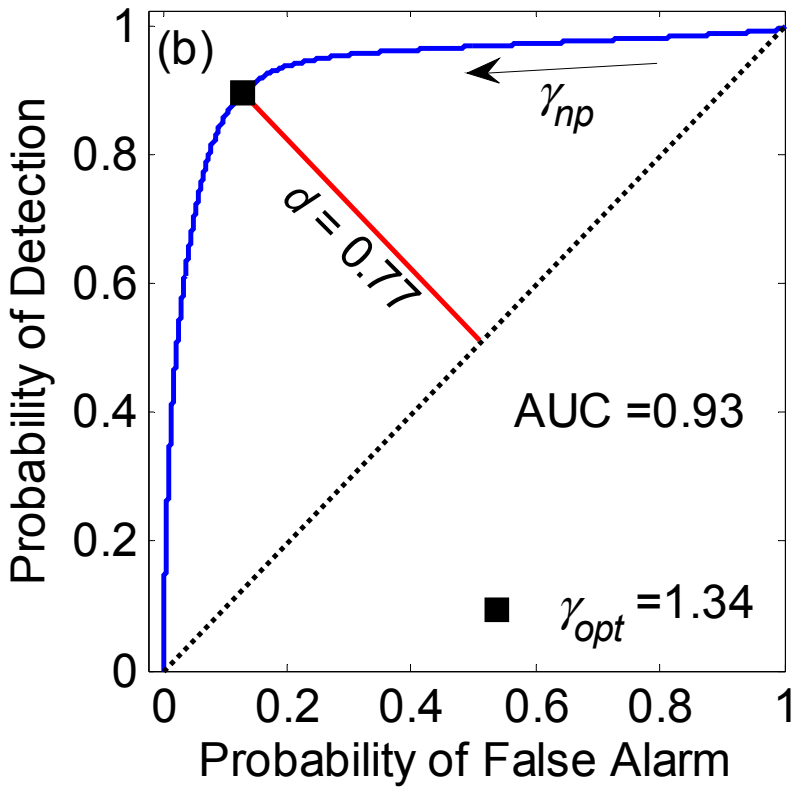

Fig. 2.1. Example of a Receiver Operating Characteristics (ROC) curve (blue). The diagonal, black, dotted line corresponds to the performance expected by chance. The maximum *d*-value indicates the best tradeoff between probabilities of detection and false alarm, and occurs at the point farthest from the chance line. AUC = Area Under the ROC curve.

An $AUC = 1$ indicates perfect detection/classification and $AUC = 0.5$ corresponds to chance performance.

The ROC curve defines the detector/classifier performance as a function of the parameter and so can be used to select its optimal value. One approach is to select the point on the ROC curve that is furthest from the chance line ($P_D = P_{FA}$) [85]. The relation

$$d = \frac{(P_D - P_{FA})}{\sqrt{2}} \Big/ \frac{1}{\sqrt{2}} = P_D - P_{FA}, \tag{2.8}$$

defines the distance of any point on the ROC curve to the chance line scaled to the range 0 to 1. Higher *d*-values indicate a better combination of P_D and P_{FA}, and therefore represent a better overall performance.

Automated methods are generally compared to a reference, manual analysis performed by expert scorers to estimate the P_D, P_{FA}, ROC curve, and *d*-value. However, there are three different approaches to compare the analyses produced by automated methods to the reference:

(i) The first is event-by-event, where events of interest (e.g., apneas) are defined as segmented in the reference, and each event is deemed detected if at least part of it was identified by the automated method.

(ii) The second is epoch-by-epoch, where the reference is used to classify epochs (i.e., fixed-length segments of continuous data) as positive if they contain the event, or negative otherwise. Note that epochs are assigned to a positive event regardless of event length. Then, a positive epoch is considered detected if at least part of it is detected as positive by the automated detector.

(iii) The third is sample-by-sample, which takes into account that each reference event consists of several data samples. This approach compares all samples produced by the automated analysis to the corresponding samples in the reference. Thus, if the automated detector only detects a portion of an event, only those samples are considered detected and the rest are counted as missed detections.

Each approach has implications on the final results. The sample-by-sample approach is the strictest, since all samples are comprehensively evaluated. In fact, by analyzing every sample the result gives a combined assessment of two properties; the ability to: (i) detect the event of interest, and (ii) determine the event start and end times. The event-by-event strategy is more lenient because it only requires a portion of the event be detected, but by doing this it does not evaluate the start and end times. Moreover, the event-by-event strategy produces results that overestimate P_D, since some of the samples need not be correctly detected to be considered correct. The epoch-by-epoch strategy suffers from the same limitations as event-by-event, but in a larger scale. This is because although one epoch may contain different types of events, it is labeled as only one type. Then, if a part of the epoch is detected by the automated method, it is considered as correct. This will overestimate P_D and misclassify significant portions of data when the events are shorter than the epoch length (usually 30 s) or span more than one epoch.

Based on this analysis, we believe that whenever possible analyses produced by automated methods should be evaluated sample-by-sample, so that results represent more accurately the actual performance and not just an approximation. This is especially important for methods trying to yield a detailed analysis representing the respiratory patterns as a function of time.

2.3.2.2. Apnea Detection

Accurate detection of apneic events is fundamental for the study of respiratory patterns. Due to the limitations of conventional manual scoring (CMS), several automated apnea detectors have been developed. Table 2.1 summarizes the different apnea detectors that will be reviewed in this section.

During a central apnea, the respiratory effort is close to zero and so the respiratory airflow and volume signals are “flat”. Based on this, detectors have been designed based on the amplitude property of respiratory signals. Macey et al. [86] used the standard deviation of abdominal respiratory signals as a metric of “flatness”, and apnea was detected when the standard deviation was less than a threshold. Similarly, Lee et al. [87] used the standard deviation of impedance pneumography signals to estimate the probability of apnea, and compared the estimated probability to a threshold to determine the presence of apnea. These methods showed very high

Method	Signals	Metrics	Strategy	P_D	P_{FA}	d	Evaluation Unit
Macey et al. [86]	ABD	Standard deviation	Threshold	0.99	0.29	0.70	–
Lee et al. [87]	IP	Standard deviation	Threshold	0.97	≥0.37	≤0.60	Event
Macey et al. [89]	ABD	"Flatness", duration, thinness, and smoothness	ANN	0.94	0.41	0.53	–
Álvarez-Estévez and Moret-Bonillo [90]	Airflow, RCG, ABD, and SAT	–	Fuzzy logic	0.87	0.11	0.76	Epoch
Han et al. [91]	NAF	Mean absolute value of second difference	Threshold	0.92	0.12	0.80	Event
Van Houdt et al. [92]	NAF	Duration, amplitude, and slope of half-breaths	Threshold	0.89	–	–	Event
De Groote et al. [80]	RCG and ABD	Raw signals from strain gauges	ANN	0.75	0.94	-0.19	Epoch
Varady et al. [93]	NAF	Instantaneous respiration amplitude, instantaneous respiration interval	ANN	0.89	0.04	0.85	Epoch
De Chazal et al. [94]	ECG	52 features from inter-beat interval, 36 from EDR signal	LDA, QDA	0.90	0.06	0.84	Epoch
Khandoker et al. [95]	ECG	Mean, variance, skewness, kurtosis, and Shannon's entropy from each set of wavelet coefficients	Symlet Wavelets, ANN	0.92	0.01	0.91	Epoch
Xie and Minn [96]	ECG and SAT	111 features (see [88] for details)	Multiple classifiers	0.79	0.15	0.64	Epoch

Table 2.1. Comparison of automated methods for detection of apnea. ABD = Abdominal respiratory movements, ANN = Artificial Neural Network, ECG = Electrocardiogram, EDR = ECG derived respiration, IP = impedance pneumography, LDA = linear discriminant analysis, NAF = nasal airflow, P_D = Probability of detection, P_{FA} = Probability of false alarm, QDA = quadratic discriminant analysis, RCG = ribcage respiratory movements, SAT = blood oxygen saturation. A '–' indicates data was not described.

P_D, but suffered from increased values of P_{FA} (see Table 2.1). Thus they detected most apneas at the cost of having many false alarms. This would require an additional step to distinguish between true apneas and false alarms to enable a realistic analysis of the incidence of apnea.

This strategy was followed by Macey et al. [89] in an effort to reduce P_{FA}. The apnea detector from [86] was coupled with a two-hidden-layer, artificial neural network (ANN) to classify segments as apnea or non-apnea. The rationale was that a very sensitive detector would detect candidate apneas, and then the features of these candidate apneas would be supplied to an ANN to refine the classification. A variety of features [89, 97, 98] related to the "flatness", duration, thinness, and smoothness of the event were tried with nonsuccess; the resulting detector had worse P_{FA} than the original detector (see Table 2.1).

A similar, two-step approach was followed by Álvarez-Estévez and Moret-Bonillo [90], who used a detector of candidate events to feed information to a fuzzy logic decision system. Their method classified 30 s epochs as apneic or non-apneic, but had only a slightly higher *d*-value than the simpler detector based on standard deviation from Macey et al. [86] (see Table 2.1).

It is likely that the high P_{FA} values from [86, 87] arose from inappropriate thresholds, since they were selected arbitrarily rather than based on any objective criteria. In contrast, Han et al. [91] selected detector thresholds based on data manually analyzed by expert scorers. This likely helped to obtain a low P_{FA} while maintaining a high P_D (see Table 2.1). The metric used was the mean of the absolute value of the second difference of the nasal airflow signal. Even though this method performed better than previous methods [86, 87], its implementation was limited by the need to estimate the second difference, a process that will amplify noise especially when signal-to-noise ratio is low.

Van Houdt et al. [92] used a different approach; the nasal airflow signal was segmented on a half-breath basis, and characterized in terms of the half-breath's: (i) duration, (ii) amplitude, and (iii) slope. Apnea was detected by comparing these metrics to thresholds. However, this method had two important limitations: (i) thresholds were determined arbitrarily, and (ii) it required estimation of the first derivative for breath segmentation, which is not robust to noise.

De Groote et al. used an ANN classifier with no pre-detection step to detect obstructive apnea in full-term infants [80]. The inputs to the ANN were 15 s segments of RCG and ABD respiratory movements measured with piezoelectric strain gauges. The ANN performed poorly (see Table 2.1); however, many of the false alarms occurred during central apneas or in segments corrupted by movement artifact. This suggests that performance of the ANN could have been significantly improved by adding a pre-processing step to systematically detect and exclude segments with central apnea or movement artifact.

Varady et al. also used an ANN classifier to detect apnea [93], but used the instantaneous respiration amplitude and instantaneous respiration interval as inputs to the ANN. This significantly improved the performance (see Table 2.1), suggesting that it is more effective to first extract representative metrics from the raw signals, and then use these metrics to classify or detect events.

Other studies attempted to detect apnea using cardiac signals such as the electrocardiogram (ECG), the photoplethysmogram (PPG), or SAT [94, 99-104]. The rationale for this is that the heart rate is modulated by respiratory sinus arrhythmia, a cardiorespiratory synchronization that shortens the inter-beat interval during inspiration, and prolongs it during expiration [105, 106]. Additionally, during apnea, hemoglobin oxygen saturation may drop, especially during prolonged events [107]. These studies followed three main approaches: (i) to extract a respiratory signal from a recorded, cardiac signal [108, 109], and then detect apnea from this signal [88, 94, 96]; (ii) to estimate a signal representing the inter-beat intervals, then extract features from this signal, and use these features with a classifier to detect apnea [88, 94, 96]; and (iii) to extract features directly from the raw signals (e.g., wavelet decomposition, number of times SAT drops below a given threshold) and input these to a classifier to detect apnea [95]. Table 2.1 shows that these methods gave performances similar to those of methods based on respiratory signals. However, while cardiac signals can detect apnea, they are not useful when trying to study multiple types of respiratory patterns such as sighs, and thoraco-abdominal asynchrony.

All these methods were evaluated event-by-event or epoch-by-epoch, strategies that overestimate P_D and do not take into account the start and end times of events, and consequently neither the

event length. This is a problem especially for apnea, since the distinction between a short pause and an apnea depends on the event length.

In summary, methods using sophisticated machine learning, fuzzy logic, or features from cardiac signals performed little better than the simpler detectors based on thresholds of standard deviation or absolute value of respiratory signals. We believe that an optimization strategy for threshold selection might yield similar or better detection than the use of specialized classifiers, since this would improve the *d*-value by maintaining a high P_D while reducing P_{FA}, as occurred for the method by Han et al. [91].

2.3.2.3. Thoraco-abdominal Asynchrony Estimation

Thoraco-abdominal asynchrony (TAA), a respiratory state where RCG and ABD movements are out of phase, is an important indicator of obstructive apnea [110]. TAA also occurs in infants during rapid eye movement (REM) sleep [111] and during anesthesia [112-115]. Thus, appropriate detection of TAA is important for management of respiratory disease [71] and infant respiratory monitoring in general.

Prisk et al. [116] evaluated different methods to measure the degree of TAA (ϕ) using simulated signals resembling RCG and ABD movements. They evaluated 3 waveform-independent methods: maximum linear correlation (i.e., Pearson correlation [4]), paradoxical motion, and cross-correlation; and 3 waveform-dependent methods: Lissajous plot analysis, signal averaging, and linear modeling.

Maximum linear correlation estimates a linear fit of RCG as a function of ABD, and obtains the correlation coefficient. Then, one signal is shifted and a new correlation coefficient is obtained. The shift at which the correlation coefficient is maximal is normalized by the signal average period to determine the degree of TAA. Paradoxical motion estimates the degree of TAA as the proportion of time the two signals are moving in opposite directions. The direction of motion of each signal is determined by estimating their time derivatives, a process that is vulnerable to high frequency noise. The cross-correlation method estimates the cross-correlation function between

two signal segments which will have its maximum at the delay between the two signals. The degree of TAA is estimated by dividing this lag by the average period of the signal segments.

The Lissajous plot analysis employs a parametric representation of respiration using the RCG and ABD signals. ABD is plotted on the x-axis, and RCG is plotted on the y-axis. If both RCG and ABD were sinusoidal, the resulting curve would be an ellipse (i.e., a "loop") [117]. The ellipse will be tilted to the right if breathing is synchronous, and tilted to the left if breathing is asynchronous [117]. To estimate the degree of TAA it is necessary to measure two horizontal distances: (i) the width of the ellipse at the midpoint on the y-axis (w_{MP}), and (ii) the total width of the curve from minimum to maximum x-axis values (w_T). The degree of TAA can be estimated from the relationship $\sin\phi = w_{MP}/w_T$, combined with the sign of the slope of the major axis of the ellipse [116, 117].

Signal averaging computes two new signals from the original RIP signals: (i) the sum of ABD plus RCG (SUM), and (ii) the difference of ABD minus RCG (DIF). If the signals are sinusoidal, the rectified average (i.e., the average of the absolute value) of both SUM or DIF would be constant for a sufficiently long time. Based on this, it is possible to establish the following set of two equations with two variables [116]:

$$\begin{aligned} \overline{|SUM|} &= C \cdot \cos(\phi/2) \\ \overline{|DIF|} &= -C \cdot \sin(\phi/2) \end{aligned}, \tag{2.9}$$

where $\overline{|SUM|}$ and $\overline{|DIF|}$ are the rectified averages of SUM and DIF respectively, and C is a constant. The TAA degree is obtained by solving the equations.

The linear modeling method hypothesizes that RCG is a linear function of ABD and its derivative with respect to time, i.e., $rcg = k_1 \cdot abd + k_2 \cdot d(abd)/dt$. If the signals are assumed to be sinusoidal, the degree of TAA will be given by $\tan(\phi) = k_2/k_1$ [116].

According to [116], maximum linear correlation performed the best with an error of less than 1^o for both sinusoidal and triangular signals and in noise-free and noisy conditions. It was closely

followed by cross-correlation, which had errors of < 3°. The remaining techniques had larger errors, especially with triangular signals and in the presence of additive noise. The main limitation in this study was that the simulated RCG and ABD signals were modeled as sinusoidal or triangular which is not realistic. Furthermore, the effects of noise were evaluated with white Gaussian noise, which is not representative of low-frequency noise found in RIP signals [4].

Brown et al. proposed a method to estimate TAA based on the Lissajous plot of RCG against ABD [71]. The parameters of the function $rcg = \alpha \cdot abd + \beta$ were estimated on 10 s segments using recursive linear regression. Each segment was classified into one of two categories: (i) synchronous-breathing (SYB), when $\alpha > 0$ for > 90 % of the time; and (ii) asynchronous-breathing (ASB), when $\alpha > 0$ for < 10 % of the time. Segments classified as ASB had significantly higher TAA (estimated with the cross-correlation method described above [116]). The main limitation of the method is that it does not provide a quantitative estimate of the degree of TAA, but only a binary classification. Moreover, segments with > 10 % and < 90 % of the time with $\alpha > 0$ were excluded from the evaluation and not classified. These segments represent threshold data that are more challenging to classify, and it is unclear how the method would behave under this circumstances.

De Groote et al. presented the mirror index as another estimator of TAA [80]. Each ABD breath was divided by its amplitude, and a similar operation was performed on RCG, and the two normalized breaths were summed to yield a SUM breath. The mirror index was defined as the area enclosed by the SUM breath, divided by the breath length. During ASB, the mirror index should approach a value of 0, while during SYB it should by higher. This index was only evaluated as a detector of obstructive apnea in infants and not as an estimator of ϕ. It had a poor performance with P_D of 0.79 and a P_{FA} of 0.89. Further analysis is necessary to determine if this mirror index is an accurate estimator of ϕ.

The most reliable method to estimate ϕ in infants to date was introduced by Motto et al. [4]. It is based on a principle similar to that of the paradoxical motion method (i.e., samples where the signals move in opposite directions have a higher TAA, while samples moving in the same

direction have a lower TAA). To estimate ϕ, the RCG and ABD signals were converted into binary signals, and the exclusive-OR (XOR) was estimated, sample-by-sample, between the two binary signals. The degree of TAA was estimated as the average value of the XOR over a sliding window. If breathing was asynchronous in the window, the output would be close to 1; conversely, if it was synchronous then the output would be close to 0. The method was compared to maximum linear correlation, which according to Prisk et al. [116] was the best option to estimate ϕ. This analysis showed that the XOR method had less bias and variance probably because it reduced noise by converting signals into binary signals and averaging, while the paradoxical motion method amplified noise by differentiating.

2.3.2.4. Movement Artifact Detection

It is essential to distinguish clean data segments from those corrupted with non-respiratory movement artifacts (MVT) for appropriate diagnosis of clinically significant abnormalities [5]. MVT in RIP occurs when the subject moves or is moved and the resulting non-respiratory movements are measured in RCG and/or ABD. When RIP is used to detect obstructive apnea by estimating TAA, MVT will generate a significant number of false alarms [80]. Thus, accurate detection of obstructive apnea using RIP requires the detection and exclusion of MVT.

Motto et al. [3] designed a MVT detector based on the hypothesis that RIP would have larger amplitudes during MVT than normal breathing. RIP amplitude was estimated in terms of its root-mean-square (RMS) value. The detector was tested on infant data, where it distinguished well between MVT and normal breathing (see Table 2.2). However, the problem of how to select the optimal detection threshold was not addressed.

Later, Aoude et al. [5] presented a RIP MVT detector specific for infant data. The detector was based on 3 observations: (i) the fundamental frequency of normal, infant breathing is band limited to 0.4 Hz to 2.0 Hz [4]; (ii) MVT usually has larger amplitude than regular breathing [3]; and (iii) MVT occurs predominantly at lower frequencies (i.e., from 0 Hz to 0.4 Hz). A metric was designed to quantify the relative power between the breathing and MVT frequency bands, and MVT was detected by comparing this metric to a threshold. This method performed better

Method	Signals	Metrics	Strategy	P_D	P_{FA}	d	Evaluation Unit
Motto et al. [3]	RCG and ABD	Root-mean-square	Threshold	>0.80	0.20	>0.60	Sample
Aoude et al. [5]	RCG and ABD	Relative power between infant breathing and MVT frequency bands	Threshold	0.86	0.14	0.72	Sample
Van Houdt et al. [92]	NAF, RCG, or ABD	Slope of half-breaths	Threshold	0.70	0.01	0.69	Event

Table 2.2. Comparison of automated methods for detection of movement artifact (MVT) in respiratory inductive plethysmography (RIP) signals. ABD = abdominal RIP signal, NAF = nasal airflow, P_D = probability of detection, P_{FA} = probability of false alarm, RCG = ribcage RIP signal.

than the one from Motto et al. [3] (see Table 2.2), but it also failed to address the optimal selection of the threshold.

Van Houdt et al. [92] developed an approach based on a breath-by-breath, period-amplitude analysis of nasal airflow or RIP signals to identify MVT. The signals were segmented in half-breaths to identify the times at which inspiration and expiration started for each breath. The slope of each half-breath was estimated as the ratio of its amplitude to its duration, and was compared to a threshold to detect MVT. The underlying hypothesis was that half-breaths corrupted by MVT would have larger slopes compared to regular half-breaths. Similar to the previous two methods [3, 5], the threshold was selected based on arbitrary criteria. This method had a slightly lower performance (see Table 2.2) than the method by Aoude et al. [5], but higher than [3]. However, the method was evaluated event-by-event, contrary to [5] and [3] which were evaluated sample-by-sample. It is likely that performance of this method would decrease if evaluated sample-by-sample.

In conclusion, the method by Aoude et al. [5] provides the best detection of MVT. The method was evaluated using the strict sample-by-sample approach, and had the highest *d*-value, indicating the best combination of P_D and P_{FA}. The three methods detected MVT by comparing metrics to thresholds, but thresholds were set based on arbitrary criteria.

2.4. Existing Infant Cardiorespiratory Datasets

Access to representative clinical data is a prerequisite for the development of tools for analysis of infant respiratory patterns and the study of POA. These data must be acquired from infants recovering from surgery and anesthesia in the immediate postoperative period, and consist on cardiorespiratory signals (e.g., RIP and SAT) recorded continuously for at least the first 12 h postoperatively [8, 14, 17, 19, 39, 52, 54-56]. There are no such data available to date.

A dataset with similar properties is the one from the Collaborative Home Infant Monitoring Evaluation (CHIME) study [118], which is a collection of overnight, cardiorespiratory signals from more than 1,000 infants. However, it is impossible to use the CHIME dataset to study the respiratory patterns and POA due to two important limitations. First, data were not acquired from

infants recovering from surgery and anesthesia, but rather from infants sleeping at home to assess their risk of SIDS. Therefore, the respiratory patterns observed in these data would not be representative of infants at risk of POA. Second, CHIME data were not recorded continuously throughout each session. Instead, only brief segments were recorded starting a few seconds before, and ending a few seconds after automatically-detected periods of slow heart rate or apnea. This recording protocol omitted a significant portion of each session, making it impossible to identify the temporal relationships between different respiratory patterns.

2.5. Summary and Thesis Rationale

Infants with postmenstrual age (PMA) ≤ 60 weeks who are recovering from surgery and anesthesia are at risk of life threatening POA [14, 17, 37, 39, 41, 45, 46]. The first POA episode may occur within the first 12 h postoperatively [14], and episodes may occur up to 72 h after surgery [39]. Thus, it is a generally accepted guideline to monitor infants at risk in hospital for at least the first 12 h after surgery, and if they experience POA in that period then continue monitoring up to 72 h.

While PMA [14, 17, 37, 39, 41, 45, 46], prematurity [15, 37, 45], and anemia [41] represent important risk factors for POA, it is not possible to predict which infants will develop POA based only on these variables. However, studies have found that the postoperative respiratory patterns of infants with POA are different from those of infants who do not exhibit POA [14, 19, 20, 24, 39]. This suggests that a comprehensive analysis of postoperative breathing patterns may provide insight about POA. We believe that this analysis should describe all respiratory patterns that an infant may display while recovering from surgery and anesthesia (e.g., regular breathing, respiratory pause, thoraco-abdominal asynchrony, sigh), while also describing the properties of these patterns (i.e., type, time of occurrence, and length). Knowledge of these properties would enable the study of the postoperative respiratory patterns and their relation to POA, and could help to identify specific infants at risk of POA and the time at which this risk no longer exists.

To enable this analysis it is important to have a representative clinical dataset from infants at risk of POA, and tools to analyze the respiratory patterns from such dataset. None of these are currently available publicly.

The only available pediatric cardiorespiratory dataset is the Collaborative Home Infant Monitoring Evaluation (CHIME) study, which deals with infants at risk of Sudden Infant Death Syndrome while sleeping at home, and not with infants recovering from surgery. Data from previous POA studies are not available, and monitoring technologies and recording periods in these studies were not standardized [17]. Therefore, there is need for a standardized dataset representative of infants at risk of POA. Based on the times of occurrence of POA [14, 17, 39], this data should comprise continuous recordings of cardiorespiratory signals starting immediately after surgery, and lasting for at least 12 h postoperatively. A significant proportion (≈ 30 %) of POA episodes may contain an obstructive component [19], which can lead to significantly larger drops in blood oxygen saturation (SAT) than central apneas [19]. Therefore infants must be monitored with a pulse oximeter to measure SAT, and a respiration sensor capable of detecting episodes of airway obstruction. We believe that the Respiratory Inductive Plethysmograph (RIP) is the best means of monitoring respiration because it is non-invasive, and can be used to detect apneas and distinguish between central and obstructive components [8, 55, 56]. We consider that other sensors of respiration such as pneumotachographs, thermistors, thermocouples, capnographs, and spirometers are much less appropriate because they must be attached to the infant's airway, which alters the natural respiratory patterns and interferes with feeding and handling.

With respect to data analysis, there is currently no available method to comprehensively analyze the respiratory patterns. Conventional manual scoring (CMS) is the most accepted method to date. It is based on several guidelines published by the American Academy of Sleep Medicine (AASM) [8, 27]. However, CMS is labor intensive, expensive, and suffers from low intra- and inter-scorer repeatability [9, 81]. Moreover, CMS only describes the occurrence of "clinically relevant" events, but does not score any other respiratory patterns such as short pauses, sighs, or thoraco-abdominal asynchrony (TAA). Thus, it is not possible to establish possible relations between these excluded patterns and POA. This is a significant limitation since sighs [7] and short pauses [82] have been linked to apnea in infants.

There have been efforts to automate the analysis of infant respiratory patterns, and methods have been designed to analyze apnea [80, 86-96, 99-104], TAA [71, 80, 116], and movement artifact

[3, 5, 92] in respiratory signals. However, these efforts represent only the initial steps towards a comprehensive analysis. We identified three important missing aspects necessary to bring the analysis to the level required for the study of POA. First, the analysis method should detect all possible respiratory patterns from infants in the recovery room, so that their relationship to POA can be studied. Current methods fail to do this, since they focus on a single pattern of interest. Second, detector thresholds should be determined based on optimization criteria, so that performance is maximized. Current methods did not address this challenge, but rather selected threshold values arbitrarily [80, 86-90, 92-96, 99-104]. Third, most automated methods available in the literature have been evaluated using an event-by-event, or epoch-by-epoch approach. As described above, these approaches tend to overestimate performance, and do not take into account the start and end times of events. Thus, we believe that automated methods should be evaluated using the sample-by-sample approach to provide a comprehensive assessment of performance.

The objective of this thesis was to address these needs by acquiring a representative dataset from infants at risk of POA, and developing automated methods to comprehensively analyze the respiratory patterns in these data accurately, consistently, quickly, and at a low cost.

3. Representative Infant Data

It is mandatory to have access to representative clinical data to develop tools for analysis of infant postoperative respiratory patterns. There are no such data available to date. A dataset with similar properties is that from the Collaborative Home Infant Monitoring Evaluation (CHIME) study [118]. The CHIME dataset is a collection of overnight, cardiorespiratory signals from more than 1,000 infants. However, this dataset is not appropriate for the development of tools for analysis of respiratory patterns because cardiorespiratory signals were not recorded continuously; only short segments related to periods of slow heart rate or apnea were stored. This recording protocol does not allow identifying the temporal relationships between different respiratory patterns.

Therefore, we set out to acquire a representative, clinical dataset from infants to enable the study of postoperative respiratory patterns. We recruited infants at risk of postoperative apnea (POA) who had received general anesthesia, and continuously recorded their respiratory patterns in the recovery room for up to 12 h, anticipating that these data would include breathing, respiratory pauses and movement artifact. This Chapter describes the details of the data acquisition and the dataset.

3.1. Study Design

Data were acquired from infants at risk of POA in the Postanesthesia Care Unit (PACU) of the Montreal Children's Hospital (MCH). Potential recruits were identified from the Operating Room booking office. Inclusion criteria were: (i) postmenstrual age (PMA) < 60 weeks at the time of surgery in preterm and former preterm infants, and < 48 weeks in full-term infants; (ii) elective surgery for inguinal herniorrhaphy; and (iii) American Society of Anesthesiology physical status 1 or 2. Exclusion Criteria were: (i) post-operative admission to the Neonatal Intensive Care Unit or Pediatric Intensive Care Unit; (ii) emergency surgery; and (iii) spinal anesthesia.

The study was approved by the Institutional Review Board (IRB) of McGill University Health Centre/MCH (approval number PED-07-30), and informed written parental consent was obtained for each infant. A second study was approved by the IRB of MCH (12-308-PED) allowing the manual analysis of the data by 3 additional manual scorers.

Since the purpose was to create a library of representative, manually-scored respiratory patterns from infants at risk of POA, and not to test a specific hypothesis, our intention was to recruit a convenience sample of infants instead of fixing a sample size.

3.2. Data Acquisition Setup

Upon arrival at the PACU, infant respibands (Ambulatory Monitoring Inc., Inductobands, Ardsley, NY, USA) were placed around the ribcage, at the nipple line, and abdomen, at the umbilicus, and interfaced with a respiratory inductive plethysmograph (RIP, Ambulatory Monitoring Inc., Battery Operated Inductotrace, Ardsley, NY, USA), to measure respiratory movements. An infant oximeter probe (Nonin 8600 Portable Digital Pulse Oximeter, Plymouth, MN, USA) was taped to a digit to measure blood oxygen saturation (SAT) and the photoplethysmography (PPG) signal. The outputs were low-pass filtered (cut-off frequency 10 Hz) with an 8-pole Bessel anti-aliasing filter (Kemo, Jacksonville, FL, USA) digitized (16 bit resolution) and sampled at 50 Hz and recorded on a computer using MATLAB (The MathWorks Inc., Natick, MA, USA) for off-line analysis. This data acquisition system was described in [2]. The system is battery operated to avoid interference with clinical equipment, and has an operating time of 12 h to 15 h [2]. No attempt was made to calibrate the RIP signals since recent manual scoring guidelines list uncalibrated RIP as a recommended sensor of respiratory effort and an alternative sensor to detect apnea [8]. We recorded data until infants were released from the recovery room (up to 12 h as permitted by the acquisition system) in accordance with the MCH practice guidelines for apnea monitoring in full-term and former preterm infants.

We continuously attended the data acquisition sessions and annotated a paper record of behavioral state of the infant (sleeping, feeding, diaper change, etc.), referenced to the clock time and recording time. We then converted these handwritten entries to an electronic entry by

scanning and manually entering the data. We also recorded demographic data as well as potential confounding clinical variables: perioperative anesthetic drug administration, analgesic, and postoperative medications.

3.3. Results

We recruited a total of 24 infants to the study. Table 3.1 shows a summary of variables recorded during data acquisition. Three recordings had to be excluded from further analysis; one recording was corrupted due to continuous handling of the infant by nurses and the parents, and the remaining two had bad data quality due to problems with the data acquisition system battery. In fact, halfway through the study we had to replace the battery since it had run out after several years of use.

Drug regimens differed among infants because anesthetic management was not standardized. At the induction of anesthesia, all infants received atropine, and 21 received propofol. One infant received a dose of propofol at the end of surgery. One infant received a second dose of atropine at the time of extubation. The maintenance anesthetic agent was either sevoflurane (n = 14) or desflurane (n = 10). An opioid was administered to 16 infants (fentanyl = 10, sufentanil = 2, remifentanil = 4). Acetaminophen was administered to 21 infants. Rocuronium was administered to 12 infants, and the muscle relaxant was antagonized with neostigmine and atropine.

Subjects received different anesthetics and perioperative drugs, including opioids; this should help to eliminate, or at least reduce, any bias induced by a specific drug, making the data library more general.

3.4. Public Availability

The complete dataset from infants at risk of POA described above has been made fully available without restriction in the Dryad Digital Repository (doi:10.5061/dryad.72dk5) [10, 11]. To our knowledge this is the only such dataset.

Variable	Value
Gender (% of males)	79
Postmenstrual Age (weeks)	43 ± 2
Birth Age (weeks)	31 ± 4
Weight (kg)	3.7 ± 1.0
Duration of surgery and anesthesia (min)	99 ± 27
Recording Time (h)	9.0 ± 2.2

Table 3.1. Summary of variables from data acquisition sessions.

Consent for publication of raw data was not requested specifically at the time the study was carried out. However, we thoroughly inspected all materials, and removed all possible identifiers (as defined in [13]) before making the data available publicly. Thus, we believe that publication of these data poses negligible risk to the privacy of study participants [10].

4. Scoring Tools for the Analysis of Infant Respiratory Inductive Plethysmography Signals

4.1. Preface

Manual scoring is considered the "gold standard" method to analyze infant respiratory data. However, conventional manual scoring (CMS) has low intra- and inter-scorer repeatability, and does not comprehensively describe the respiratory patterns, but only short segments regarded as "clinically relevant". This has limited the study of postoperative apnea (POA) and its relation to abnormal postoperative respiratory patterns. While the objective of this thesis was to develop automated methods for analysis of respiratory patterns, it was still necessary to develop an improved, high quality, manual scoring method to be used as a "gold standard" reference for evaluation of the automated methods.

In this Chapter I describe the development of a set of tools for manual scoring of infant respiratory data, designed to address the limitations of CMS. These tools support the scoring of all samples in long data records, provide a fully automated training protocol, improve intra- and inter-scorer repeatability, and incorporate ongoing quality control procedures to maintain this repeatability throughout a study. This Chapter also presents the results of a validation study demonstrating that respiratory data can be efficiently analyzed with high intra- and inter-scorer repeatability using these tools.

C. A. Robles-Rubio, G. Bertolizio, K. A. Brown, and R. E. Kearney, "Scoring Tools for the Analysis of Infant Respiratory Inductive Plethysmography Signals," *PLoS ONE*, vol. 10, p. e0134182, 2015. Digital Object Identifier: 10.1371/journal.pone.0134182.

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (www.nserc-crsng.gc.ca, grant NSERC RGPIN 1051-13), and in part by the Queen Elizabeth Hospital of Montreal Foundation Chair in Pediatric Anesthesia, McGill University Faculty of Medicine (www.mcgill.ca/medicine/faculty-medicine). CARR was supported in part by a scholarship for graduate studies from the Mexican National Council for Science and Technology (www.conacyt.gob.mx). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

4.2. Abstract

Infants recovering from anesthesia are at risk of life threatening Postoperative Apnea (POA). POA events are rare, and so the study of POA requires the analysis of long cardiorespiratory records. Manual scoring is the preferred method of analysis for these data, but it is limited by low intra- and inter-scorer repeatability. Furthermore, recommended scoring rules do not provide a comprehensive description of the respiratory patterns. This work describes a set of manual scoring tools that address these limitations. These tools include: (i) a set of definitions and scoring rules for 6 mutually exclusive, unique patterns that fully characterize infant respiratory inductive plethysmography (RIP) signals; (ii) RIPScore, a graphical, manual scoring software to apply these rules to infant data; (iii) a library of data segments representing each of the 6 patterns; (iv) a fully automated, interactive formal training protocol to standardize the analysis and establish intra- and inter-scorer repeatability; and (v) a quality control method to monitor scorer ongoing performance over time. To evaluate these tools, three scorers from varied backgrounds were recruited and trained to reach a performance level similar to that of an expert. These scorers used RIPScore to analyze data from infants at risk of POA in two separate, independent instances. Scorers performed with high accuracy and consistency, analyzed data efficiently, had very good intra- and inter-scorer repeatability, and exhibited only minor confusion between patterns. These results indicate that our tools represent an excellent method for the analysis of respiratory patterns in long data records. Although the tools were developed for the study of POA, their use extends to any study of respiratory patterns using RIP (e.g., sleep apnea, extubation readiness). Moreover, by establishing and monitoring scorer repeatability, our tools enable the analysis of large data sets by multiple scorers, which is essential for longitudinal and multicenter studies.

4.3. Introduction

Anesthesia enhances the susceptibility to apnea in infants [14-17, 19], leading to Postoperative Apnea (POA) events that may be life threatening, so infants require continuous cardiorespiratory monitoring [14, 15, 35]. POA events are rare with most occurring in the initial postoperative

hours, but a delayed onset, as late as 12 hours after surgery, has been reported [14, 17, 19]. Thus, any comprehensive study of POA requires the analysis of long data records.

Measuring infant respiration for extended periods of time requires a sensor that is well tolerated during both sleep and wakefulness. The initial studies of POA monitored respiration with thoracic impedance [14, 22, 52], the sensor of respiration most commonly used clinically in Postanesthesia Care Units (PACU). However, this sensor has important limitations leading to missed apneas, as both obstructive apnea and cardiogenic oscillations may often be misinterpreted as breathing [30]. Consequently, thoracic impedance is not recommended for research applications. The American Academy of Sleep Medicine (AASM) recommends the use of an airflow sensor (e.g., oronasal thermistor, or nasal pressure) to measure respiration and detect apnea [8]. However, airflow measurements require that sensors be attached to the face. These sensors are poorly tolerated by infants during recovery from surgery as they interfere with both sleep and feeding.

The AASM guidelines also designate the respiratory inductive plethysmograph (RIP) as an alternative sensor for apnea detection [8]. RIP uses two elastic bands that encircle the torso to measure ribcage (RCG) and abdominal (ABD) respiratory movements. These bands are well tolerated by infants and do not interfere with clinical care or the infant's behavioral state. RIP is the standard sensor for respiratory effort [8] in polysomnography and cardiorespiratory studies. It is also used to study respiration in other research applications including: prediction of extubation success in mechanically ventilated infants [119, 120], study of sudden infant death syndrome [118], and investigations of asthma [121] and bronchopulmonary dysplasia [122]. We have developed a data acquisition system that incorporates RIP sensors to monitor respiration, and a digital pulse oximeter to measure blood oxygen saturation (SAT) and photoplethysmography (PPG) [2], for the study of respiratory behavior of infants at risk of POA.

The investigation of POA using these data requires a consistent, reliable analysis method that fully characterizes the respiratory behavior of infants. The AASM endorses manual scoring as the "gold standard" for the study of apnea, and has published a set of rules to standardize the manual detection of apneas using RIP signals [8]. However these rules have 4 important

limitations. First, they assume that the RIP signals are calibrated; that is, the RCG and ABD signals are scaled so that their sum is proportional to tidal volume. This process is valid for a fixed spinal angle and constant posture [54], but becomes inaccurate when the measurement conditions and/or breathing patterns change [28, 79]. Consequently, the RIP calibration is likely to change throughout a long recording session invalidating the accuracy of the calibrated sum, making its use questionable. Second, the AASM rules only define clinically relevant apnea events, but do not define other respiratory patterns such as short respiratory pauses, thoraco-abdominal asynchrony, sighs, and normal breathing. Yet, these other patterns are relevant to the comprehensive study of respiratory behavior, since there is evidence that POAs are associated with abnormal respiratory patterns [14]. Indeed, we have found that an increased frequency of respiratory pauses, longer than 2 s, was associated with POA [82]. Third, the AASM rules must be applied by certified sleep laboratory technicians. As a result the analysis is costly and not widely available, since many sleep laboratories have long waiting times [123]. This severely constrains the amount of data that can be analyzed. Fourth, even when the AASM rules are applied by certified sleep laboratory technicians, the results have low intra- and inter-operator repeatability [9]. This adversely affects studies where multiple scorers are needed (e.g., large datasets, longitudinal, multicenter), because the repeatability of the analysis decreases with the number of scorers.

Advancement of the study of POA requires that these limitations be addressed. To do so we believe it is necessary to: (i) adapt the manual scoring rules to analyze uncalibrated RIP data; (ii) define a comprehensive set of RIP patterns; (iii) provide a computer-aided, scoring tool to improve accuracy and consistency, and reduce the time required for manual analysis; and (iv) develop a training and evaluation strategy to standardize the analysis and improve intra- and inter-operator repeatability. This Chapter describes a comprehensive set of tools developed to address these needs. These tools comprise 5 components: (i) a clear, comprehensive set of definitions and scoring rules for 6 mutually exclusive RIP patterns, (ii) a computer aided tool for efficient manual scoring, (iii) a library of data segments representing each of the 6 RIP patterns, (iv) a formal training protocol for scorers to standardize performance, and (v) a method to monitor the ongoing performance of scorers.

This Chapter is organized as follows. Section 4.4 describes the 5 manual scoring tools introduced above. Section 4.5 describes the methods used to evaluate these tools. Section 4.6 reports the results obtained by applying the tools to representative data from infants recovering from anesthesia. These results demonstrate that use of our tools produces efficient and accurate scoring with high intra- and inter-scorer repeatability regardless of operator expertise. Section 4.7 discusses the findings, and Section 4.8 provides concluding remarks.

4.4. Tools for Manual Scoring

4.4.1. Pattern Definitions and Scoring Rules

Our objective was to define a comprehensive set of respiratory inductive plethysmography (RIP) patterns that would provide a complete description of the respiratory behavior on a continuous, sample-by-sample basis. To this end, we carried out an extensive literature review related to the scoring rules for infant RIP data. Key sources included: (i) the Infant Sleep Apnea section of the revised International Classification of Sleep Disorders: Diagnostic and Coding Manual from the American Academy of Sleep Medicine (AASM) [35]; (ii) the updated AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications [8]; (iii) a series of articles on manual scoring published in the Journal of Clinical Sleep Medicine [124-131]; (iv) publications on POA in infants [7, 14, 15, 17, 19]; and (v) publications on thoraco-abdominal synchrony in infants [117, 122, 132]. This led us to define 6, mutually exclusive, unique patterns that would comprehensively characterize RIP signals. These patterns are: synchronous-breathing (SYB), asynchronous-breathing (ASB), sigh (SIH), respiratory pause (PAU), movement artifact (MVT), and unknown (UNK). Table 4.1 describes each pattern in detail, and provides the scoring rules for the unambiguous assignment of each data sample to one of the 6 patterns.

Pattern	Definition	Scoring Rule	Example
Synchronous-breathing (SYB)	Quasi-sinusoidal breathing patterns in RCG and ABD, where the inspiration and expiration movements of RCG and ABD are in phase.	Phase difference of less than 90°.	Fig. 4.3
Asynchronous-breathing (ASB)	Quasi-sinusoidal breathing patterns in RCG and ABD, where the RCG and ABD movements are out of phase.	Phase difference of 90° or more.	Fig. 4.4
Sigh (SIH)	A breath with considerably larger amplitude and duration than preceding breaths.	Breath amplitude and duration twice that of the epoch's average breath in both RCG and ABD.	Fig. 4.5
Movement artifact (MVT)	A period during which both RCG and ABD signals are corrupted by movements not related to respiration.	RCG and ABD display a chaotic, non-sinusoidal, low frequency motion.	Fig. 4.6
Respiratory pause (PAU)	A period where respiratory movements are absent in both RCG and ABD.	RCG and ABD have amplitudes less that 10 % of those of the preceding normal breath. A PAU begins at the start of inspiration of the first breath that is clearly reduced, and ends with the start of inspiration of the first breath whose amplitude returns to the epoch's average breath amplitude. If the start or end time of a PAU differs between RCG and ABD, the priority is given to the signal with higher SNR. All respiratory pauses are scored regardless of duration. Special cases: (i) PAU following SIH: RCG and ABD have amplitudes of less than 10 % of that of the breath preceding the sigh in both signals. (ii) PAU following MVT: RCG and ABD have amplitudes of less than 10 % of that of the breath that follows the pause in both signals.	Fig. 4.1 Fig. 4.7 Fig. 4.8
Unknown (UNK)	Any other pattern arising from technical problems (e.g., loss of a connector, high noise), or ambiguous patterns (e.g., MVT during SYB, different patterns in RCG and ABD).	RCG and/or ABD do not correspond to any other pattern.	Fig. 4.9

Table 4.1. Unique, mutually-exclusive patterns of respiratory inductive plethysmography and their scoring rules. RCG = ribcage, ABD = abdomen, SNR = signal-to-noise ratio.

4.4.2. RIPScore

RIPScore is an interactive computer application with a graphical user interface developed to support the efficient, manual scoring of RIP signals on a sample-by-sample basis. RIPScore is a redesign, and re-engineering of a rudimentary, prototype, manual scoring interface described in [6].

4.4.2.1. Main Screen

RIPScore displays data in 30 s epochs, and allows the scorer to segment the signals and assign a RIP pattern to each segment. Fig. 4.1 shows the main screen of RIPScore which comprises these main components:

(A) *RIP Pattern*: a color-coded bar showing the RIP pattern assigned by the scorer at each time;

(B) *Signals*: plots of the cardiorespiratory signals including ribcage (RCG), abdomen (ABD), photoplethysmograph (PPG), and blood oxygen saturation (SAT). Clicking on a breath from RCG or ABD plots three horizontal cursors, one at the estimated breath's amplitude, and two at ± 90 % of that amplitude. Note that these cursors are not an exact amplitude reference for the epoch because they do not take into account low frequency trends frequently observed in RIP signals [5];

(C) *Notes*: text boxes showing time stamped notes made during data acquisition, and comments entered by the scorer during analysis;

(D) *Segment and Epoch Control*: text boxes showing the start and end times for the current segment (highlighted in red in *Signals*); command buttons to add a "Comment" or "Delete" the RIP pattern assigned to the current segment; command buttons to scroll through epochs ("Previous", "Next"), and a text box with the start time of the current epoch;

(E) *Lissajous Figure*: a plot of RCG versus ABD for the current segment to aid the user in evaluating thoraco-abdominal synchrony. During breathing, the plot will be an ellipse tilted to the right for a phase less than 90 degrees, a circle for a phase of 90 degrees, and an ellipse tilted to the left for a phase greater than 90 degrees;

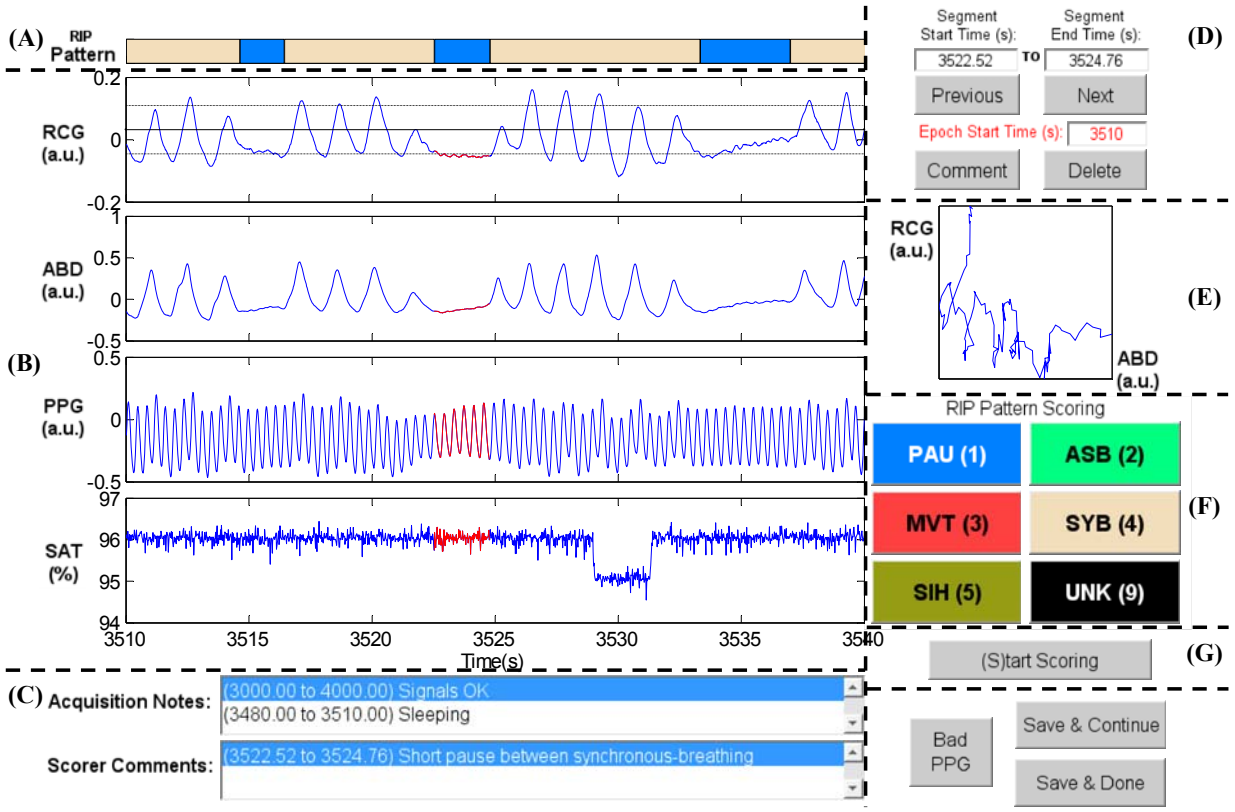

Fig. 4.1. Elements of the RIPScore interface. (A) Respiratory Inductive Plethysmography (RIP) Pattern; (B) Signals from ribcage (RCG), abdomen (ABD), photoplethysmograph (PPG), and blood oxygen saturation (SAT); (C) Notes; (D) Segment and Epoch Control; (E) Lissajous Figure; (F) RIP Pattern Scoring; and (G) Mode Control. The epoch shows a representative example of Pause (PAU). The quasi-sinusoidal pattern in RCG and ABD stops during the PAU highlighted in red. The horizontal dotted cursors in RCG show an estimated variation of ± 90 % of the amplitude of the breath preceding the PAU. Note that these cursors do not take into account low frequency trends, and so are only an approximate reference. a.u. = arbitrary units.

(F) *RIP Pattern Scoring*: color-coded command buttons that assign a RIP pattern to the current segment; each button may also be activated by hitting the corresponding keyboard "hot-key" defined by the character in parenthesis for each button (e.g., the hot-key for Pause is '1');

(G) *Mode Control:* command button to switch between scoring and visualization mode.

4.4.2.2. Operating Modes

RIPScore has 4 operating modes: Visualization/Review, Scoring, Training, and Evaluation. These modes support different aspects of the scoring process.

Visualization/Review Mode supports viewing the signals and reviewing the RIP patterns and annotations assigned throughout the record. In this mode, the "Previous" and "Next" buttons scroll the data in 20 s increments. Entering a value in "Epoch Start Time" moves the epoch display to that value. The *RIP Pattern Scoring* buttons move the data to the next segment assigned to that pattern.

Clicking a segment on the *RIP Pattern* bar selects the segment, highlights the segment in *Signals*, plots the corresponding *Lissajous Figure*, and updates the segment start and end time text boxes. The "Comment" command can be used to assign a comment to the segment, while the "Delete" command removes the RIP pattern assigned to it.

Scoring Mode supports manual scoring. When activated, the cursor changes to crosshairs, the display moves to the first unscored segment, the segment start is set to the first unscored sample, and RIPScore prompts the user to select the end of the segment. The selected *Signals* segment is highlighted in red, and RCG and ABD are plotted in the *Lissajous Figure*. The scorer then assigns a RIP pattern to the segment using a *RIP Pattern Scoring* button or its hot-key; the segment's assigned pattern, start and end time, and a timestamp are stored. The *RIP Pattern* bar is updated; and the display moves to the start of the next, unscored segment. This procedure continues until the scorer stops (by selecting the"(S)top Scoring" button) or all data have been scored. RIPScore then returns to Visualization/Review mode.

Training Mode supports the training of scorers by having users analyze simulated data with known RIP patterns. The interface is similar to that in Scoring Mode with the addition of an *Actual Pattern* bar for scored segments. If the trainee assigns an incorrect pattern to a segment, RIPScore displays an error message and provides the trainee with the opportunity to review the scored segment and reassign the pattern. Conversely, if the trainee assigns the correct pattern, RIPScore updates the *Actual Pattern* bar and allows the trainee to continue. A Training Mode session ends once the trainee has either: (i) scored the complete record, or (ii) correctly scored 5 patterns of each type consecutively.

The simulated infant RIP records used in Training Mode are generated by concatenating, i.e., linking together, signal segments with known RIP patterns to yield continuous signals. Fig. 4.2 illustrates the concatenation method, which consisted of the following 4 steps:

(i) two input signal segments were selected to be concatenated;
(ii) the 2 signal segments were aligned with an overlap (transition window T) of N_T samples; that is, the last N_T samples of the first segment overlapped the first N_T samples of the second segment;
(iii) the samples of the first segment in the transition window were gradually attenuated by multiplying them by a decaying sigmoid factor that varied from 1 to 0 over the length of the window; samples of the second segment were gradually amplified by multiplication with a sigmoid factor that increased from 0 to 1 over the window length; the modified signals in the transition window were then added to yield a smooth transition; and
(iv) the output signal consisted on the first segment up to the start of T, followed by the transition, and then by the second segment starting after T.

The concatenation method overlapped the input segments to produce a smooth transition. This was done to avoid transition artifacts, which could generate sharp transients that do not resemble natural RIP patterns.

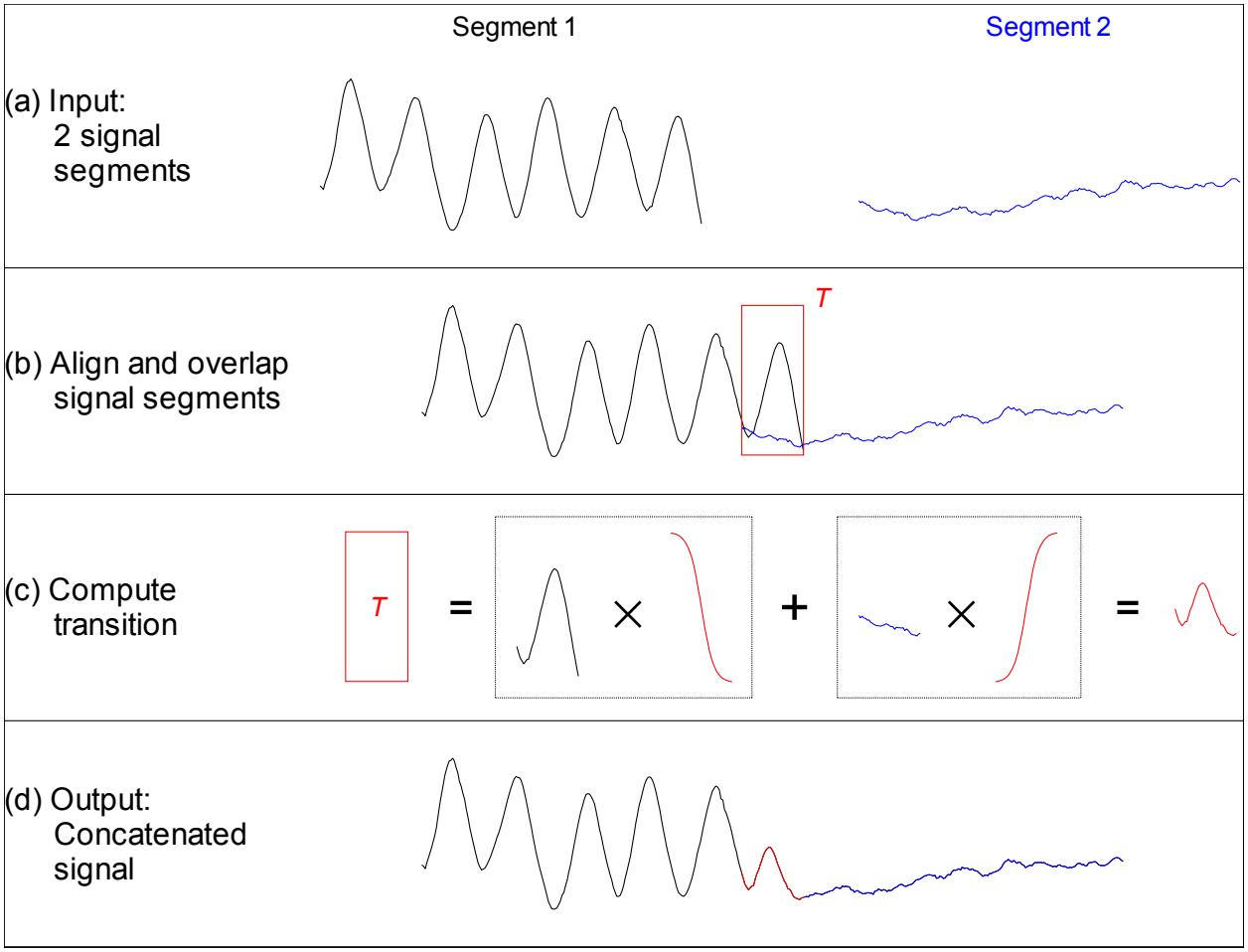

Fig. 4.2. Concatenation of signal segments. (A) Sample input segments. (B) Input segments are aligned and overlapped over a transition window T. (C) The output during this window is computed by gradually attenuating the end of the first segment, gradually incrementing the start of the second segment, and adding the two parts to yield a smooth transition. (D) The output signal consists on the first segment up to the start of T, followed by the transition, followed by the second segment starting after T.

RIPScore uses two types of simulated data, and investigators are required to configure which type to use before scoring sessions start. Type I "simulated-pattern" data was based on signals generated using a breath-by-breath time-series model of infant breathing; other RIP patterns were simulated by manipulating these signals as described in [85]. Type II "true-pattern" data comprised segments of real data whose RIP pattern was determined during a reference analysis (REF) performed by one of the authors (KAB) as described below. Type II data were more complex and realistic than Type I because they incorporated the inherent variability of real infant breathing.

A new, 1 hr long, Training Mode data record is generated for each training session as follows:

(i) segments of each RIP pattern category are simulated and stored in a list, until the total length of data is > 1.5 hr;
(ii) the list of simulated segments is re-ordered randomly;
(iii) the list is examined to ensure that contiguous segments have different RIP patterns, if two contiguous segments have the same pattern, the second segment is pushed to the end of the list;
(iv) the list is truncated to the first N segments whose total length is 1 hr; and
(v) the segments on the list are concatenated as described in Fig. 4.2.

Evaluation Mode is used to evaluate a scorer's accuracy and consistency. In this mode, the user analyzes a simulated data record with an interface similar to Training Mode, but with no feedback. Upon completion, RIPScore: (i) estimates the accuracy and consistency of the scorer; (ii) stores the accuracy and consistency values, the simulated data record, and the assigned RIP patterns; (iii) displays the accuracy and consistency to the scorer; and (iv) reveals the *Actual Pattern* bar in Review Mode so that the scorer can compare their assigned patterns to the actual, simulated patterns.

Data for Evaluation Mode are generated as follows:

(i) the first 30 min of data segments are simulated and stored in a list as for the Training data;
(ii) the list is duplicated;

(iii) the duplicate list is re-ordered randomly, and contiguous segments with equal RIP patterns pushed to the end;
(iv) the two lists are joined, and the segments concatenated.

Thus, in the evaluation data record each simulated segment appears in both the first and second half but in a different, random order.

Performance is assessed in terms of the accuracy and consistency of the assigned RIP patterns. Accuracy is measured as the agreement between patterns assigned by the trainee and the actual pattern. Consistency is measured as the agreement between the patterns assigned to the same segments in the first and second half of the evaluation record. Agreement is quantified using the Fleiss' kappa (κ) statistic [133, 134] computed on a sample-by-sample basis as in [85, 135]. This kappa implementation generalizes the traditional Cohen's κ statistic [136] to evaluate agreement between multiple scorers when classifying observations into two or more categories.

4.4.2.3. Sample Patterns in RIPScore

Examples of the RIP patterns and special cases defined in Table 4.1 are illustrated in the following figures.

- Synchronous-Breathing (SYB, Fig. 4.3): the selected breaths in RCG and ABD (in red) are in phase, and the Lissajous plot is an ellipse tilted to the right;
- Asynchronous-Breathing (ASB, Fig. 4.4): the selected breaths are out of phase, and the Lissajous plot is elliptical and tilted to the left;
- Sigh (SIH, Fig. 4.5): the dotted horizontal cursor in RCG provides an approximate reference showing that the sigh has an amplitude of more than 190 % of that of the preceding breath, with a duration longer than that of the other breaths;
- Movement Artifact (MVT, Fig. 4.6): low-frequency motion corrupts both RCG and ABD;
- Pause (PAU, Fig. 4.1): the pause at the middle of the epoch has an amplitude of less that 10 % of that of the preceding breath, as evidenced by the horizontal cursor in RCG;
- PAU which follows a SIH (Fig. 4.7): the horizontal cursors in the ABD signal show approximate reference amplitudes for the breath preceding the sigh; it is clear that the

sigh's amplitude is much larger, and that at least part of the pause's amplitude is below the 10 % dotted line;

- PAU which follows a MVT (Fig. 4.8): the horizontal cursor in RCG suggests that the amplitude during the pause is less than 10 % of that of the breath that follows the pause;
- Unknown (UNK, Fig. 4.9): RCG and ABD have different patterns; RCG shows low-frequency movement artifact, while ABD shows breathing.

4.4.3. Library of Segments with Known Patterns

A library containing "true-pattern" data segments representative of each of the 6 RIP patterns was created for use in RIPScore Training and Evaluation Modes.

4.4.3.1. Infant Data

The library was built using data acquired from 24 infants (19 male, birth age 31 ± 4 weeks, postmenstrual age 43 ± 2 weeks, weight 3.7 ± 1.0 kg) recruited for a prospective POA study. Inclusion criteria were: (i) postmenstrual age < 60 weeks at the time of surgery in preterm infants, and < 48 weeks in term infants, (ii) elective surgery for inguinal herniorrhaphy, and (iii) American Society of Anesthesiology physical status 1 or 2. Exclusion Criteria were: (i) post-operative admission to the Neonatal Intensive Care Unit or Pediatric Intensive Care Unit, (ii) emergency surgery, and (iii) spinal anesthesia. The anesthetic technique was not standardized.

Data were acquired in the Postanesthesia Care Unit (PACU) of the Montreal Children's Hospital using a custom-built monitoring system [2]. Upon admission to the PACU, infant respibands (Inductobands, Ambulatory Monitoring Inc., Ardsley, NY, USA) were placed around the ribcage (at the nipple line) and abdomen (at the umbilicus) and interfaced with a Respiratory Inductive Plethysmograph (Battery Operated Inductotrace, Ambulatory Monitoring Inc., Ardsley, NY, USA). An infant oximeter probe (Nonin 8600 Portable Digital Pulse Oximeter, Nonin Medical Inc., Plymouth, MN, USA) was taped to a digit. The outputs were low-pass filtered (cut-off frequency 10 Hz) with an 8-pole, anti-aliasing, Bessel filter (Kemo, Jacksonville, FL, USA), sampled at 50 Hz, and stored. Subsequent, off-line analysis was performed using MATLAB (The MathWorks Inc., Natick, MA, USA). No attempt was made to calibrate the RIP signals.

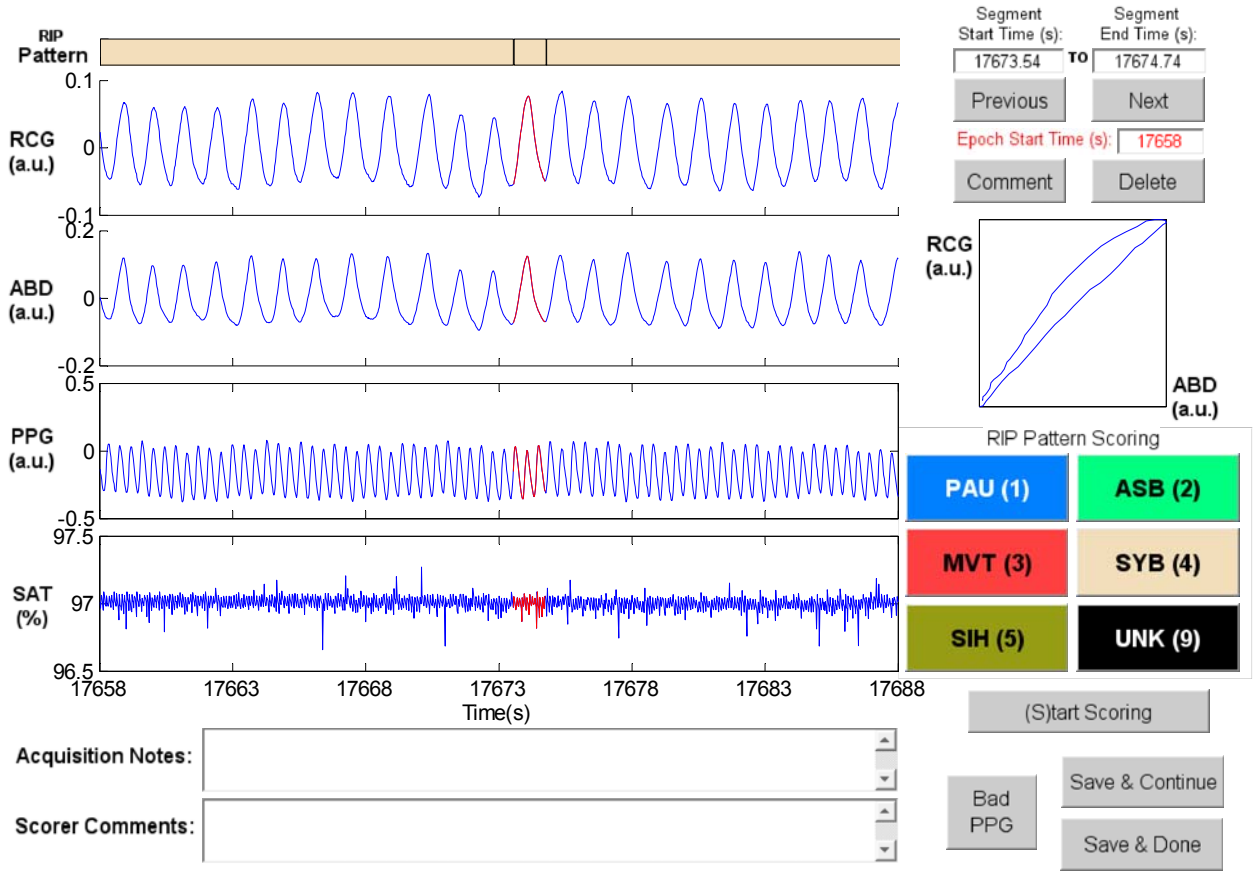

Fig. 4.3. Representative example of Synchronous-Breathing (SYB). The ellipse in the Lissajous plot of ribcage (RCG) against abdomen (ABD) is tilted to the right. PPG = photoplethysmograph, SAT = blood oxygen saturation, a.u. = arbitrary units.

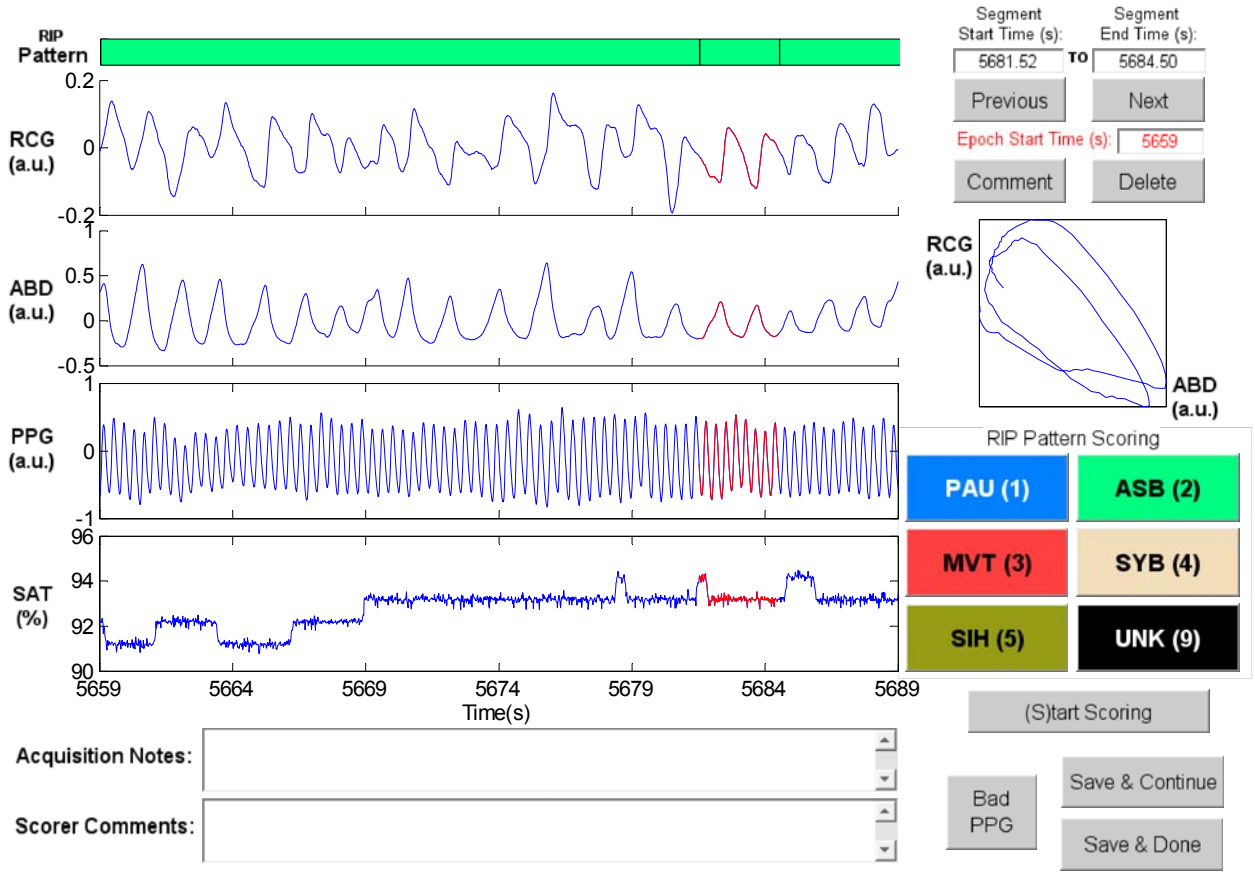

Fig. 4.4. Representative example of Asynchronous-Breathing (ASB). The Lissajous plot of ribcage (RCG) against abdomen (ABD) for the segment highlighted in red shows ellipses tilted to the left. PPG = photoplethysmograph, SAT = blood oxygen saturation, a.u. = arbitrary units.

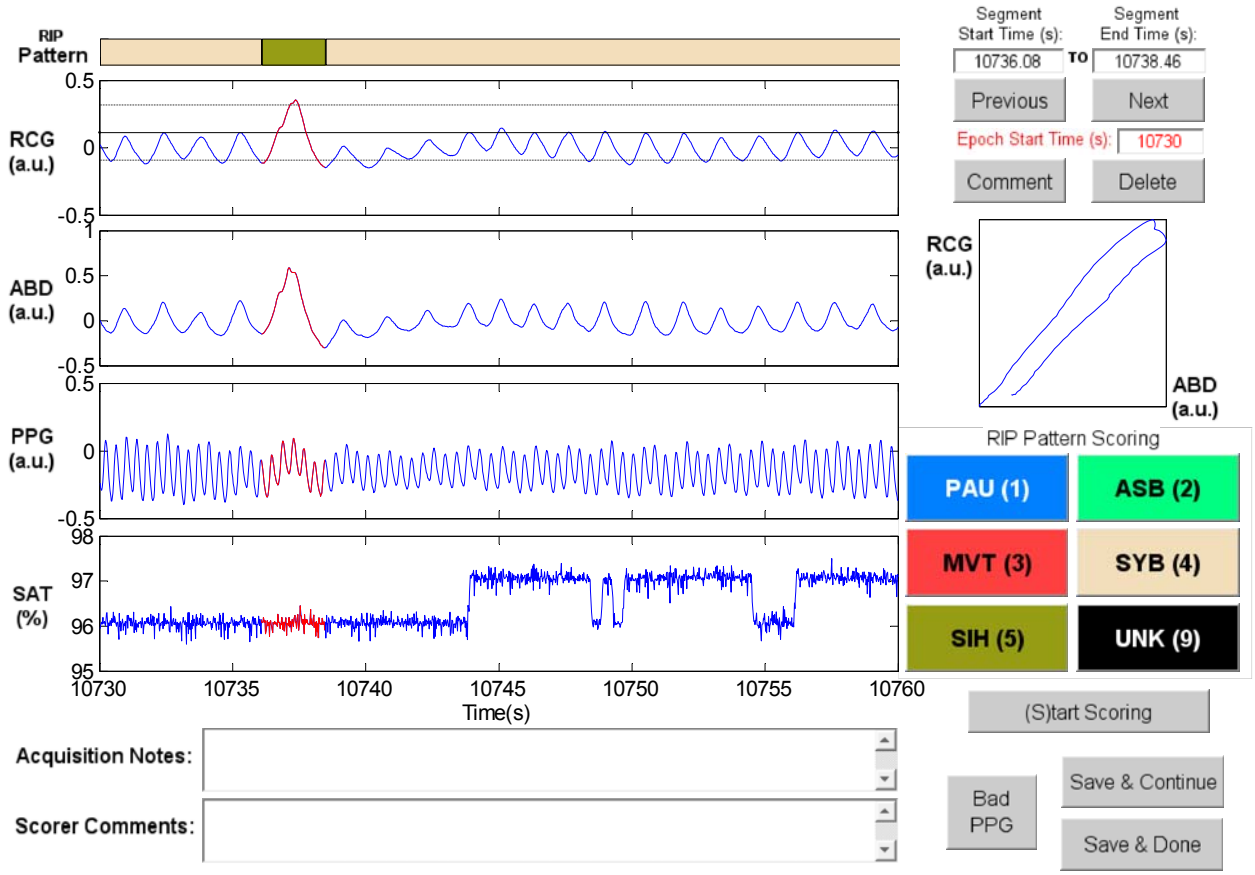

Fig. 4.5. Representative example of Sigh (SIH). The SIH highlighted in red has larger amplitude and longer duration than the other breaths. The horizontal dotted cursors in the ribcage (RCG) signal show an estimated variation of ± 90 % of the amplitude of the breath preceding the SIH. Note that these cursors are not an exact amplitude reference. Also, the Lissajous plot shows an ellipse tilted to the right. ABD = abdomen, PPG = photoplethysmograph, SAT = blood oxygen saturation, a.u. = arbitrary units.

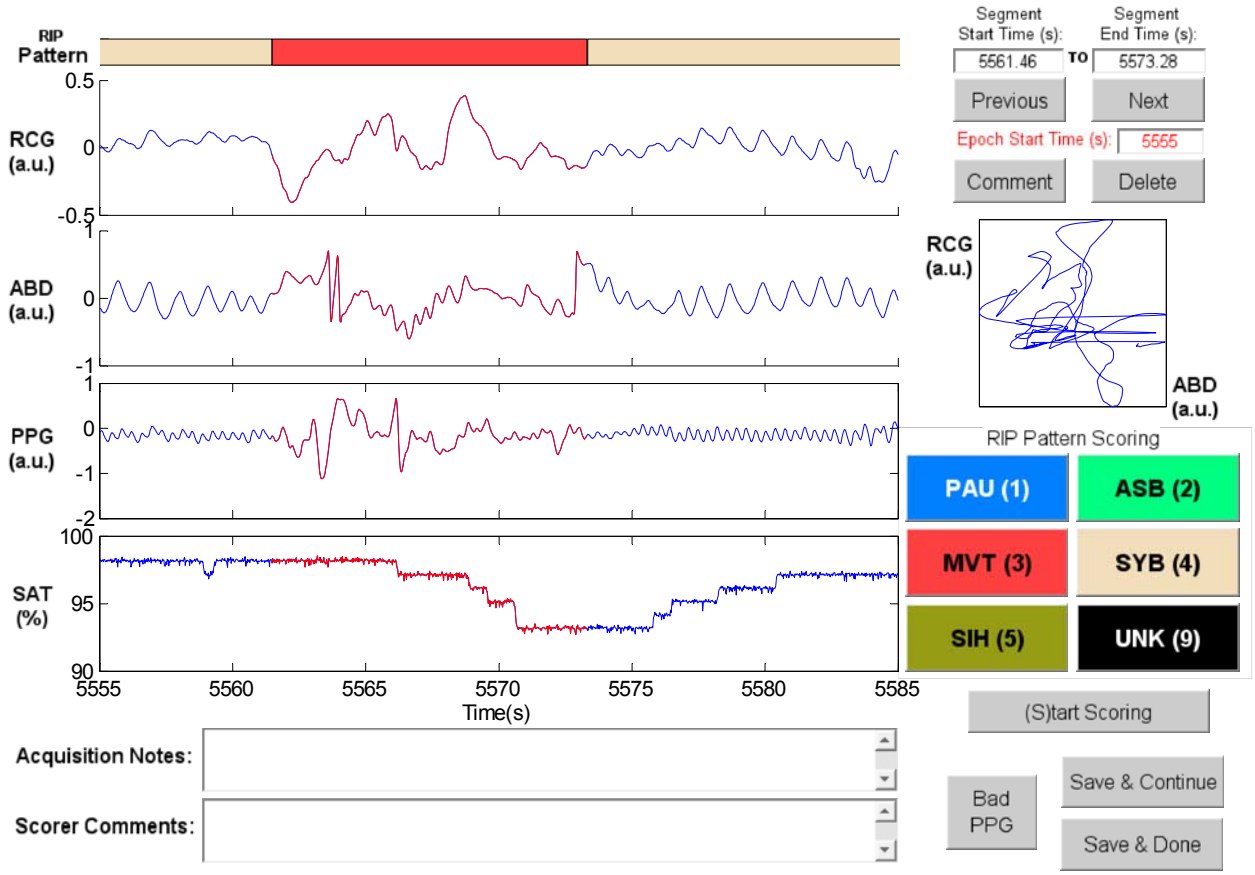

Fig. 4.6. Representative example of Movement Artifact (MVT). The MVT in the ribcage (RCG) and abdomen (ABD) signals is highlighted in red. PPG = photoplethysmograph, SAT = blood oxygen saturation, a.u. = arbitrary units.

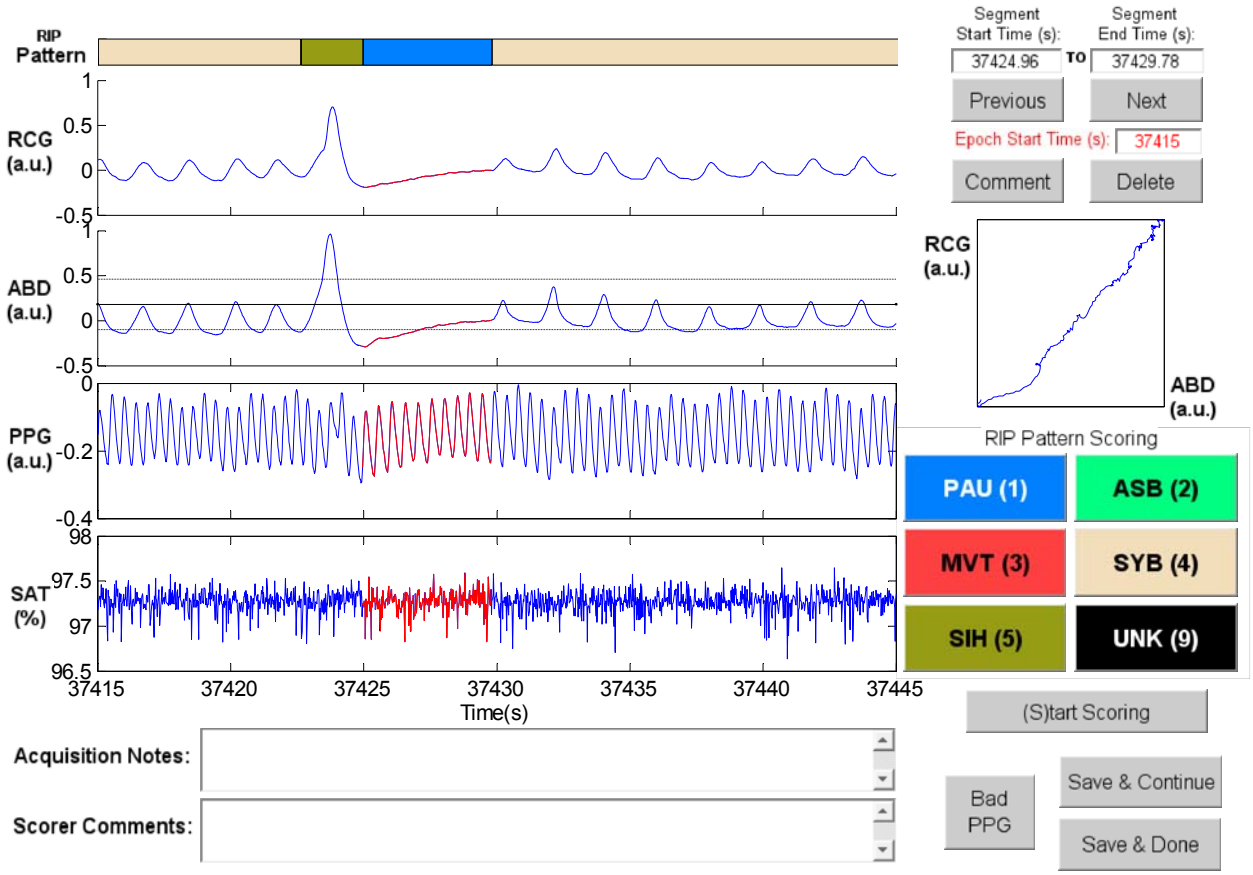

Fig. 4.7. Representative example of a Pause (PAU) which follows a Sigh (SIH). The horizontal dotted cursors in the abdomen (ABD) signal show an estimated variation of ± 90 % of the amplitude of the breath that precedes the SIH. Note that these cursors are not an exact amplitude reference. RCG = ribcage, PPG = photoplethysmograph, SAT = blood oxygen saturation, a.u. = arbitrary units.

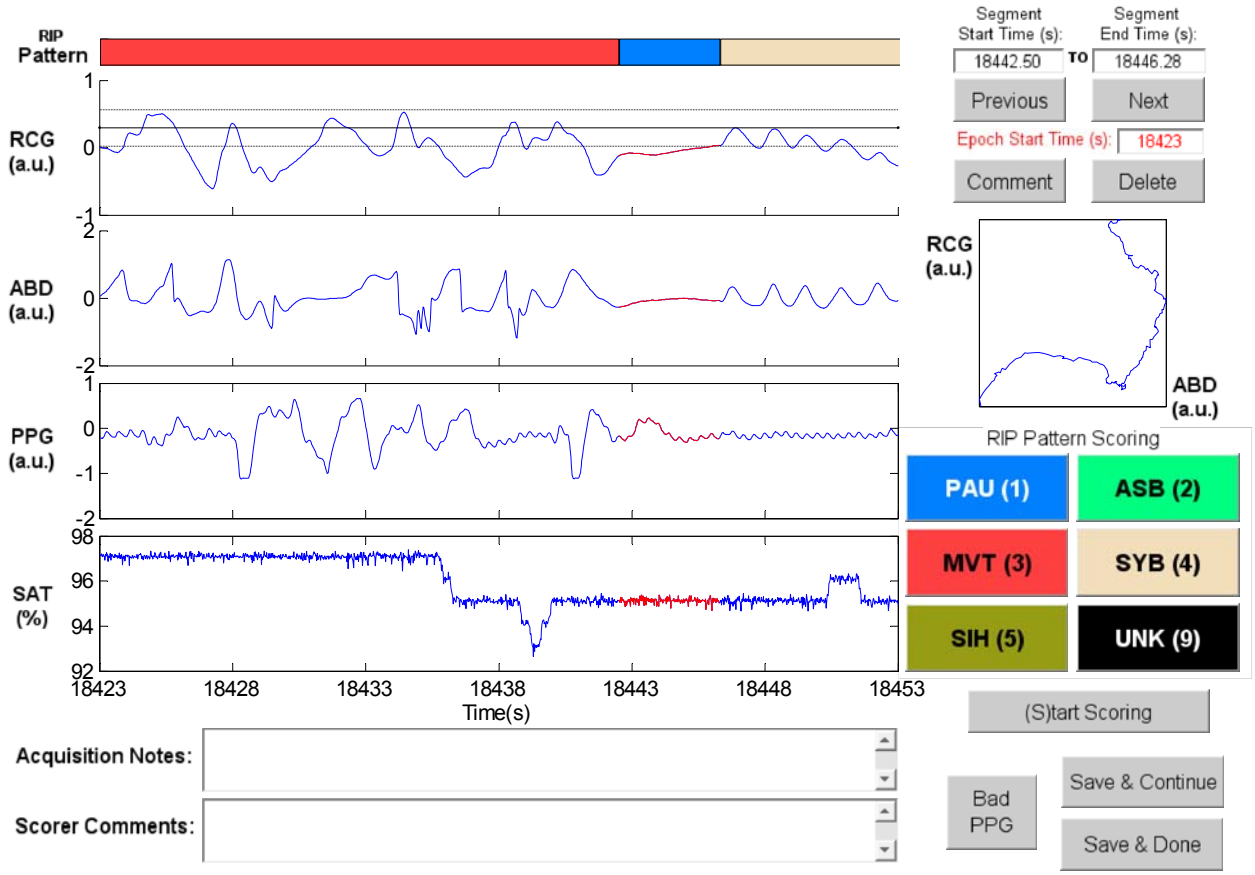

Fig. 4.8. Representative example of a Pause (PAU) which follows a Movement Artifact (MVT). The horizontal dotted cursors in the ribcage (RCG) signal show an estimated variation of ± 90 % of the amplitude of the breath that follows the PAU. Note that these cursors are not an exact amplitude reference. ABD = abdomen, PPG = photoplethysmograph, SAT = blood oxygen saturation, a.u. = arbitrary units.

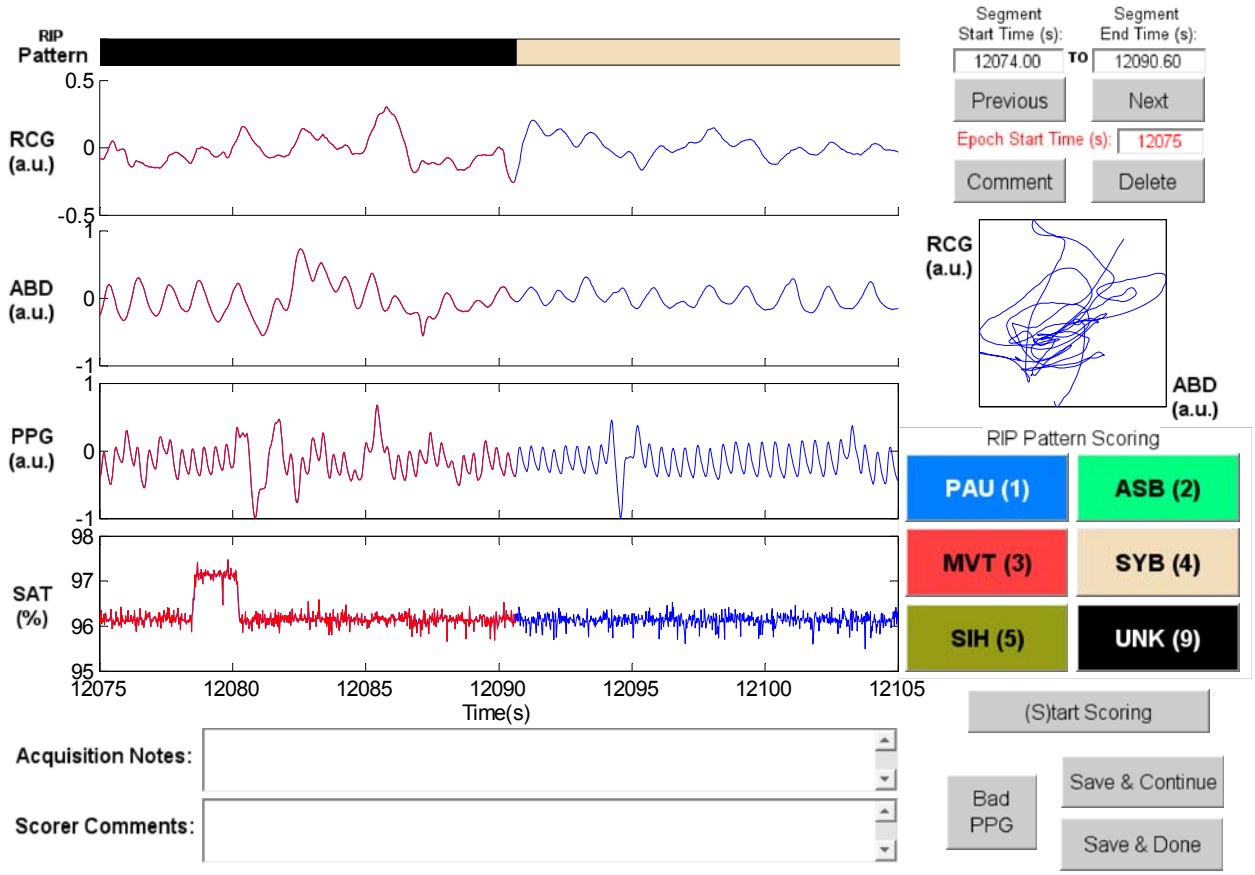

Fig. 4.9. Example of Unknown (UNK). It is not possible to determine the pattern in the selected segment (red) because the ribcage (RCG) signal shows a low-frequency, chaotic pattern, while the abdomen (ABD) signal has a quasi-sinusoidal breathing pattern with an additional low-frequency movement component. PPG = photoplethysmograph, SAT = blood oxygen saturation, a.u. = arbitrary units.

Recordings were 9.0 ± 2.2 hr long. Subsets of these data have been used in previous work [135, 137-139].

Recording sessions were continuously attended, and a paper record of the infant's behavioral state, i.e., sleeping, feeding, diaper change, etc., was kept, referenced to the clock time and recording time. These handwritten entries were transcribed to an electronic text file and displayed as acquisition *Notes* in RIPScore. Demographic data and relevant clinical variables, including anesthetic and analgesic drug regimen, were recorded.

4.4.3.2. Ethics Statement

The study was approved by the Institutional Review Board of the McGill University Health Centre / Montreal Children's Hospital (approval numbers PED-07-30, and 12-308-PED). Written, informed parental consent was obtained for each infant recruited to the study. Consent for publication of raw data was not requested specifically at the time the study was carried out. However, all materials have been thoroughly inspected, and all possible identifiers (as defined in [13]) were removed before the data were made available publicly. Thus, we believe that publication of these data poses negligible risk to the privacy of study participants.

4.4.3.3. Reference Manual Analysis

One of the authors (KAB) served as the reference scorer (REF). REF has extensive experience in the manual scoring of infant cardiorespiratory data, participated in the data acquisition, and contributed to the development of RIPScore.

REF used RIPScore to analyze the full records of 23 infants in two independent instances; the order in which the data records were analyzed was randomized between instances. One record was excluded because the infant was continuously handled by nurses and parents throughout the recording session. REF's overall intra-scorer repeatability, measured with the Fleiss' kappa statistic [133, 134], was "substantial" ($\kappa = 0.80$) [140]. Samples where REF assigned the same RIP pattern in the two instances were considered to be correct and defined the "true-pattern" for these samples.

This reference scoring task was very labor intensive and required 8 months to complete. For this reason, data were partitioned into two subsets: (i) a validation subset used to evaluate the performance of scorers, and (ii) a library of "true-pattern" segments used to generate the Type II "true-pattern" simulated data. Fig. 4.10 summarizes how the validation subset and the "true-pattern" segment library were created.

The validation subset comprised data from 21 infants, truncated to a maximum of 20,000 s per record, representing a 54 % of the complete data set. Records from 2 infants that were analyzed by REF were excluded due to bad quality in the recordings. To ensure that the validation subset was representative, the proportion of "true-pattern" samples assigned to each RIP pattern was computed for both the complete and truncated data records. The Wilcoxon signed rank test [141] indicated that the proportions were not significantly different as Table 4.2 shows.

The library of "true-pattern" segments was created from remaining data and comprised 16,285 segments.

4.4.4. Training Protocol

All scorers underwent a common training protocol, using RIPScore Training and Evaluation Modes, to standardize the analysis and performance of scorers using our tools.

Fig. 4.11 shows a block diagram of the training protocol. Training had 2 levels, each having two stages: training and evaluation. Trainees started at Level 1, where they were familiarized with RIPScore, the 6 mutually exclusive RIP pattern definitions, and the scoring rules, by analyzing Type I "simulated-pattern" records (Fig. 4.12A). Each level began with a training stage where trainees scored data in RIPScore Training Mode. Upon completing the training stage, their accuracy and consistency were evaluated using RIPScore Evaluation Mode. If their performance was adequate (see Fig. 4.11) they advanced to Level 2 of training, if not, they repeated the Level 1 training stage.

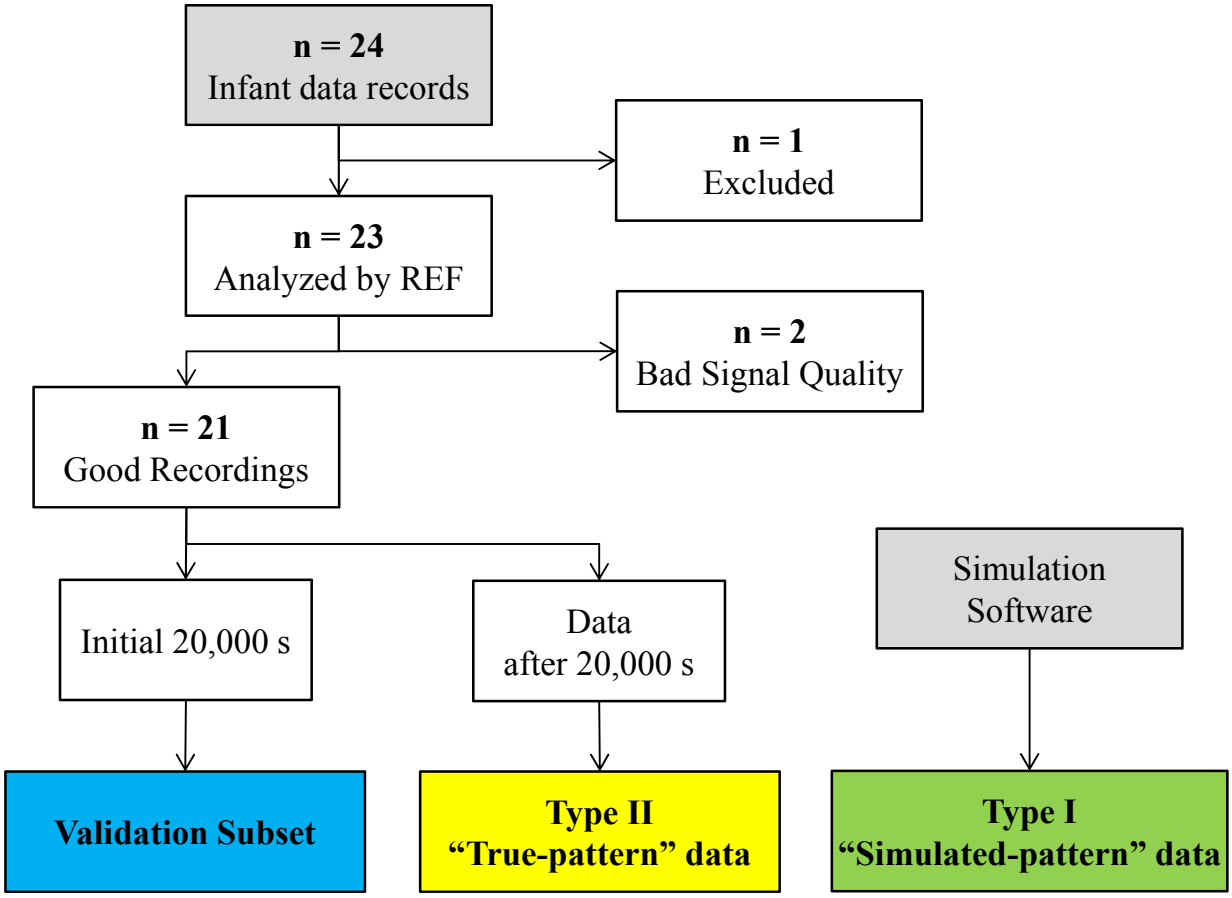

Fig. 4.10. Study Data Flowchart.

Pattern	Complete Record	Truncated, Validation Record	p-value
SYB	0.73 [0.08]	0.75 [0.06]	0.13
ASB	0.03 [0.05]	0.02 [0.05]	0.13
SIH	0.01 [0.00]	0.01 [0.00]	0.28
PAU	0.02 [0.03]	0.02 [0.02]	0.25
MVT	0.12 [0.03]	0.12 [0.06]	0.15
UNK	0.08 [0.04]	0.08 [0.04]	0.39

Table 4.2. Proportion of "true-pattern" samples in the records used to create the validation data subset. Results presented as median [interquartile range]. SYB = synchronous-breathing, ASB = asynchronous-breathing, SIH = sigh, PAU = respiratory pause, MVT = movement artifact, UNK = unknown.

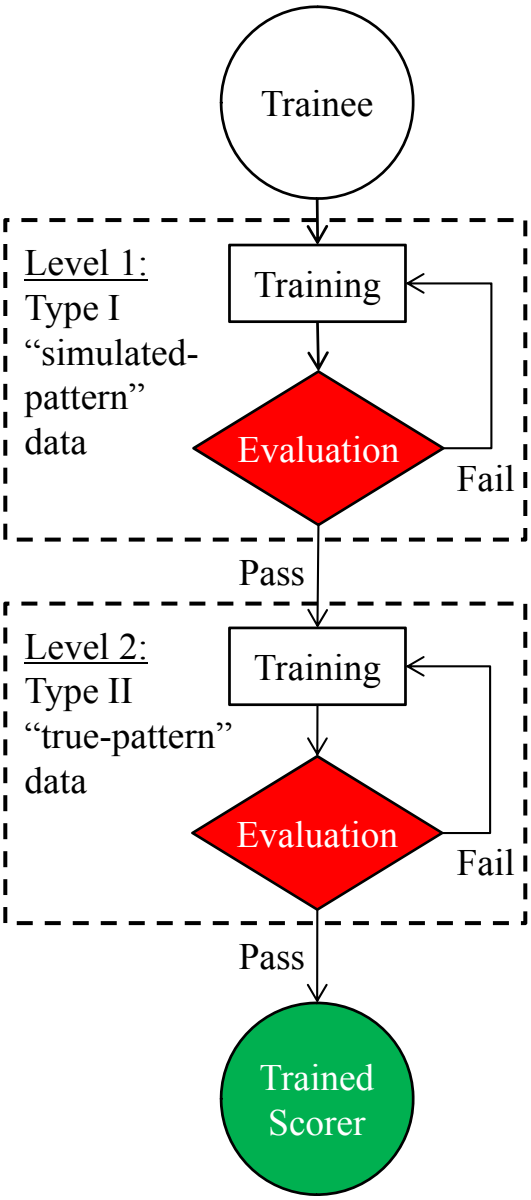

Fig. 4.11. Scorer training protocol. Criteria to successfully complete levels: (A) Level 1, the trainee obtained accuracy and consistency values of $\kappa \geq 0.8$; and (B) Level 2, the trainee obtained accuracy and consistency values of $\kappa \geq 0.8$ on two consecutive sessions.

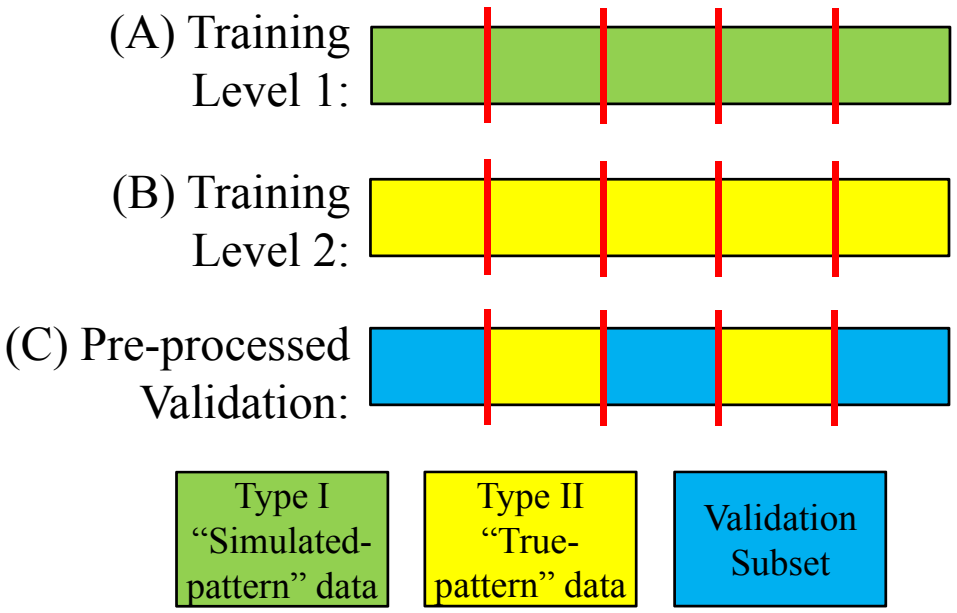

Fig. 4.12. Data formats. (A) Type I, and (B) Type II data segments were concatenated to generate the training records. (C) Validation records were pre-processed such that Type II segments were inserted into the validation subset. Red vertical lines indicate the concatenation point.

Level 2 training proceeded in a similar manner except that the data analyzed were the more realistic Type II "true-pattern" data records (Fig. 4.12B). Training was completed after successful completion of the Level 2 evaluation stage (see Fig. 4.11).

Reference values of performance were obtained by having REF analyze two sessions of each training level. These analyses showed that REF had excellent consistency and accuracy values ranging from $\kappa = 0.76$ to $\kappa = 0.89$.

4.4.5. Monitoring of Scorers for Quality Control

Scorer accuracy and consistency were evaluated on a record-by-record basis using a quality control method based on the pre-processing phase described next.

4.4.5.1. Pre-processing

The validation dataset was pre-processed by inserting Type II "true-pattern" segments into each data record (Fig. 4.12C). Thus, for this pre-processing phase, a total of 152 segments (1,000 s worth of data) were selected from the "true-pattern" segment library, such that each RIP pattern was equally represented. The distribution of these 152 segments was: 25 SYB, 26 ASB, 27 SIH, 22 PAU, 27 MVT, and 25 UNK.

For each data record in the validation subset, the 152 segments were randomly ordered and inserted into the first 3 hrs of the record at randomly selected times. These "true-pattern" segments were then randomly re-ordered, and inserted into the last 3 hr of the record at random times. Segments were inserted by splitting the data record (see Fig. 4.12C), and concatenating the segment as in Fig. 4.2. Thus each of the 21 pre-processed data records contained two copies of the 152 "true-pattern" segments.

These inserted "true-pattern" segments were then used to evaluate scorer accuracy and consistency using the same methods as in RIPScore's Evaluation Mode.

4.5. Evaluation of the Manual Scoring Tools

The manual analysis tools were evaluated by examining the performance of three scorers in the analysis of the pre-processed validation data subset.

4.5.1. Scorer Recruitment and Training

The three scorers had quite different backgrounds and experience in the analysis of respiratory data. The first (SC1) was a pediatric anesthesiologist with expertise in infant respiratory physiology, who participated in data acquisition and is a co-author (GB). The second (SC2) was a senior respiratory pediatric sleep laboratory technician with extensive experience in manual scoring of pediatric cardiorespiratory data. The third (SC3) was a computer network analyst with a master's degree in telecommunications but no clinical expertise. All three scorers were trained using the protocol.

4.5.2. Validation of the Manual Analysis Tools

The three scorers analyzed the entire, pre-processed, validation data subset in two independent, blinded instances; the order of the data records was randomized between instances and between scorers. Scorer performance was evaluated in terms of the following parameters.

4.5.2.1. Accuracy and Consistency

The two copies of the 152 "true-pattern" segments inserted in each data record were analyzed to evaluate the scorers' ongoing accuracy and consistency.

4.5.2.2. Scoring Rate

The time required to score a data record was estimated by summing the difference between the timestamps of consecutive scores. Differences greater than 2 min were excluded because they likely resulted from interruptions in the analysis. The overall scoring rate was estimated as the ratio of the length of a data record (in data hours) to the hours required to score it. Pattern-specific scoring rates were estimated as the ratio of the total length of segments assigned to a RIP pattern to the time required to score those segments.

4.5.2.3. Intra- and Inter-Scorer Repeatability

Intra- and inter-scorer repeatability of the RIP patterns assigned to the validation data were assessed using the Fleiss' kappa (κ) statistic [133, 134] on a sample-by-sample basis.

4.5.2.4. Confusion Analysis

Confusion in the scoring of the 6 RIP patterns $\Theta = \{SYB, ASB, SIH, PAU, MVT, UNK\}$ was assessed by computing the confusion matrix $\mathbf{P}$ whose elements $P_{i,j}$ gave the conditional probability that a sample with consensus pattern i would be scored as pattern j. A sample x_k was assigned a consensus RIP pattern $Cn(x_k) \in \Theta$ if it was assigned that pattern in the absolute majority (4 or more) of the 6 scoring iterations. Samples without consensus pattern were excluded from the confusion analysis. Thus, to estimate $P_{i,j}$ for each scorer, the N_i samples with consensus pattern i were identified. Then, N_j, the number of times the N_i samples had been assigned to pattern j, was determined. Finally, the conditional probability was estimated as $P_{i,j} = N_j / N_i$. Confusion matrices were computed for each scorer separately, and also as a group.

To assess the effects of segment length, confusion matrices were also computed after excluding scored segments shorter than a threshold (varied from 0 s to 20 s).

4.5.3. Statistical Analysis

Bootstrapping [142] with 100 resamples was used to estimate the standard deviation of the κ values and the confusion matrix probabilities. Values of κ were interpreted according to the intervals proposed in [140]: $\kappa < 0$ = poor, $0 \le \kappa \le 0.2$ = slight, $0.2 < \kappa \le 0.4$ = fair, $0.4 < \kappa \le 0.6$ = moderate, $0.6 < \kappa \le 0.8$ = substantial, and $0.8 < \kappa \le 1$ = almost perfect. Random selections were drawn from a uniform distribution where all instances had equal probability of being selected.

4.6. Results

4.6.1. Training

Tables 4.3 and 4.4 show the accuracy and consistency of the scorers for each training session and level. All scorers reached the required Level 1 performance ($\kappa \geq 0.8$) after the first session. None of the scorers reached the required performance in the first Level 2 session; SC1 and SC3 had low accuracy, and SC1 and SC2 had low consistency. Scorer performance improved with training and all 3 achieved the required level of accuracy and consistency ($\kappa \geq 0.8$) in sessions 2 and 3 of Level 2, completing the training protocol requirements.

4.6.2. Accuracy and Consistency

Fig. 4.13 documents the performance of the scorers as a function of the number of records scored. Fig. 4.13A shows that the overall scoring accuracy was substantial and nearly constant throughout the scoring effort for all three scorers (SC1: $\kappa = 0.66 \pm 0.02$, SC2: $\kappa = 0.74 \pm 0.02$, SC3: $\kappa = 0.67 \pm 0.03$). Consistency (Fig. 4.13B) was high throughout for SC1 ($\kappa = 0.79 \pm 0.03$) and SC2 ($\kappa = 0.79 \pm 0.02$); SC3 ($\kappa = 0.77 \pm 0.05$) started slightly lower, but quickly reached a level similar to the other scorers.

Analysis of pattern-specific accuracy and consistency revealed some substantial differences between scorers for 3 RIP patterns: PAU, MVT, and UNK. For PAU, Fig. 4.14 shows that two scorers had high, nearly constant levels of accuracy (SC1: $\kappa = 0.76 \pm 0.06$, SC2: $\kappa = 0.72 \pm 0.06$) and consistency (SC1: $\kappa = 0.73 \pm 0.07$, SC2: $\kappa = 0.78 \pm 0.06$). In contrast, SC3, the scorer with non-clinical background, had lower accuracy ($\kappa = 0.34 \pm 0.14$) and consistency ($\kappa = 0.44 \pm 0.11$). For MVT (Fig. 4.A1), the three scorers had similar consistency, but a range of accuracies, with SC2 having the highest ($\kappa = 0.75 \pm 0.03$), followed by SC3 ($\kappa = 0.65 \pm 0.07$), and SC1 with the lowest ($\kappa = 0.53 \pm 0.02$). For UNK (Fig. 4.A2), the accuracy of SC2 ($\kappa = 0.54 \pm 0.07$) and SC3 ($\kappa = 0.46 \pm 0.06$) were moderate, while that of SC1 was poor ($\kappa = 0.03 \pm 0.05$). As would be expected the consistency of SC1 for UNK was much lower ($\kappa = 0.29 \pm 0.09$) than those of SC3 ($\kappa = 0.66 \pm 0.11$), and SC2 ($\kappa = 0.58 \pm 0.06$).

Scorer	Level 1	Level 2		
	Session 1	*Session 1*	*Session 2*	*Session 3*
SC1	0.94	0.72	0.82	0.81
SC2	0.94	0.81	0.86	0.87
SC3	0.94	0.79	0.82	0.81

Table 4.3. Training accuracy. Level 1 = Type I "simulated-pattern" data. Level 2 = Type 2 "true-pattern" data. Performance was measured using the Fleiss' κ statistic [133]. The standard deviation was < 0.01 in all cases.

Scorer	Level 1	Level 2		
	Session 1	*Session 1*	*Session 2*	*Session 3*
SC1	0.89	0.74	0.86	0.81
SC2	0.90	0.76	0.83	0.84
SC3	0.93	0.86	0.85	0.80

Table 4.4. Training consistency. Level 1 = Type I "simulated-pattern" data. Level 2 = Type 2 "true-pattern" data. Performance was measured using the Fleiss' κ statistic [133]. The standard deviation was < 0.01 in all cases.

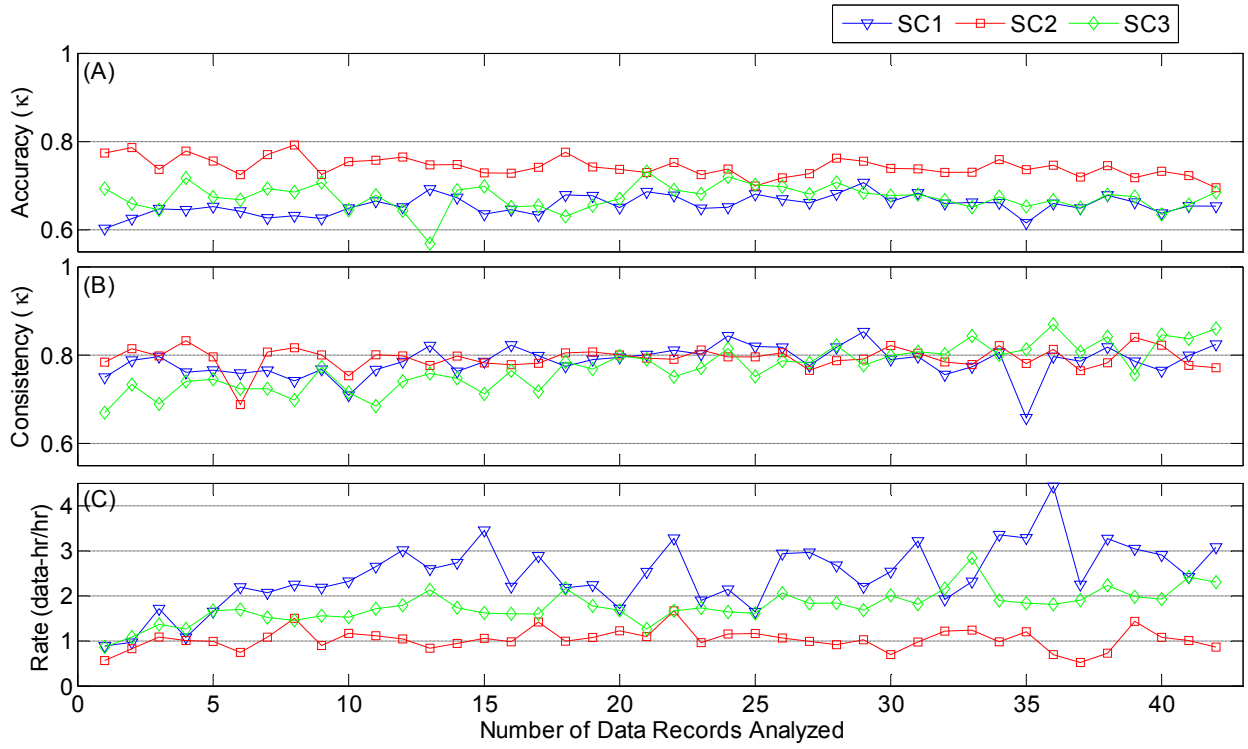

Fig. 4.13. Overall scoring performance. (A) Accuracy (Fleiss' κ); (B) consistency (Fleiss' κ); and (C) rate (hours of data per hour of scoring) as a function of number of data records analyzed. SC1 was a pediatric anesthesiologist; SC2 was an experienced sleep laboratory scorer; and SC3 was a data networks analyst with no clinical experience. Standard deviation of each accuracy and consistency point was < 0.01.

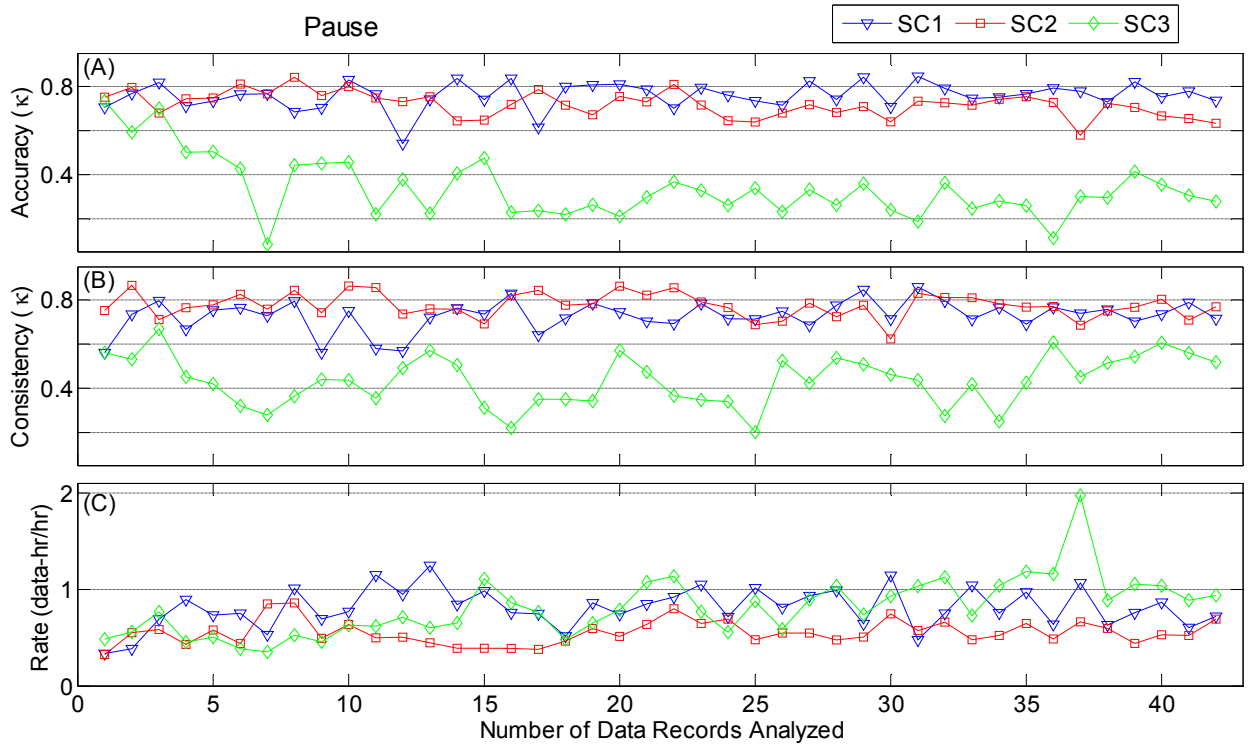

Fig. 4.14. Evaluation of manual scoring of Pause. (A) Accuracy (Fleiss' κ); (B) consistency (Fleiss' κ); and (C) rate (hours of data per hour of scoring) as a function of number of data records analyzed. Results are shown for the 42 data records analyzed (21 files scored twice).

The 3 scorers had similar accuracy and consistency for SYB, ASB, and SIH (Figs. 4.A3, 4.A4, and 4.A5).

4.6.3. Scoring Rate

Fig. 4.13C demonstrates some significant differences in scoring rate among the scores. All three scorers began scoring at a rate of 1 data-hr/hr, but SC1 and SC3 gradually increased the scoring rate by two- to three-fold throughout the study. In contrast, SC2 maintained a constant rate throughout. Analysis of the pattern-specific rates showed that the increase in scoring rate was primarily associated with SYB (Fig. 4.A3), and MVT (Fig. 4.A1), while scoring rates for ASB (Fig. 4.A4), SIH (Fig. 4.A5), PAU (Fig. 4.14), and UNK (Fig. 4.A2) were fairly constant throughout.

4.6.4. Repeatability

Each scorer analyzed the pre-processed validation subset in two independent, randomized instances. Intra-scorer repeatability was assessed by comparing the RIP patterns each scorer assigned to the same data in the two instances. Table 4.5 shows that the overall intra-scorer repeatability was very good; the scorer who participated in data acquisition SC1 had the highest repeatability ($\kappa = 0.84$), followed by the sleep laboratory technician SC2 ($\kappa = 0.77$), and the non-clinical scorer SC3 ($\kappa = 0.72$). The pattern with the highest intra-scorer repeatability was SYB ($0.84 \le \kappa \le 0.89$), and the pattern with the lowest intra-scorer repeatability was UNK ($0.49 \le \kappa \le 0.56$).

Inter-scorer repeatability was computed for each of the 8 unique analysis combinations (each combination comprised one analysis iteration from each of the 3 scorers, and each scorer performed 2 iterations). Table 4.6 reports the result as mean ± standard deviation. The overall inter-scorer repeatability was $\kappa = 0.65$. The RIP pattern with most repeatability was SYB ($\kappa = 0.81$), and the repeatability on PAU was substantial ($\kappa = 0.65$).

Scorer	Overall	SYB	ASB	SIH	PAU	MVT	UNK
SC1	0.84	0.89	0.78	0.73	0.79	0.88	0.49
SC2	0.77	0.86	0.79	0.58	0.78	0.76	0.56
SC3	0.72	0.84	0.70	0.67	0.74	0.64	0.53

Table 4.5. Intra-scorer repeatability. Repeatability was measured using the Fleiss' κ statistic [133]. Standard deviation was < 0.01 in all cases. SYB = synchronous-breathing, ASB = asynchronous-breathing, SIH = sigh, PAU = respiratory pause, MVT = movement artifact, UNK = unknown.

Overall	SYB	ASB	SIH	PAU	MVT	UNK
0.65 ± 0.02	0.81 ± 0.01	0.69 ± 0.01	0.53 ± 0.01	0.65 ± 0.02	0.58 ± 0.04	0.28 ± 0.03

Table 4.6. Inter-scorer repeatability of scorers SC1, SC2, and SC3. Repeatability was measured using the Fleiss' κ statistic [133]. Results are presented as mean ± standard deviation. SYB = synchronous-breathing, ASB = asynchronous-breathing, SIH = sigh, PAU = respiratory pause, MVT = movement artifact, UNK = unknown.

4.6.5. Confusion Analysis

Table 4.7 presents the proportion of samples assigned to each consensus RIP pattern in the validation dataset. There was a consensus for 90 % of the samples; with the most common pattern being SYB (65 %), and the least frequent being SIH (1 %). For completeness, we computed the pattern proportions for the remaining 10 % of samples with no consensus even though these data were not used in the confusion analysis. We found that the majority (60 %) of the non-consensus samples were scored as either UNK or MVT, and the rest were: SYB 22 %, ASB 8 %, SIH 3 %, and PAU 7 %. We later found that the proportion of samples without consensus pattern could be reduced to 5 % if all samples scored as MVT were to be re-assigned to UNK.

Fig. 4.15 shows the confusion matrix for the full data set (3 scorers combined for all segment lengths). It is evident that there was no systematic confusion of samples with consensus pattern of SYB, ASB, PAU, or SIH. A significant confusion was evident between UNK and MVT (Fig. 4.15F). The confusion matrices for the individual scorers showed similar results (see Figs 4.A6, 4.A7, and 4.A8).

Note that segment length had no effect on the confusion matrix for SC2 and SC3, but for SC1, confusion of PAU varied with segment length. Fig. 4.16 illustrates that SC1 confused PAU segments longer than 15 s with UNK, and this confusion increased with segment length.

4.7. Discussion

This Chapter describes a novel set of tools for the manual analysis of infant respiratory inductive plethysmography (RIP) data. The tool set includes 5 components:

(i) A set of clear, concise definitions of RIP patterns, and scoring rules based on uncalibrated RIP data. These definitions and rules make it possible to fully characterize an infant's respiratory behavior across extended periods of time, thus enabling the analysis of long data records required for the study of Postoperative Apnea (POA).

Consensus Pattern	Number of Samples	Proportion
SYB	12,877,448	0.65
ASB	859,835	0.04
SIH	145,352	0.01
PAU	632,694	0.03
MVT	2,606,271	0.13
UNK	810,583	0.04
None	2,017,540	0.10

Table 4.7. Proportion of consensus patterns for the confusion analysis. SYB = Synchronous-breathing, ASB = asynchronous-breathing, SIH = sigh, PAU = respiratory pause, MVT = movement artifact, UNK = unknown.

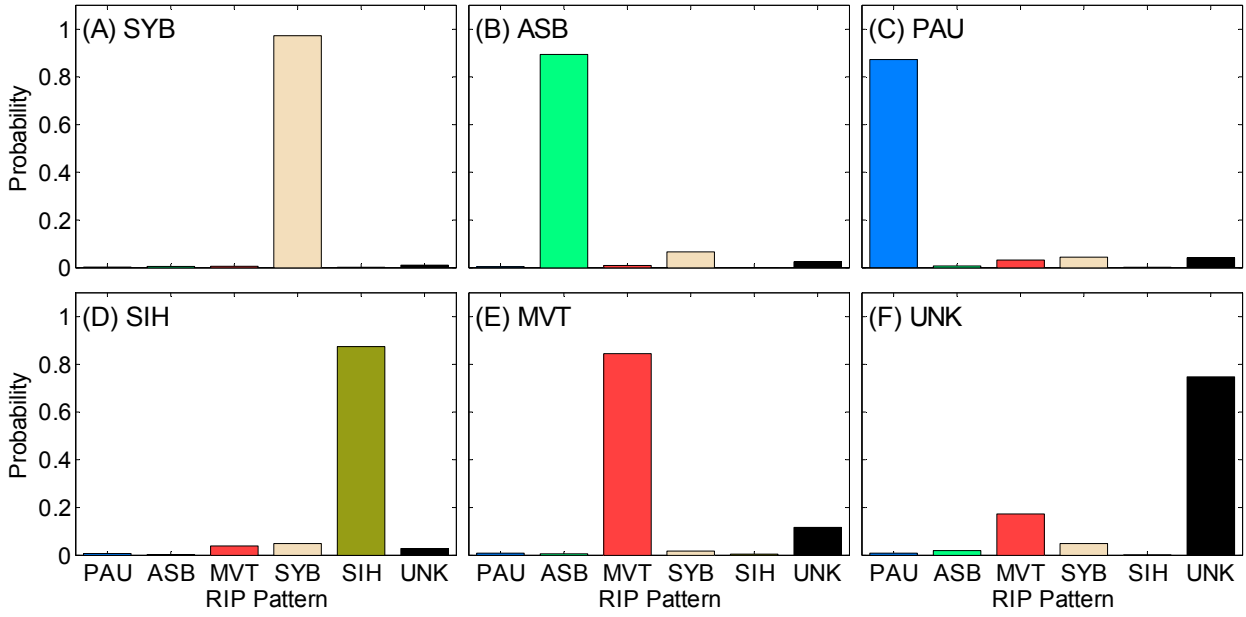

Fig. 4.15. Confusion matrix. Conditional probability of each respiratory inductive plethysmography (RIP) pattern for samples with the consensus pattern of: (A) synchronous-breathing (SYB), (B) asynchronous-breathing (ASB), (C) pause (PAU), (D) sigh (SIH), (E) movement artifact (MVT), and (F) unknown (UNK). When there is no confusion, the consensus pattern has a probability of 1 and the others have probabilities of 0. During total confusion all patterns have equal probabilities. Standard deviations of all probabilities were < 0.01.

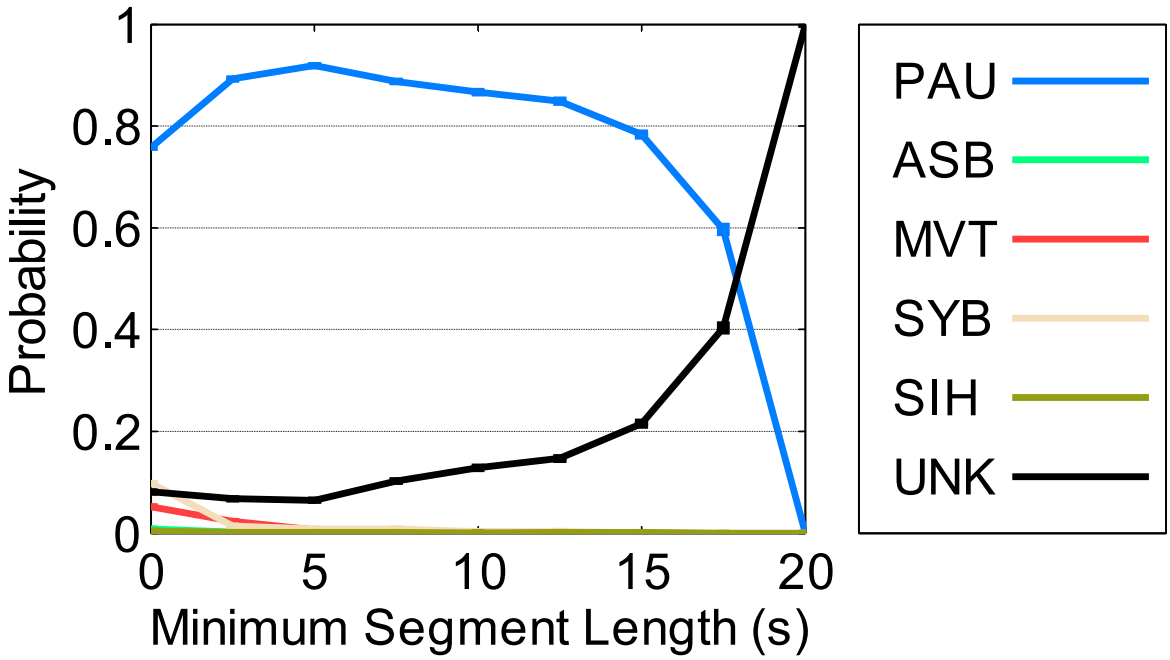

Fig. 4.16. Confusion of SC1 on samples with consensus pattern of pause as a function of segment length. SYB = synchronous-breathing, ASB = asynchronous-breathing, SIH = sigh, PAU = pause, MVT = movement artifact, UNK = unknown. A probability of 1 for PAU indicates no confusion. Lower PAU probabilities indicate increased confusion. Standard deviations of all probabilities were < 0.01.

(ii) An interactive, computer application (RIPScore) that supports the application of the scoring rules to infant data in an efficient manner. RIPScore incorporates the capability to track the rate at which scorers analyze data; providing the objective measurement of the time required to analyze a dataset.

(iii) A library of "true-pattern" segments representing each of the 6 RIP patterns, used for training, assessment of scorer performance, and development of evaluation methods.

(iv) A formal training protocol based on the interactive, completely automated RIPScore Training and Evaluation Modes. This protocol allows scorers from varied backgrounds to become proficient with RIPScore and the scoring protocol, and reach a standardized performance level similar to that of an expert. This training protocol obviates the requirement of certified sleep laboratory technicians, helping to reduce analysis costs, while increasing the feasibility of recruiting new scorers.

(v) A method to monitor the ongoing performance of scorers over time. This quality control measure allows the monitoring of scorers throughout the study to ensure they maintain a standardized performance. An advantage of this method is the early identification of underperforming scorers, which might allow for corrective action to assure the analysis quality.

The validation experiment demonstrates that analysis with these tools is accurate, efficient, and has high intra- and inter- scorer repeatability. These characteristics make our tools appropriate for studying respiratory conditions where large datasets (e.g., POA), and multiple scorers (e.g., longitudinal, multicenter trials) are a necessity.

4.7.1. Comparison to Existing Manual Scoring Tools

Commercially available scoring software is designed to analyze data based on the AASM scoring rules [8]. Using this software, scorers analyze data records and detect clinically relevant respiratory events such as central, obstructive, and mixed apnea. This analysis does not provide a comprehensive description of respiratory behavior as a function of time, because it focuses only on detecting and scoring isolated segments of data. As a result, the AASM analysis ignores potentially informative data segments. For example short respiratory pauses are not considered,

even though they are more frequent in infants with POA than in controls [82]. Additionally, the AASM rules require scorers to scroll throughout long records and visually detect candidate events. This strategy is prone to fatigue, leading to missed detections and increased variability.

In contrast, analysis with RIPScore requires that signals are analyzed continuously, on a sample-by-sample basis. An advantage of this continuous analysis is that the complete data record is classified. As a result, the instantaneous respiratory pattern is fully characterized as a function of time, enabling a comprehensive signals and systems analysis approach to the study of disorders of respiration such as POA. Additionally, the focus of scorers is changed from visual detection of events to classification of data segments. This design requires scorers to analyze all data segments and so it is not possible to miss events. Moreover, contrary to the AASM rules, our tools impose no arbitrary segment length definitions that may exclude short but relevant segments [82].

4.7.2. Training of Scorers

RIPScore provides an interactive Training Mode that familiarizes trainees with the interface, provides practice in scoring with immediate feedback using simulated data, and evaluates their performance. Three scorers with very varied backgrounds were trained in this way. All trainees reached the desired performance after four 2-hour training/evaluation sessions. Thus, by the end of training, all 3 scorers regardless of their clinical expertise, reached a standardized performance similar to that of the experienced reference scorer (REF). This implies that for large projects requiring multiple scorers, it should be possible to efficiently train a cadre of naive scorers to have performance similar to that of an expert.

4.7.3. Accuracy and Consistency

The scorers used our tools to carry out a comprehensive manual analysis of the pre-processed validation dataset, comprising 21 infant data records that incorporated quality control segments with known “true-patterns”; a total of 125 hours of data were manually analyzed twice per scorer. The ongoing accuracy and consistency of each scorer was assessed by analyzing the RIP patterns assigned to the quality control “true-pattern” segments. All scorers maintained a high,

relatively constant overall accuracy throughout the analysis of the 42 data records. The consistency of the two scorers with clinical expertise (SC1 and SC2) was nearly constant throughout, while the consistency of the third, non-clinical scorer (SC3) quickly rose to a level similar to that of the other two scorers after 10 data records. The high, nearly constant values of overall accuracy and consistency are evidence that the training protocol was effective, since scorers were able to achieve and maintain the desired performance level throughout.

It is noteworthy that for the PAU pattern, SC3 had lower accuracy and consistency for most of the data records, suggesting that a minimum clinical expertise with infant respiratory patterns may be necessary to maintain the desired performance. Figs. 4.14A and 4.14B suggest that even though the PAU-specific performance of SC3 was lower than expected, the initial 3 values of accuracy and consistency were likely influenced by training since they matched the values of SC1 and SC2. It was until after the third record that the performance of SC3 dropped. It is possible that an intervention at this point might have mitigated deterioration in PAU-specific performance.

4.7.4. Scoring Rate

We measured the rate at which scorers analyzed infant data throughout the study. Scoring was efficient, occurring at a rate of at least 1 hr of data analyzed in 1 hr. Scorers with no previous scoring experience gradually increased their rate, with no loss of either accuracy or consistency. In contrast, the sleep laboratory technician (SC2) maintained a constant rate. We believe that the design of the RIPScore Scoring Mode interface, which only required a single cursor selection and one key stroke to score a segment, facilitated this efficient analysis rate.

4.7.5. Repeatability of the Manual Analysis

The repeatability analysis showed that the two scorers with clinical background had very good intra-scorer repeatability, similar to that of REF. The scorer with no clinical expertise had a slightly lower intra-repeatability but it was still substantial.

The inter-scorer repeatability was very good in most categories. Indeed, the overall inter-scorer repeatability was much higher ($\kappa = 0.65$) than that reported between expert scorers from sleep laboratories using conventional scoring tools ($\kappa = 0.31$) [9]. For the particular pattern of PAU, intra- ($0.74 \leq \kappa \leq 0.79$) and inter-scorer ($\kappa = 0.65$) repeatability were substantial, which is relevant for the study of apnea. UNK was the pattern with lowest repeatability. Intra- and inter-scorer repeatability were also low for SIH, the only pattern requiring a breath-by-breath manual analysis.

4.7.6. Confusion of Patterns

Analysis of the confusion among RIP patterns found that SYB, ASB, SIH, and PAU were not often confused with other patterns. MVT and UNK were frequently confused with each other. This was the main reason for the low repeatability of UNK. This was expected since UNK grouped ambiguous patterns and segments of low signal quality. Even though this was a misclassification, both MVT and UNK correspond to corrupted data segments meant to be excluded from further analyses.

Additionally, we evaluated the influence of segment length on confusion, and found that segment length was a factor for only one scorer (SC1), who confused PAU segments longer than 15 s with UNK. A possible explanation is that SC1 might have interpreted long periods without respiratory movements as missing data resulting from technical problems, rather than as long PAU segments.

4.7.7. Implementation and Availability

RIPScore was implemented in MATLAB (The MathWorks Inc., Natick, MA, USA), compiled as a standalone application, and installed on the scorers' personal computers for the validation study. RIPScore and the pre-processing algorithm have been made available as open source, free of charge software; the manual and complete function repository are in GitHub (www.github.com/McCRIBS). The standalone application is available from the authors upon request.

4.7.8. Future Work

A difference between the manual scoring tools presented in this work and the AASM methodology is that respiratory behavior is classified in terms of 6 mutually exclusive patterns, instead of the occurrence of respiratory events such as apnea. At present, no direct link has been established between the 6 patterns and respiratory events. However, the patterns could be post-processed to identify respiratory events. For instance, a PAU with duration longer than a threshold (e.g., 15 s) would define a central apnea. Similarly, a combination of PAU with ASB would define a mixed obstructive apnea. Future work is necessary to evaluate the utility of a secondary set of rules based on pattern post-processing for the identification of clinically relevant respiratory events.

A direct application of the tools presented in this Chapter is the study of POA, and its relation to postoperative respiratory patterns. There is a variety of evidence suggesting that infants who experience POA have abnormal postoperative respiratory patterns [14, 24, 82]. Based on this, one could hypothesize that postoperative respiratory patterns may have information that is predictive of POA. The manual scoring tools from this Chapter could be used to investigate this hypothesis because they provide the means needed to comprehensively describe the respiratory patterns. Thus, for example, it would be straightforward to extract features from the manual scoring results related to information of the respiratory patterns such as the frequency of pauses, the proportion of time spent in each pattern, the relative proportion of synchronous- versus asynchronous-breathing, or the temporal sequence of patterns. Future work will investigate these and other features extracted from the respiratory patterns, and their ability to predict POA.

4.7.9. Significance

The tools for manual scoring introduced in this Chapter provide a comprehensive framework for the analysis of infant RIP data. These tools offer a significant advance in the study of respiratory behavior by providing: a comprehensive analysis method for large data sets, a means for the training and standardization of scorers, a method for the ongoing monitoring of scorer consistency and accuracy, and open source access to software and data sets.

Comprehensive Analysis: The tools provide a clear, concise definition of RIP patterns, and a software application (RIPScore) to locate these patterns along data records. The analyzed data record represents a sample-by-sample characterization of respiratory behavior as a continuous time series of patterns. All data points are classified and thereby significant segments are not missed. This approach facilitates the study of respiratory behavior from a signals and systems perspective by enabling the study of the temporal correlation between POA and the varied respiratory patterns (e.g., the relation between pause frequency and POA). The development of models that predict POA occurrence becomes possible, and preemptive interventions to enable preventive actions may follow.

Training & Standardization: The tools can be used to train any person to be a scorer, regardless of background, to achieve a standardized performance level similar to that of an expert. The ability to quickly train new scorers recruited from varied backgrounds increases the availability of potential scorers, thus helping to reduce the analysis cost by obviating the need for certified sleep laboratory technicians.

Monitoring of Scorer Performance: Another major contribution of this work is that the manual scoring tools make possible multicenter and longitudinal studies requiring multiple scorers. Conventional scoring tools have heretofore limited these types of study because of a low intra- and inter- scorer repeatability [9]. Intra-scorer repeatability is important to ensure that scorers maintain consistency throughout the period of data analysis. Inter-scorer repeatability is necessary to maintain the consistency of results among multiple scorers. The quality control method introduced in this work evaluates the ongoing scorer performance on a record-by-record basis. This quality control tool can identify underperforming scorers at any time throughout the duration of the study. This timely identification enables investigators to take corrective actions (e.g., additional training, scorer replacement) to maintain the desired performance. This ability will in turn help to reduce intra- and inter-scorer variability.

Open Source Access: Importantly, all the tools presented in this work are openly available to researchers interested in the analysis of respiratory patterns using RIP, and the study of POA. In addition to the RIP pattern definitions, scoring rules, representative examples, and training

protocol described in this manuscript; the software, including RIPScore and the pre-processing method for quality control, are freely available (www.github.com/McCRIBS). Finally, the library of "true-pattern" data segments, the complete dataset from infants at risk of POA, the training sessions, and analysis results from the 4 scorers are available from the Dryad Digital Repository (doi:10.5061/dryad.72dk5).

4.8. Conclusion

The tools presented in this work provide an excellent framework for study of infant respiratory behavior because they: (i) classify all respiratory patterns as a time series, (ii) standardize scorer performance using a training protocol which employs simulated data, (iii) monitor scoring repeatability by providing an ongoing quality control supervision of scorers, and (iv) are openly available and can be readily used in any study involving RIP.

4.9. Supporting Information

This section presents the additional figures published as supporting information in [10].

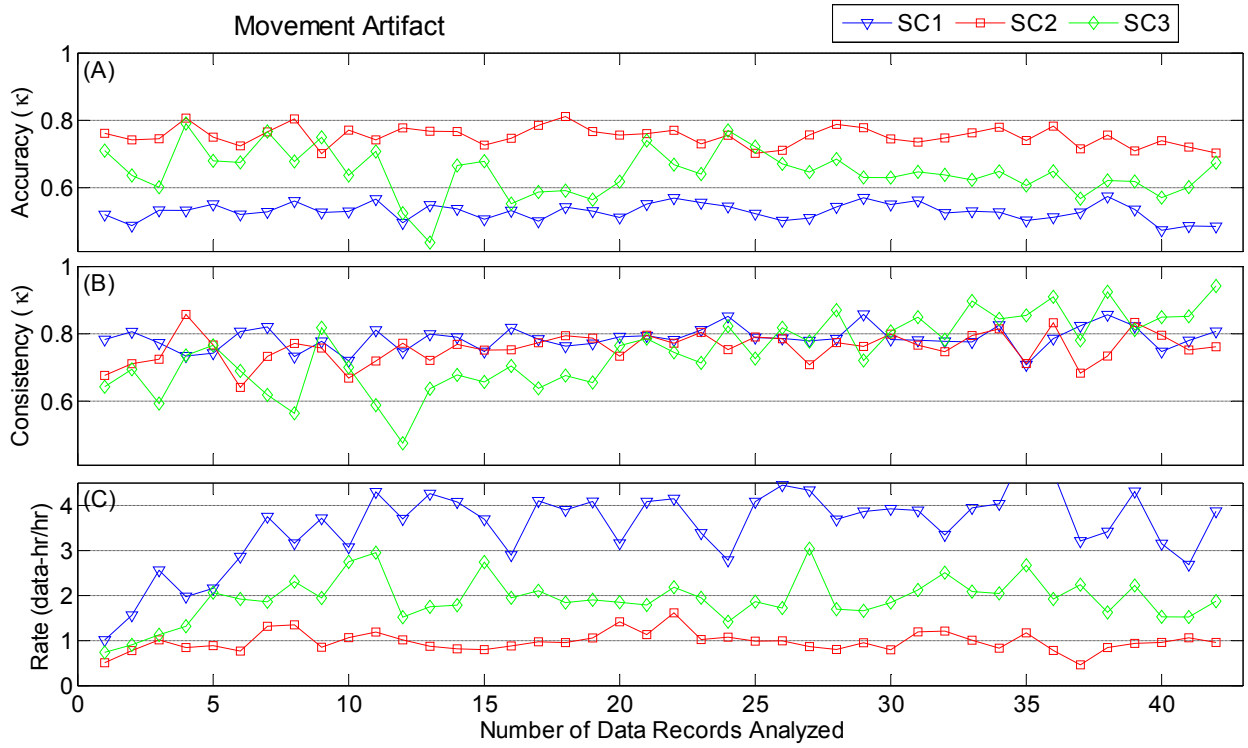

Fig. 4.A1. Evaluation of manual scoring of Movement Artifact. (A) Accuracy (Fleiss' κ); (B) consistency (Fleiss' κ); and (C) rate (hours of data per hour of scoring). Results are shown for the 42 data records analyzed (21 files scored twice).

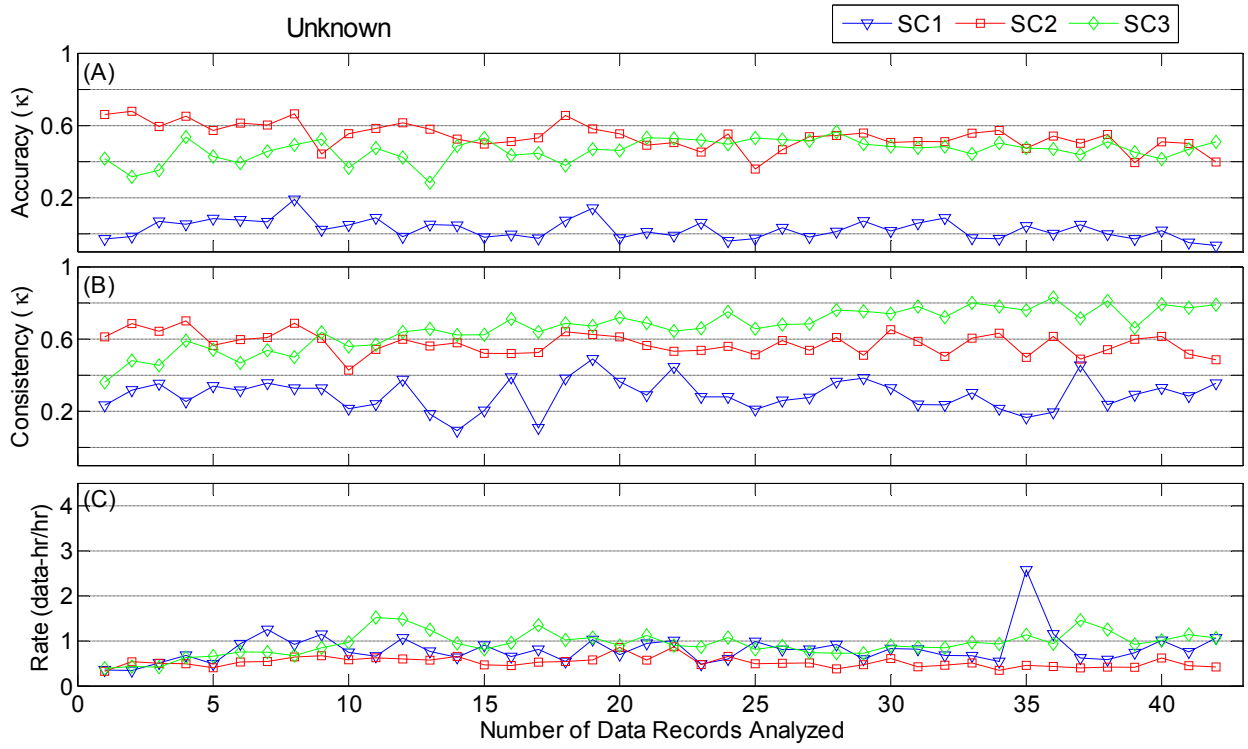

Fig. 4.A2. Evaluation of manual scoring of Unknown. (A) Accuracy (Fleiss' κ); (B) consistency (Fleiss' κ); and (C) rate (hours of data per hour of scoring). Results are shown for the 42 data records analyzed (21 files scored twice).

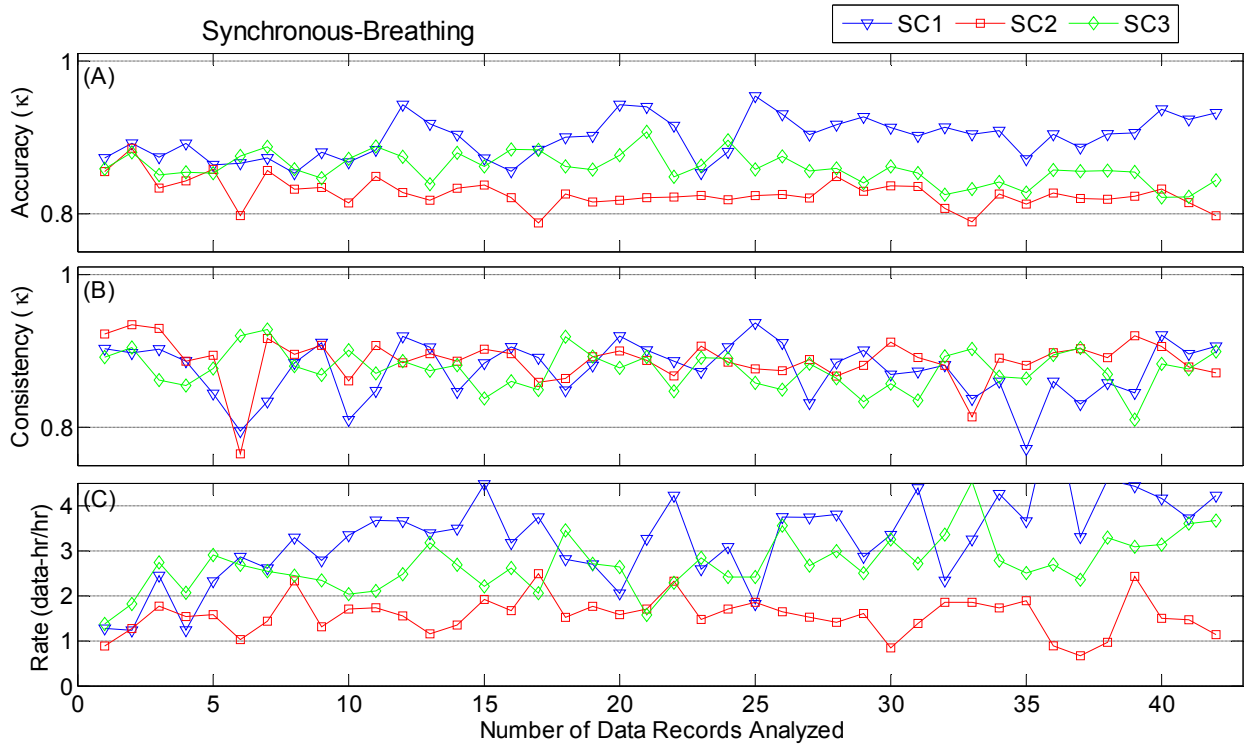

Fig. 4.A3. Evaluation of manual scoring of Synchronous-Breathing. (A) Accuracy (Fleiss' κ); (B) consistency (Fleiss' κ); and (C) rate (hours of data per hour of scoring). Results are shown for the 42 data records analyzed (21 files scored twice).

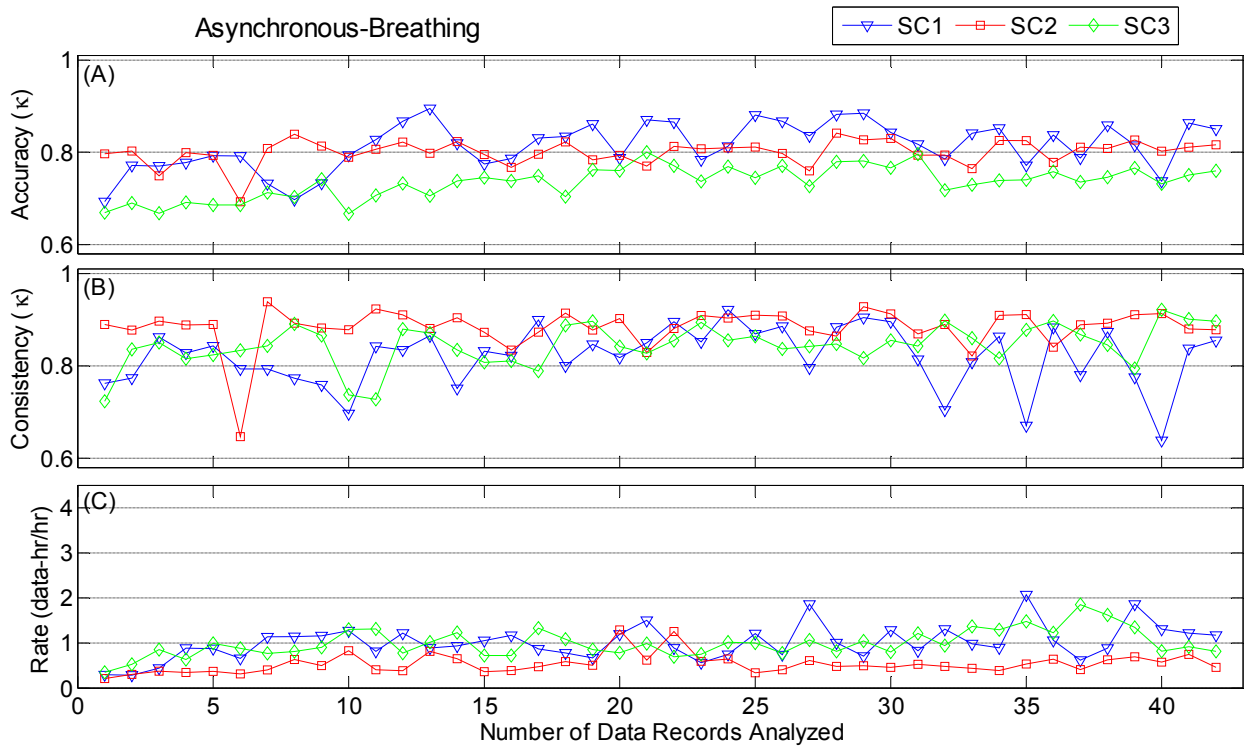

Fig. 4.A4. Evaluation of manual scoring of Asynchronous-Breathing. (A) Accuracy (Fleiss' κ); (B) consistency (Fleiss' κ); and (C) rate (hours of data per hour of scoring). Results are shown for the 42 data records analyzed (21 files scored twice).

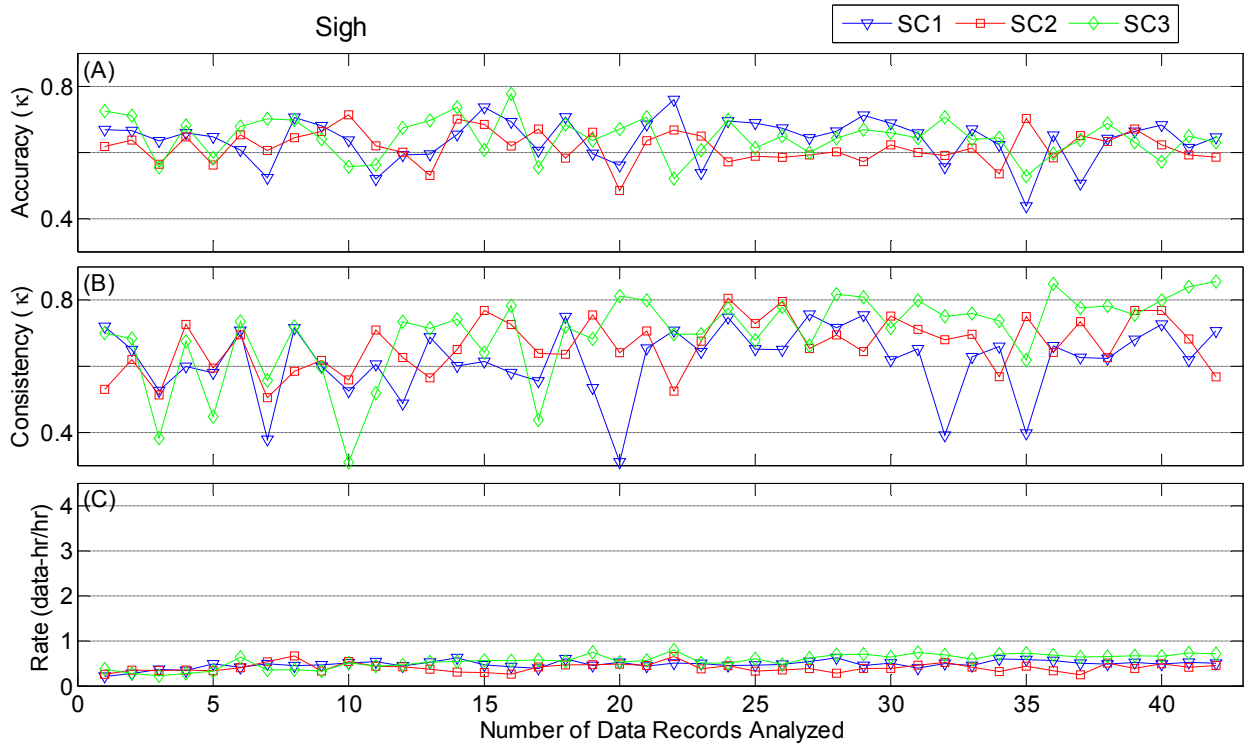

Fig. 4.A5. Evaluation of manual scoring of Sigh. (A) Accuracy (Fleiss' κ); (B) consistency (Fleiss' κ); and (C) rate (hours of data per hour of scoring). Results are shown for the 42 data records analyzed (21 files scored twice).

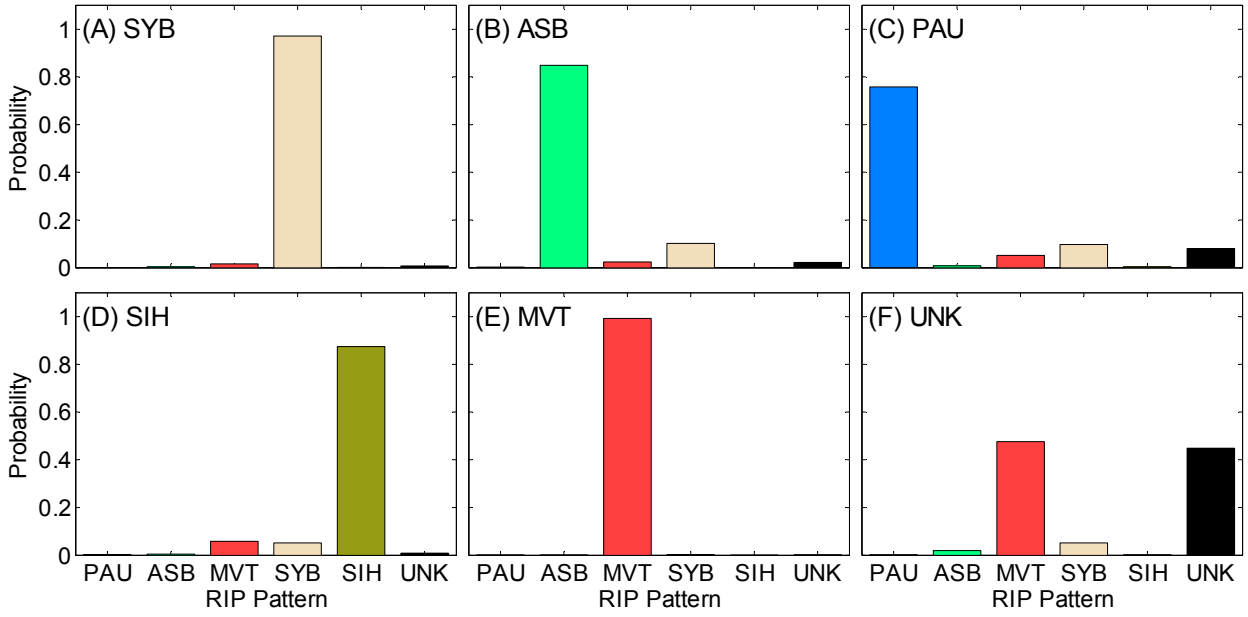

Fig. 4.A6. Individual confusion matrix of scorer SC1. Conditional probability of each respiratory inductive plethysmography (RIP) pattern for samples with the consensus pattern of: (A) synchronous-breathing (SYB), (B) asynchronous-breathing (ASB), (C) pause (PAU), (D) sigh (SIH), (E) movement artifact (MVT), and (F) unknown (UNK). When there is no confusion, the consensus pattern has a probability of 1 and the others have probabilities of 0. During total confusion all patterns have equal probabilities. Standard deviations of all probabilities were < 0.01.

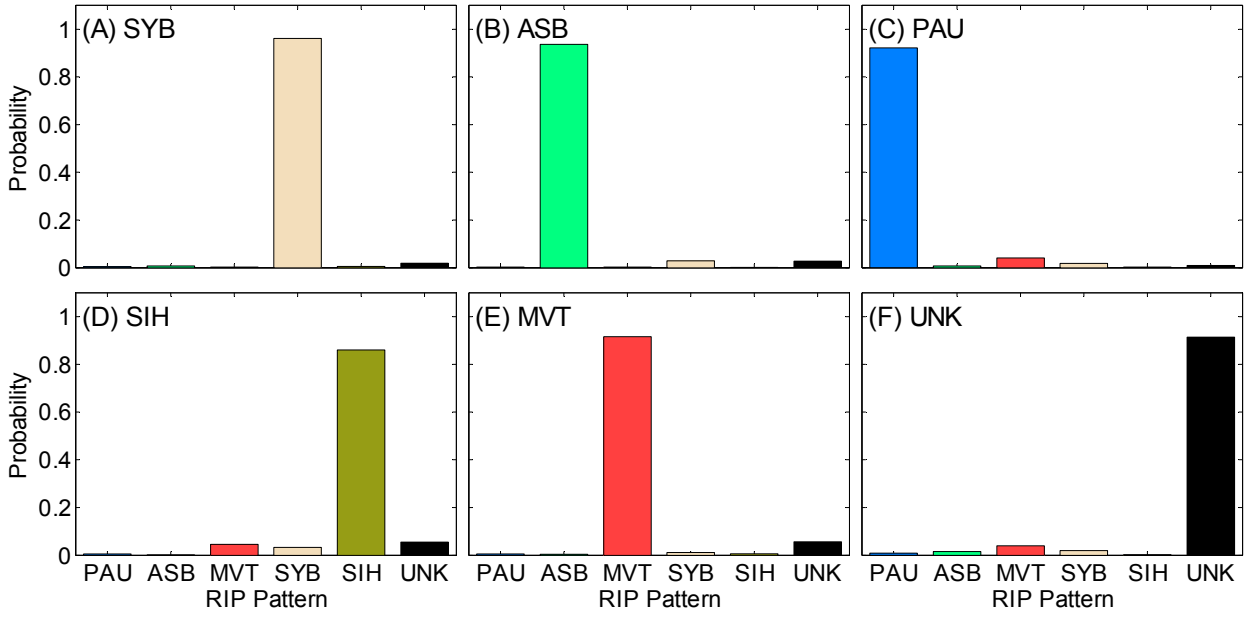

Fig. 4.A7. Individual confusion matrix of scorer SC2. Conditional probability of each respiratory inductive plethysmography (RIP) pattern for samples with the consensus pattern of: (A) synchronous-breathing (SYB), (B) asynchronous-breathing (ASB), (C) pause (PAU), (D) sigh (SIH), (E) movement artifact (MVT), and (F) unknown (UNK). When there is no confusion, the consensus pattern has a probability of 1 and the others have probabilities of 0. During total confusion all patterns have equal probabilities. Standard deviations of all probabilities were < 0.01.

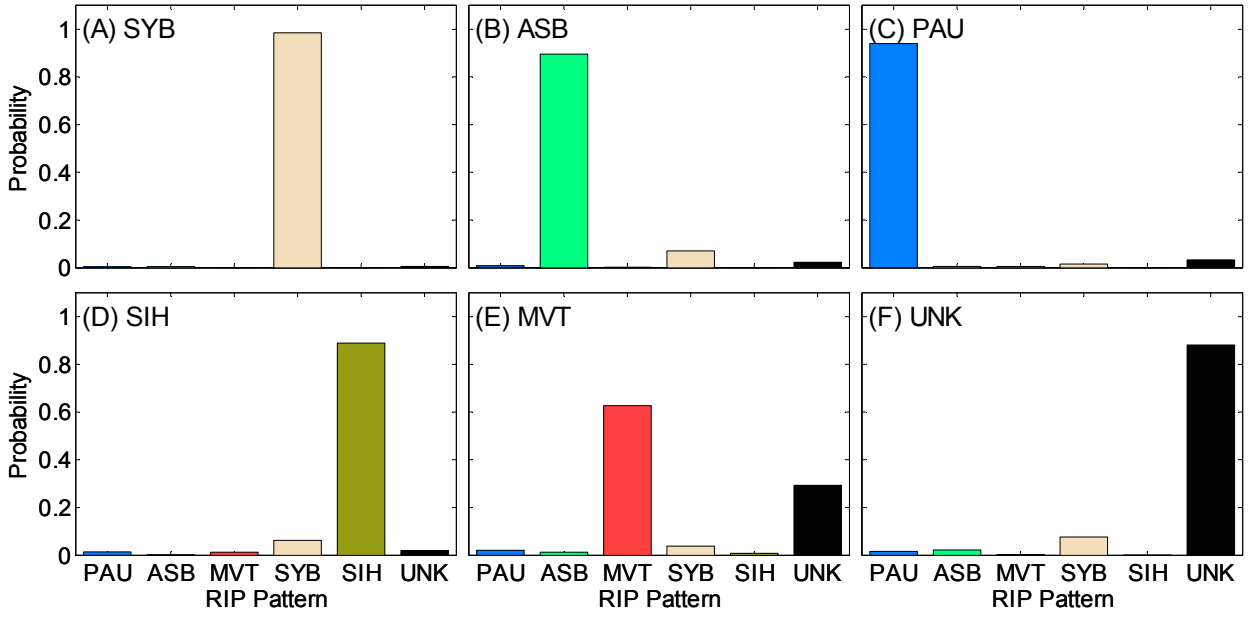

Fig. 4.A8. Individual confusion matrix of scorer SC3. Conditional probability of each respiratory inductive plethysmography (RIP) pattern for samples with the consensus pattern of: (A) synchronous-breathing (SYB), (B) asynchronous-breathing (ASB), (C) pause (PAU), (D) sigh (SIH), (E) movement artifact (MVT), and (F) unknown (UNK). When there is no confusion, the consensus pattern has a probability of 1 and the others have probabilities of 0. During total confusion all patterns have equal probabilities. Standard deviations of all probabilities were < 0.01.

5. Improving Manual Scoring of Respiratory Patterns using Expectation-Maximization

5.1. Preface

In the previous Chapter I presented a set of tools for manual scoring of infant respiratory data that yield high intra- and inter-scorer repeatability. However, the results from any one scorer will be subjective and will have some inherent variability. One approach to reduce this variability might be to have multiple scorers analyze the data, and combine the individual scores in a meaningful fashion. In this Chapter I explored this hypothesis, and developed a procedure based on Expectation-Maximization (EM) to optimally combine multiple analyses of the respiratory pattern. I demonstrated that the EM estimator performs significantly better than individual scorers, and the more traditional majority vote approach. The results constitute a better estimate of a "gold standard", since respiratory pattern estimates are highly accurate, very consistent, and much less subjective than those from individual scorers.

This Chapter is a manuscript that I will submit for publication:

C. A. Robles-Rubio, K. A. Brown, and R. E. Kearney, "Improving Manual Scoring of Respiratory Patterns using Expectation-Maximization," to be submitted to *IEEE Trans Biomed Eng*.

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada. The work of C. A. Robles-Rubio was supported in part by the Mexican National Council for Science and Technology. C. A. Robles-Rubio and K. A. Brown were supported in part by the Queen Elizabeth Hospital of Montreal Foundation Chair in Pediatric Anesthesia.

5.2. Abstract

Conventional manual scoring (CMS) is the preferred method to analyze respiratory data. However, it is limited by low intra- and inter-scorer repeatability. We recently developed a set of tools to assist manual scoring, which improved the repeatability of the analysis. However, its results were still limited by the inherent subjectivity and variability of any one scorer. To mitigate this, it is possible to have multiple scorers analyze the same respiratory data repeatedly, and combine the results in some fashion. A common approach is to use a simple majority vote (MV). However, MV has two important limitations: (i) the majority may be determined by less than 50 % of the votes in situations where most of the scorers disagree; and (ii) all scorers are assumed to have similar performance. This paper presents a method that addresses these limitations to obtain a better estimate of the most likely respiratory patterns. This is accomplished using an expectation-maximization (EM) method to combine results from multiple, manual scorers weighted by their individual performance.

The accuracy of the EM method was compared to those of individual scorers (IS), and the MV approach in a study that simulated the performance of real, manual scorers. The simulation evaluated the accuracy with which EM and MV estimated the true respiratory patterns from analyses by multiple, simulated scoring sequences. The accuracy of both methods improved with the number of scoring sequences, but EM was significantly better than IS and MV. In fact, with only 5 scoring sequences EM had the same accuracy and less variability than did MV with 25 scoring sequences.

We then applied the EM method to the results of the manual analysis of 21 data records from infants. These data contained quality control segments with known, “true” patterns that were used to assess the accuracy and consistency of the EM, MV, and IS estimators. Each record had two copies of these quality control segments, one in each half of the recording. For each record, accuracy was measured as the agreement between the estimator and the “true” pattern, and consistency was measured as the agreement between the estimator and itself in the two copies of the quality control segments. EM estimates were significantly more accurate than those from

either IS or MV. EM and MV had similar consistency, which was significantly higher than that of IS.

The excellent accuracy and consistency of EM means that it can be used to produce objective, "gold standard" analyses of respiratory behavior that dramatically reduce the effects of intra- and inter-scorer variability. This "gold standard" analysis opens the door for large multi-institutional or longitudinal studies, by making it possible to obtain reliable results with a reasonable number of scorers.

5.3. Introduction

Infants recovering from surgery and anesthesia are at risk of life-threatening Postoperative Apnea (POA) [1-3]. The majority of these POA events occur within 2 h after surgery, but the onset of apnea may be delayed up to 12 h postoperatively [14]. To date there is no way to predict the risk of POA in a specific infant, and so all infants with postmenstrual age (PMA) ≤ 60 weeks are hospitalized for monitoring for a minimum of 12 hours after surgery [14]. It has been hypothesized that characteristics of the respiratory patterns can predict the risk of POA [7, 24]; thus, investigators have analyzed the respiratory behavior of these infants in an attempt to identify possible predictors. To this end, cardiorespiratory signals were acquired using respiratory inductive plethysmography (RIP), including ribcage (RCG) and abdomen (ABD) respiratory movements, as well as pulse oximetry, including photoplethysmography (PPG) and blood oxygen saturation (SAT) [2, 7, 85, 135, 139]. These signals were then analyzed to determine the occurrence of POA and other cardiorespiratory events.

The preferred method to analyze these cardiorespiratory signals has been conventional manual scoring (CMS), with a focus on the visual detection of respiratory events based on a set of rules defined by the American Academy of Sleep Medicine (AASM) [8] . We recently developed a set of tools to assist manual scoring [10] that includes a set of definitions for 6 unique, mutually exclusive patterns that fully describe the RIP signals; a library of representative data segments;

and a quality assurance component to monitor scorer performance that involves pre-processing of the data by inserting segments with known "true" patterns (see Table 5.1 for definition of the term "true" pattern). We showed that the use of these tools improves intra- and inter-scorer repeatability.

Nonetheless, any manual scoring method will be limited by the inherent subjectivity of human scorers. Employing multiple scorers may help eliminate individual bias but will add additional variability. Fig. 5.1 illustrates this by showing the patterns assigned to the same data segment by three different scorers. The pattern sequences are generally similar but there is both intra-scorer variability, evident in the pairs of scores assigned by the same scorer, and inter-scorer variability, evident in the results from the different scores. For the samples assigned the same pattern by all scorers (e.g., gray-shaded segments on Fig. 5.1), it is reasonable to assume that the assigned pattern is the most likely pattern (see Table 5.1). However, for the majority of the record there was some disagreement among scorers (e.g., non-shaded segments on Fig. 5.1) and the most likely pattern is not evident. This raises the question of how to combine the results from multiple scorers to estimate the most likely respiratory pattern.

A simple approach would be to use the majority vote (MV), where samples are assigned the pattern receiving the most votes [10, 143]. This approach has two important limitations: (i) when there is much disagreement among scorers the majority may not be absolute, and so the final pattern would be determined by a minority of votes; and (ii) votes from all scorers are weighted equally regardless of their performance [144]. A more informed estimate should take into account the individual performance of each scorer and weight their votes accordingly. This can be achieved using Expectation-Maximization (EM) [145], where votes are weighted based on estimates of individual scorer performance [144, 146]. Indeed, this approach has been used to evaluate the performance of annotators of patient records [146], and also to evaluate the segmentation of medical images [147].

Term	Definition
Most likely pattern	It is not possible to determine the ground truth respiratory patterns in real, clinical data. In the absence of a ground truth, the most likely pattern is the one that has the highest probability of being correct.
"True" pattern	An estimate of the most likely pattern in real, clinical data, obtained from the analysis of a very experienced, expert scorer. For details see [10].
Assigned pattern	The pattern assigned by a manual scorer to a sample of data
Estimated pattern	The pattern estimated by any of the estimators described in this paper.
True pattern	The simulation ground truth

Table 5.1. Definition of terms used by this paper to describe respiratory patterns.

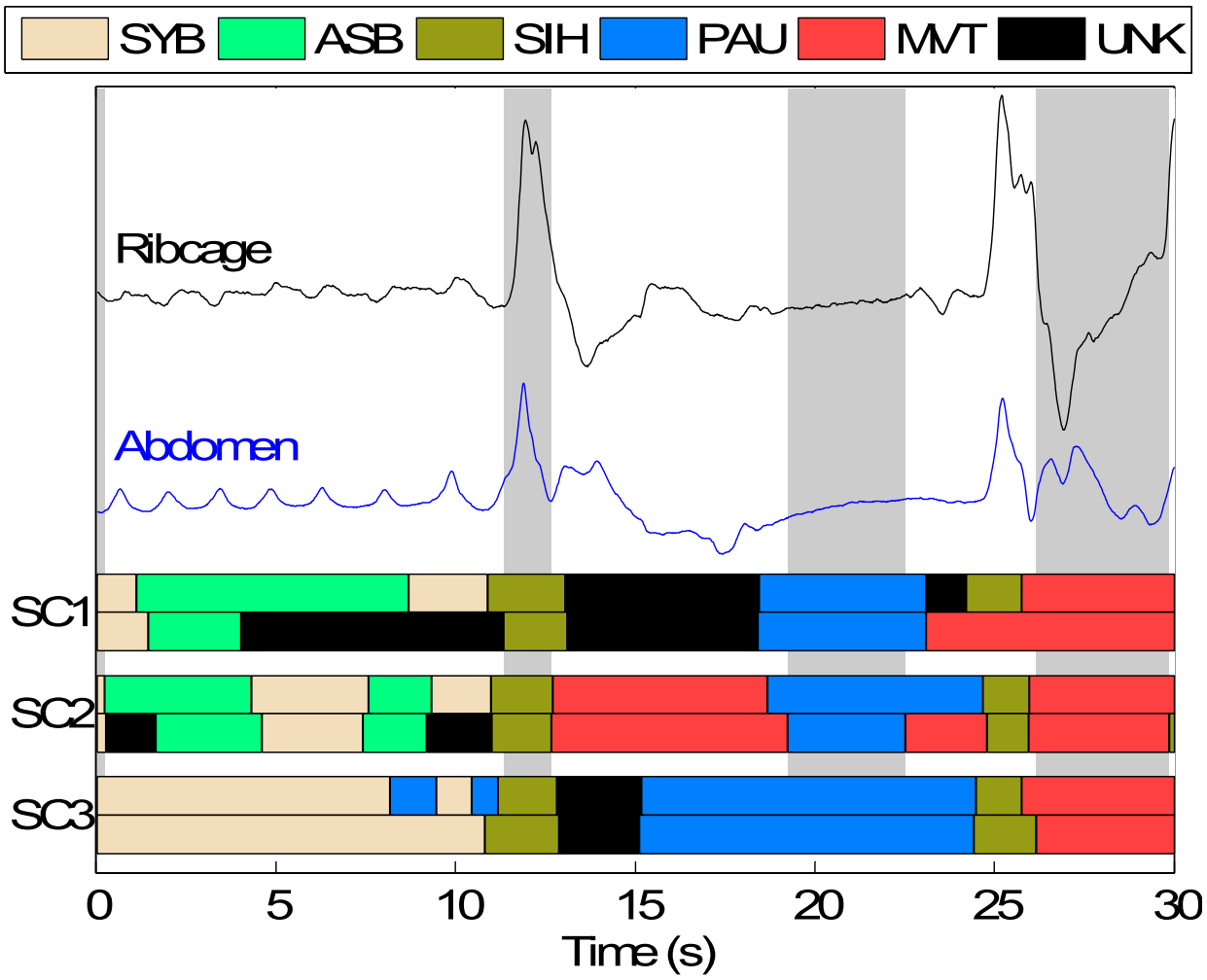

Fig. 5.1. Results of the manual analysis of a typical of respiratory data performed by three scorers (SC1, SC2, and SC3), twice each. Gray-shading indicates the same respiratory pattern was assigned to the segment in all 6 scoring sequences, representing the most likely pattern. Ribcage and abdomen are respiratory inductive plethysmography measurements in arbitrary units. Assigned patterns are color coded as: SYB = Synchronous-breathing, ASB = asynchronous-breathing, SIH = sigh, PAU = respiratory pause, MVT = movement artifact, UNK = unknown.

This paper explores how to combine the patterns assigned by multiple scorers to estimate the most likely respiratory patterns. To this end, we evaluated the performance of three estimators: (i) individual scorers (IS), (ii) MV, and (iii) EM. The paper is developed as follows: Section 5.4 describes the MV and EM estimators; Section 5.5 describes the materials used for the evaluation; Section 5.6 describes the simulation experiment used to examine the accuracy of the three estimators; Section 5.7 demonstrates the application of the estimators to data from infants at risk of POA; Section 5.8 discusses the results; and Section 5.9 provides some concluding remarks.

5.4. Estimation of Most Likely Respiratory Patterns using Expectation-Maximization

This section describes an expectation-maximization (EM) method to estimate the most likely respiratory patterns from multiple scores. It is adapted from the method by Raykar et al. [144], which estimates the ground truth from a set of multiple noisy segmentations. Consider a set of N samples of respiratory inductive plethysmography (RIP) data acquired from the ribcage and abdomen of an infant, which were manually analyzed by R scorers to yield $S \geq R$ individual sequences of respiratory patterns [10]. These sequences classify samples into one of $C = 6$ unique, mutually exclusive RIP patterns: synchronous-breathing (SYB), asynchronous-breathing (ASB), sigh (SIH), movement artifact (MVT), respiratory pause (PAU), and unknown (UNK). The EM algorithm estimates $T[n]$, the most likely pattern of each sample as follows:

(i) Initialize $W_c[n]$, the probability that sample n has the pattern $c \in \{1, 2, \ldots, C\}$, as the proportion of sequences that assign that pattern to it:

$$W_c^0[n] = \frac{1}{S}\sum_{s=1}^{S} I\left(\hat{T}_{IS}^s[n], c\right), \tag{5.1}$$

where $\hat{T}_{IS}^s[n]$ is the pattern assigned by scoring sequence s to sample n, and

$$I\left(\hat{T}_{IS}^s[n], c\right) = \begin{cases} 1 & \text{if } \hat{T}_{IS}^s[n] = c \\ 0 & \text{otherwise.} \end{cases}$$

The pattern with the highest probability is the majority vote (MV) estimate

$$\hat{T}_{MV}[n] = \arg\max_{c} \left\{ W_c^0[n] \right\}. \tag{5.2}$$

This initialization defines the basic MV estimator and the pattern probabilities W_c^0 for the first iteration ($i = 1$) of the EM estimator.

(ii) Estimate the marginal probability of each pattern within the dataset as

$$P_c^i = \frac{1}{N} \sum_{n=1}^{N} W_c^{i-1}[n]. \tag{5.3}$$

(iii) Estimate the confusion matrix $\mathbf{Q}^{s,i}$ for each sequence s as

$$Q_{c',c}^{s,i} = \frac{1}{N P_c^i} \sum_{n=1}^{N} W_c^{i-1}[n] I\left(\hat{T}_{IS}^{s}[n], c' \right), \tag{5.4}$$

where the index $c = 1, 2, \ldots, C$ spans the most likely patterns, and $c' = 1, 2, \ldots, C$ those assigned in the sequence. Each element $Q_{c',c}^{s,i}$ is an estimate of the conditional probability that sequence s assigns the pattern c', to samples whose most likely pattern is c.

(iv) Refine the estimate $W_c[n]$, the probability that sample n has the most likely pattern c, by re-weighting the votes of each sequence based on its confusion matrix as

$$W_c^i[n] = \frac{W_c^{*,i}[n]}{\sum_{c'=1}^{C} W_{c'}^{*,i}[n]}, \tag{5.5}$$

where

$$W_c^{*,i}[n] = P_c^i \prod_{s=1}^{S} Q_{\hat{T}_{IS}^{s}[n],c}^{s,i}. \tag{5.6}$$

is proportional to the probability that sample n has the pattern c taking into account the patterns assigned by each sequence and their probability of being correct.

(v) Increment the index $i = i + 1$, and repeat from (ii) until changes in the estimates of the confusion and sample probability matrices are smaller than some convergence error ε. The EM algorithm has been demonstrated to converge to a local optimum [148].

(vi) Set the estimate of the most likely pattern of each sample to that with the highest probability

$$\hat{T}_{EM}[n] = \arg\max_{c} \left\{ W_c^{i-1}[n] \right\}. \tag{5.7}$$

5.5. Clinical Dataset and Manual Analysis

A portion of the materials presented in [10], which are openly available from the Dryad Digital Repository (doi:10.5061/dryad.72dk5) [11], were used to evaluate the estimation methods. This section describes this material briefly.

5.5.1. Infant Data

Data were acquired from 21 infants at risk of POA (16 male, birth age 31 ± 4 weeks, postmenstrual age 43 ± 2 weeks, weight 3.6 ± 1.0 kg) immediately after surgery in the postanesthesia care unit of the Montreal Children's Hospital. The study was approved by the Institutional Review Board of the McGill University Health Centre / Montreal Children's Hospital (approval numbers PED-07-30, and 12-308-PED). Written, informed parental consent was obtained for each infant recruited.

The signals acquired were ribcage (RCG) and abdomen (ABD) respiratory inductive plethysmography (RIP). Signals were low-pass filtered at 10 Hz, sampled at 50 Hz, and stored. No attempt was made to calibrate the signals. Record lengths for each infant varied from 3.9 h to 12 h; the total length of data was 191 hr.

5.5.2. Data Pre-processing

Each data record was pre-processed using the McGill CardioRespiratory Infant Behavior Software (McCRIBS) [12]. This pre-processing truncated each record to a maximum of 20,000 s, and inserted two copies of 152 data segments, with known "true" patterns, into each record at

random times, once in the 1st half of the record, and then again in the 2nd half. The same 152 "true" pattern segments were inserted into all data records and comprised: 25 SYB, 26 ASB, 27 SIH, 22 PAU, 27 MVT, and 25 UNK segments. Details of this pre-processing are described in [10]. These "true" pattern segments were used to design the simulation and to evaluate the accuracy and consistency of the algorithms.

5.5.3. Manual Data Analysis

Three trained scorers with varied backgrounds analyzed all 21 pre-processed data sets in two, independent, randomly ordered instances [10]. This yielded 6 pattern sequences describing the respiratory behavior for each data record.

5.6. Evaluation of Performance with Simulated Data

A simulation experiment was used to examine the estimation accuracy of individual scoring (IS), majority vote (MV), and expectation-maximization (EM).

5.6.1. Simulation Method

To do this we modeled the performance of the simulated scorers based on an estimate of $\mathbf{Q}^{real}$, the combined confusion matrix of the 3 manual scorers. The elements of this matrix, $Q_{c',c}^{real}$, were estimated from the "true" pattern segments as the proportion of samples with "true" pattern c that were assigned the pattern c'. Table 5.2 shows the values of $\mathbf{Q}^{real}$.

We generated one scoring sequence for each simulated scorer. Thus, each scoring sequence s was assigned a specific confusion matrix, termed $\mathbf{Q}^{s,sim}$, generated by perturbing $\mathbf{Q}^{real}$. The diagonal values of $\mathbf{Q}^{s,sim}$, i.e., $Q_{c,c}^{s,sim}$, representing the probabilities that the assigned patterns were correct, were sampled from a unimodal beta distribution with parameters α and β set as

$$\alpha = \begin{cases} 2 & \text{if } Q_{c,c}^{real} \le 0.5 \\ \dfrac{1}{1-Q_{c,c}^{real}} & \text{otherwise} \end{cases} \tag{5.8}$$

		"True" Pattern					
		SYB	*ASB*	*SIH*	*PAU*	*MVT*	*UNK*
Assigned Pattern	*SYB*	0.90	0.02	0.06	0.02	0.02	0.10
	ASB	0.06	0.81	0.01	0.00	0.01	0.03
	SIH	0.00	0.00	0.69	0.00	0.03	0.01
	PAU	0.01	0.01	0.00	0.58	0.02	0.01
	MVT	0.01	0.06	0.15	0.09	0.80	0.39
	UNK	0.02	0.10	0.08	0.30	0.12	0.46

Table 5.2. Values of $\mathbf{Q}^{real}$, the combined confusion matrix of 3 real, manual scorers. SYB = Synchronous-breathing, ASB = asynchronous-breathing, SIH = sigh, MVT = movement artifact, PAU = respiratory pause, and UNK = unknown.

and

$$\beta = \begin{cases} \dfrac{1}{Q_{c,c}^{real}} & if \;\; Q_{c,c}^{real} < 0.5 \\ 2 & otherwise \end{cases}, \tag{5.9}$$

such that the mode of the resulting distribution was equal to $Q_{c',c}^{real}$.

The off-diagonal elements of the simulated confusion matrix were assigned the probabilities

$$Q_{c',c}^{s,sim} = \frac{Q_{c',c}^{real}\left(1 - Q_{c,c}^{s,sim}\right)}{\left(1 - Q_{c,c}^{real}\right)}, \quad \forall c' \neq c \,. \tag{5.10}$$

Simulated scoring sequences were generated as follows:

(i) The samples of the 152 "true" pattern segments were pooled to form a collection of true pattern samples.

(ii) A realization of the true pattern vector $\mathbf{T}^{sim}$ of length $N = 10,000$ samples, was generated by randomly sampling, with replacement, samples from the collection.

(iii) A simulated scoring sequence $\hat{\mathbf{T}}_{IS}^{s,sim}$ was generated by assigning each sample a value sampled from the probability distribution $f\left[c' \mid T^{sim}[n]\right]$, i.e.,

$$\hat{T}_{IS}^{s,sim}[n] \sim f\left[c' \mid T^{sim}[n]\right], \tag{5.11}$$

where the $T^{sim}[n]$ was the true pattern of sample n, and $f\left[c' \mid T^{sim}[n]\right]$ was defined as

$$f\left[c' \mid T^{sim}[n]\right] = Q_{c',T^{sim}[n]}^{s,sim}, \qquad c' = 1,\ldots,6\,, \tag{5.12}$$

so that the probability of $\hat{T}_{IS}^{s,sim}[n]$ being set to pattern c' given true pattern $T^{sim}[n]$ was equal to $Q_{c',T^{sim}[n]}^{s,sim}$. This procedure was repeated for as many scoring sequences as needed.

Monte Carlo simulations using 10,000 realizations of $\mathbf{T}^{sim}$ were carried out to evaluate the accuracy of the estimators. For each realization, a total of S simulated scorers with one scoring sequence each were generated and the true patterns were estimated using the IS $\left(\hat{\mathbf{T}}_{IS}^{s,sim}\right)$, MV $\left(\hat{\mathbf{T}}_{MV}^{sim}\right)$, and EM $\left(\hat{\mathbf{T}}_{EM}^{sim}\right)$ methods. The accuracies of $\hat{\mathbf{T}}_{IS}^{s,sim}$, $\hat{\mathbf{T}}_{MV}^{sim}$, and $\hat{\mathbf{T}}_{EM}^{sim}$ were assessed as the agreement between the estimated and true patterns using the Fleiss' κ statistic [133] in each realization, resulting in one vector of 10,000 accuracy values per estimator. These vectors were summarized by their median and interquartile range (IQR).

The standard deviation of the median and IQR were then estimated using the bootstrap method [142]. Thus, a resampled accuracy vector for a given estimator was generated by sampling with replacement from the original accuracy vector, and its median and IQR were computed. The procedure was repeated 10,000 times to estimate the standard deviation of the median and IQR.

5.6.2. Simulation Results

The convergence of the EM algorithm was examined using the Monte Carlo setup described above, fixing the number of simulated scorers to 3 to yield $S = 3$ scoring sequences, and measuring accuracy as a function of the number of iterations. Fig. 5.2 shows that the change in the average accuracy of EM was positive for all iterations demonstrating that accuracy increased with the number of iterations. The change in accuracy became almost 0 by the 40th iteration.

Next, the effect of the number of scorers on the accuracy was evaluated by running another Monte Carlo simulation, this time by varying S from 3 to 25 and keeping the number of EM iterations constant at a value of 50. Fig. 5.3A shows the median accuracy of both EM and MV increased with the number of sequences, but the increase was both larger and more rapid for EM. An important remark is that using MV alone improved the performance of the individual scorers. Fig. 5.3B shows that the IQR of accuracy of all 3 estimators decreased as the number of sequences increased. EM had the lowest IQR for any number of sequences, while MV had the highest variability. EM reached near perfect accuracy by 25 scorers, having a median of 1 and an IQR approaching 0.

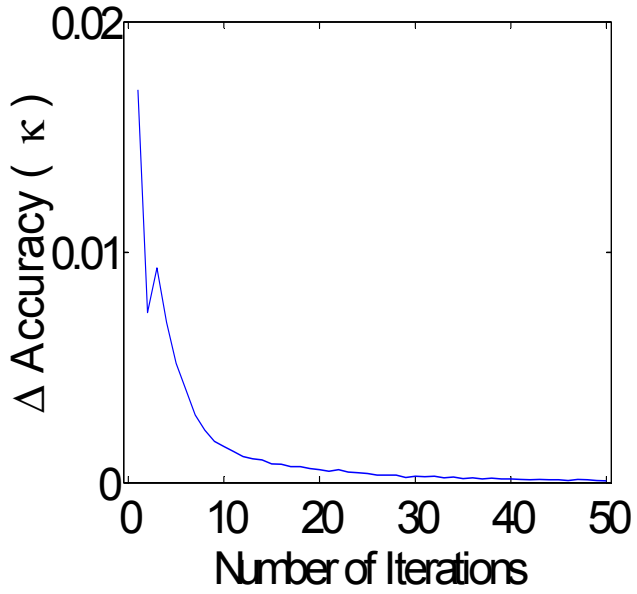

Fig. 5.2. Change in accuracy of Expectation-Maximization (EM) as a function of the number of iterations.

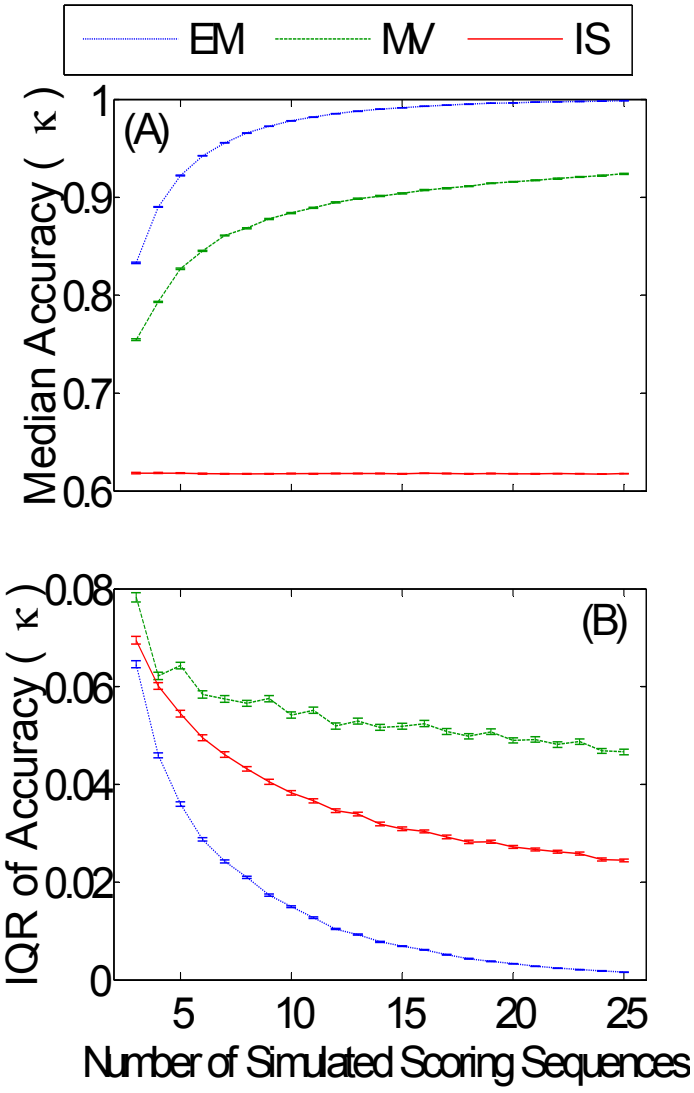

Fig. 5.3. Simulation of the effects of number of scorers (one scoring sequence per scorer). (A) Median and (B) Interquartile Range (IQR) of accuracy of Expectation-Maximization (EM), Majority Vote (MV), and Individual Scoring (IS) estimates as functions of the number of simulated scoring sequences.

5.7. Evaluation of Performance with Clinical Data

The performance of the estimators was also evaluated using the results from the manual analyses performed by the 3 scorers. The two copies of the 152 "true" pattern segments inserted into the 21 data records were used to evaluate the accuracy and consistency of individual scoring (IS), majority vote (MV), and expectation-maximization (EM). Accuracy was assessed as the agreement between the "true" patterns and the estimated patterns. Consistency was assessed as the agreement between the patterns estimated in the first and second copies of the "true" pattern segments. Overall and pattern-specific agreements were evaluated using the Fleiss' κ statistic [133]. The significance of differences in accuracy and consistency between the estimators was evaluated using the Wilcoxon rank sum test [141].

5.7.1. Method Convergence

First, we examined the number of iterations needed for the EM algorithm to converge. To do this we carried out 10 EM iterations, and evaluated the overall accuracy and consistency of $\hat{\mathbf{T}}_{EM}$ in each iteration. Fig. 5.4 shows both overall accuracy and consistency became almost constant by the 6th iteration. Accuracy changed most increasing from 0.74 to almost 0.8, while consistency remained almost constant.

5.7.2. Accuracy and Consistency

Next, we compared the accuracy and consistency of the IS, MV, and EM estimates. Fig. 5.5A shows the results. The accuracy of EM estimates $(\kappa = 0.79[0.03])$ was significantly higher than those from either IS $(\kappa = 0.68[0.07])$ or MV $(\kappa = 0.74[0.03])$. The consistency of EM $(\kappa = 0.83[0.02])$ and MV $(\kappa = 0.83[0.03])$ were similar, while IS $(\kappa = 0.79[0.04])$ was somewhat lower. It is also noteworthy that the variability in accuracy and consistency of the EM estimates was lower than those from IS, as evidenced by the narrower IQRs.

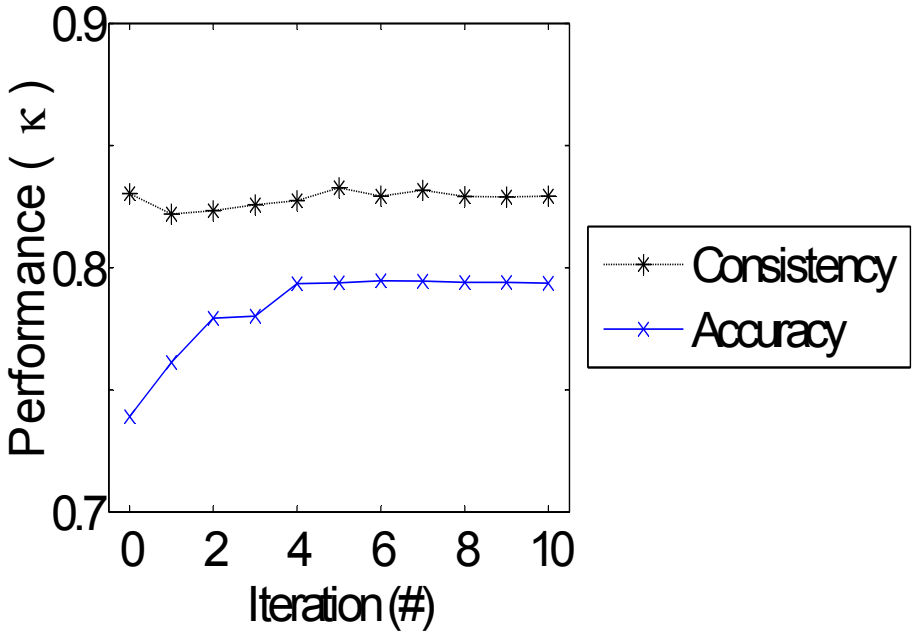

Fig. 5.4. Performance of the Expectation-Maximization estimator as a function of the number of iterations in "true" pattern clinical data.

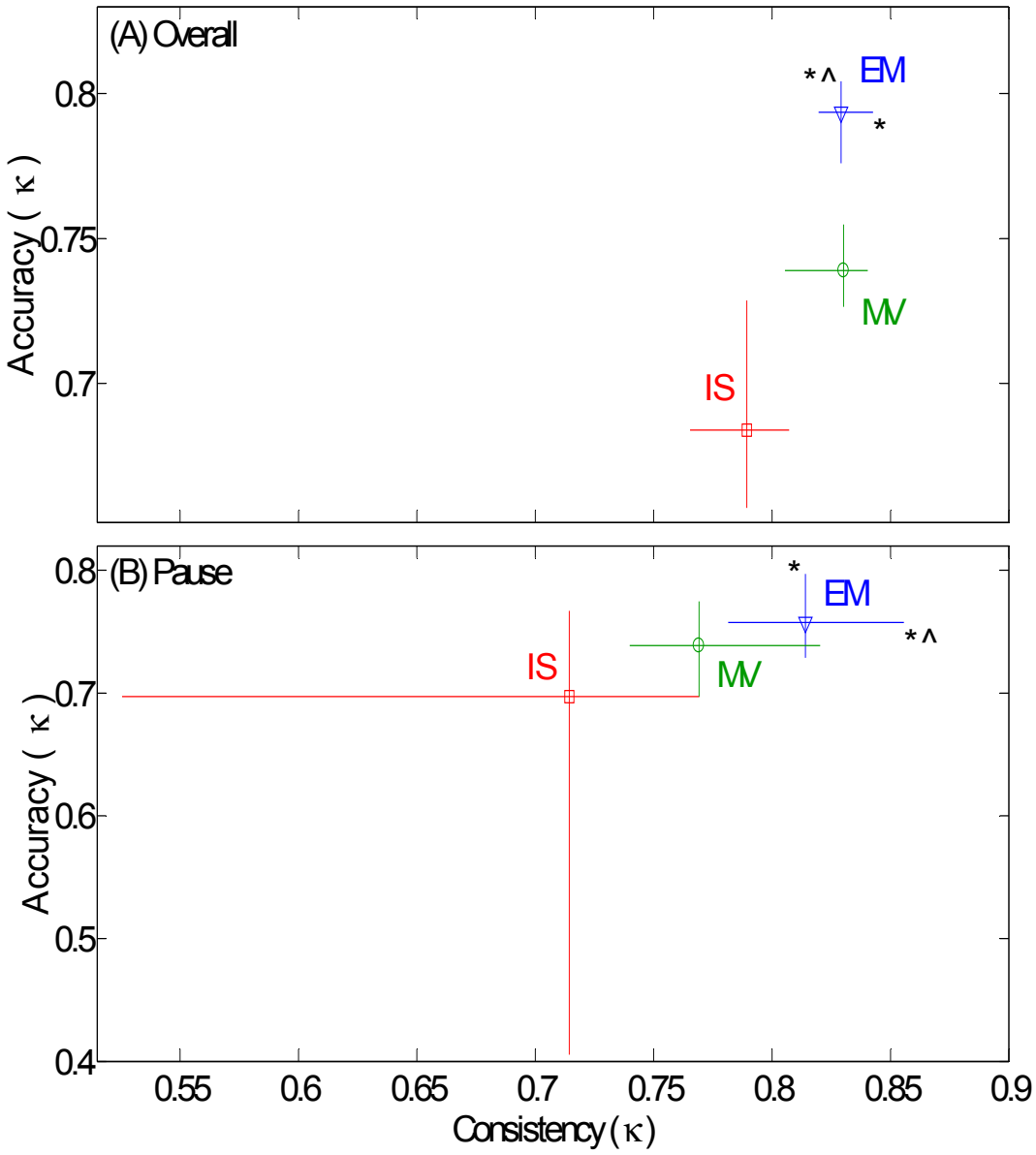

Fig. 5.5. Consistency and accuracy of respiratory pattern estimators applied to "true" pattern clinical data: individual scoring (IS, red-square), majority vote (MV, green-circle), and expectation-maximization (EM, blue-triangle). The points represent median values, and the bars interquartile ranges. P-values < 0.05 between EM and IS are indicated by '*', and between EM and MV by '^'.

Fig 5.5B shows a similar comparison for the PAU pattern, since pauses are particularly relevant for the study of POA. EM estimates were significantly more consistent $(\kappa = 0.81[0.07])$ than those from IS $(\kappa = 0.71[0.24])$ and MV $(\kappa = 0.77[0.08])$, and were significantly more accurate $(\kappa = 0.76[0.07])$ than those from IS $(\kappa = 0.70[0.36])$. The median accuracy of EM was greater than that of MV $(\kappa = 0.74[0.08])$, but the variability was too large to achieve statistical significance.

5.8. Discussion

We presented an expectation-maximization (EM) method to combine the results from multiple scorers of respiratory behavior, and applied it to the analysis of data from infants. The method estimates the most likely respiratory pattern by weighting the contributions from each scorer according to their individual performance; it was adapted from the ground truth estimator in [144], and uses expectation-maximization (EM) to iteratively refine the estimates. Application of the method to both simulated and real data demonstrated the method to be more accurate and consistent than individual scorers (IS), or majority vote (MV).

5.8.1. Simulation Analysis

We used Monte Carlo simulations to compare the performances of EM, IS and MV in estimating the true respiratory pattern.

The use of MV improved the accuracy of true-pattern estimates from that of IS, because the method reduced the inherent subjectivity of manual scoring by averaging the votes to obtain the estimates. Thus, it is reasonable to say that MV was an appropriate starting point in our search for a true-pattern estimator.

Using EM to refine the true-pattern estimates further improved the accuracy. This improvement was evident for any number of scorers. Accuracy of EM increased faster and higher than that of MV with the number of scorers. The analysis revealed that the same accuracy was obtained with the 5-scorer EM and the 25-scorer MV. This means that EM achieved a given accuracy with

much less scoring effort than MV, and this boost in accuracy required only post-processing of the manual analyses, with no additional monetary costs.

The variability of accuracy estimates, evaluated using the interquartile range (IQR), improved with the number of scorers for all 3 estimators. The IQR of EM estimates decreased much faster than those of IS and MV. In fact, EM had the lowest IQR at any number of scorers, and with 7 scorers had lower IQR than IS and MV with 25 scorers.

In summary, the EM estimator was much better than IS and MV for any number of scorers, and was able to achieve a given level of accuracy with much less effort than the other 2 estimators.

5.8.2. Evaluation with Real Infant Data

We then applied the estimators to 125 hrs of real infant data, and evaluated them using the two copies of the 152 segments with known "true" patterns that were inserted in each data record. MV had better accuracy and consistency than IS, similar to that found in the simulation study. Since MV had better performance than IS, it was the most appropriate starting point for EM. Overall accuracy increased with the EM iteration, while consistency remained mostly constant. Once EM converged, it had significantly higher accuracy than either IS or MV. The EM results were also more consistent than those from IS, maintaining the consistency gained by MV. The performance of the EM estimator was excellent; it had a consistency of $\kappa > 0.8$, and an accuracy of $\kappa \approx 0.8$. Both EM and MV produced estimates more objective than IS, because they reduced subjectivity by averaging assigned patterns from multiple scorers. As a consequence, the estimates had lower IQR values as the number of scorers increased. However, the EM approach was significantly better than MV. This was due to the informed combination of manual analyses by weighting the contribution from each scorer by their individual performance.

These results with clinical data confirmed what was found in the simulation study. The EM estimator represented an excellent choice to combine analyses from multiple, manual scorers, since it had the highest accuracy and consistency. Additionally, they also showed that performance of EM improved with the number of iterations, until the method reached convergence.

The pattern-specific analysis showed that EM had an excellent accuracy and consistency for PAU, making it the method of choice to study POA. Further analysis revealed that the patterns Unknown (UNK), and Movement Artifact (MVT) had the largest gains in performance with the EM estimator. UNK is a pattern that groups ambiguous patterns, bad signal segments, and any other pattern not defined by the other 5 patterns. In [10] we found that UNK had the most intra- and inter-scorer disagreement and the highest confusion with other patterns, especially with MVT. By using EM, we were able to significantly reduce the ambiguity of UNK, and improve the distinction between MVT and UNK.

5.8.3. "Gold Standard"

EM classified the respiratory patterns better than MV and IS, in both simulated and clinical data. In fact, EM was the only method to have excellent consistency $(\kappa = 0.83)$ and near-excellent accuracy $(\kappa = 0.79)$. Thus, the use of EM to estimate the most likely respiratory patterns should be considered a better "gold standard" than manual scoring because it has better accuracy and consistency and reduces subjectivity and variability by combining the patterns assigned by several, manual scorers. Fig. 5.6 shows an example of applying EM to estimate the "gold standard" respiratory patterns, using the example from the introduction (i.e., Fig. 5.1).

5.8.4. Possible Limitations

With the MV method more than one pattern might have the highest number of votes, i.e., there might be ties. An unbiased strategy to deal with ties is to randomly select one of the patterns holding the majority at every sample that this occurs. In this work we used the MATLAB function 'max' to determine the MV. When there are ties, the 'max' function returns the index of the pattern that is found first, instead of randomly selecting one of the tied patterns. Only 5 % of the clinical data set used in this work had ties, so the bias that might have been added to the performance of MV by this implementation is minimal.

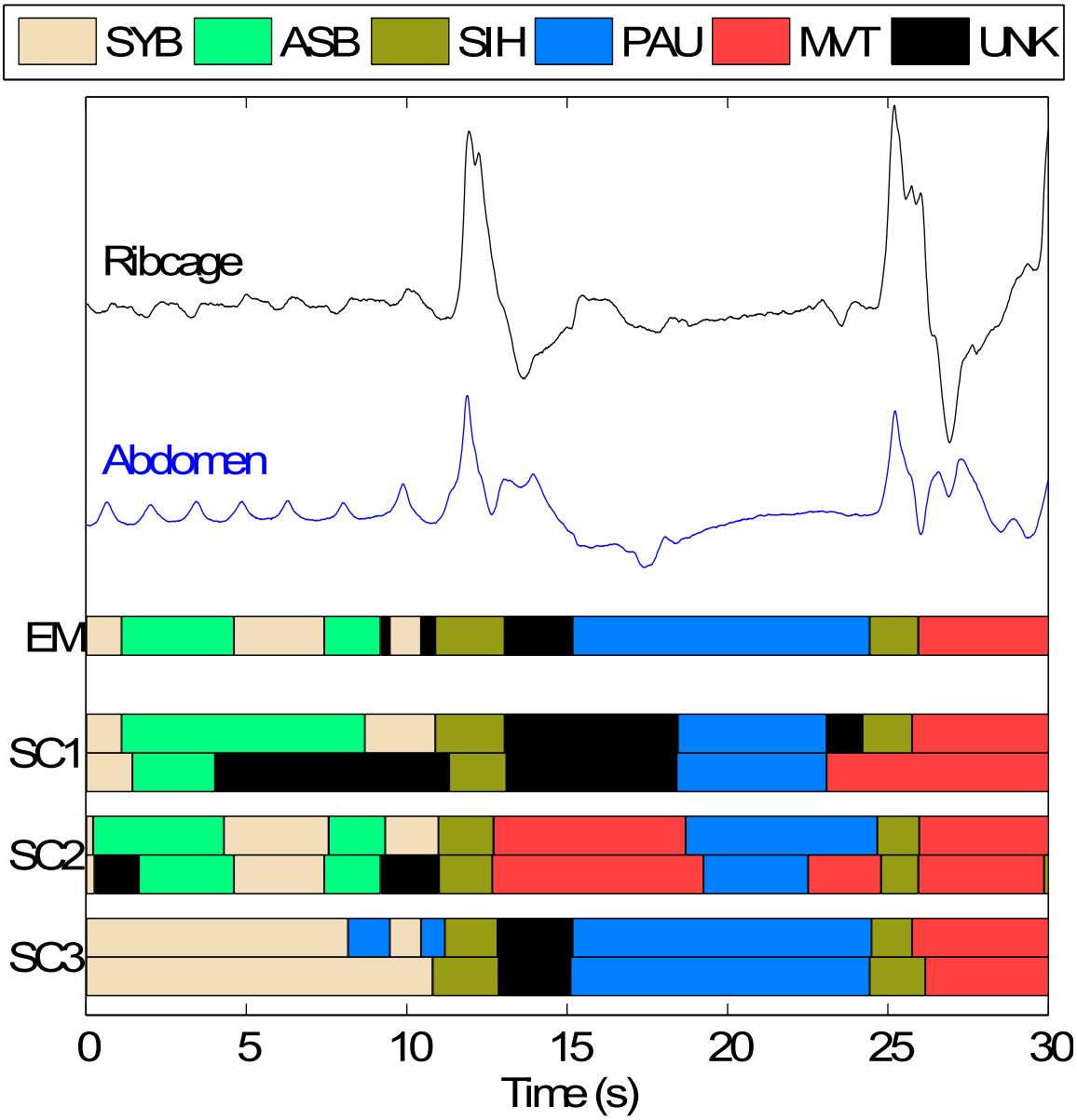

Fig. 5.6. Respiratory patterns on a sample epoch of respiratory data as estimated by Expectation-Maximization (EM) and 3 manual scorers (SC1, SC2, and SC3). Ribcage and abdomen are respiratory inductive plethysmography measurements in arbitrary units. SYB = Synchronous-breathing, ASB = asynchronous-breathing, SIH = sigh, PAU = respiratory pause, MVT = movement artifact, UNK = unknown.

A limitation of the simulation study could be the manner in which the scorers were simulated. The confusion matrix of each simulated scorer was modeled after the combined confusion observed in real manual scorers. The simulation of each individual scorer's confusion matrix required the use of unimodal beta distributions. The unimodal beta distribution was selected because it is defined in the domain $[0,1]$, which spans all possible probability values. Also, it is bell-shaped so most probable values are located around the mode. To verify that the beta distribution was an appropriate model, we estimated the confusion matrix for each scoring sequence in each data record (for a total of 126 matrices), and looked at the probability distribution of each diagonal element. We found that the distribution in 4 out of 6 patterns (ASB, SIH, PAU and MVT) had a shape similar to a unimodal beta (see Figure 5.7), justifying the use of such distribution.

A problem in the evaluation of performance with real respiratory data is that the ground truth respiratory patterns are not known. Consequently, we evaluated accuracy using the patterns assigned by a very experienced, independent, expert, manual scorer (REF) as reference. In addition, only segments assigned the same pattern by REF in two, independent scoring sessions were used. Consequently we believe that these "true" segments provide an unbiased basis for comparison. This contention is supported by the results presented in Table 5.2 where in all cases the probability of assignment was greatest for the "true" pattern.

Furthermore, the results of the consistency analysis support the effectiveness of the EM analysis since these estimates were not dependent on the analysis performed by REF, but only on the agreement between the patterns assigned to the first and second copies of the 152 "true" pattern segments.

5.8.5. Implications for Analysis of Respiratory Data

Respiratory data is generally analyzed using conventional manual scoring (CMS), based on the guidelines provided by the American Academy of Sleep Medicine (AASM) [8]. We recently developed a set of tools to assist manual scoring whose use improves the analysis repeatability from that obtained with CMS [10]. Nevertheless, the results from any one scorer will always

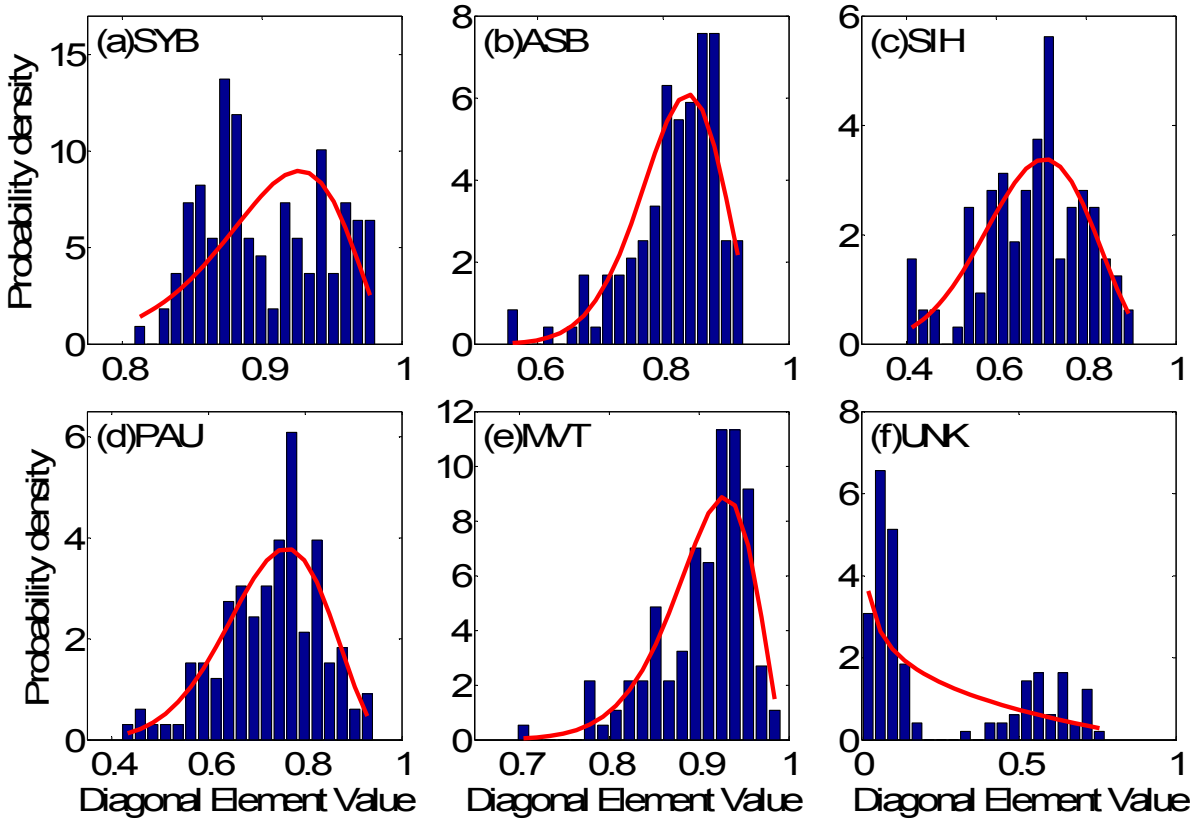

Figure 5.7. Probability density of the diagonal value of the confusion matrix estimated from real, manual scorers for each respiratory pattern type. (a) Synchronous-breathing (SYB), (b) asynchronous-breathing (ASB), (c) sigh (SIH), (d) respiratory pause (PAU), (e) movement artifact (MVT), and (f) unknown (UNK). Bars correspond to the histogram of the actual data estimated with 20 bins, and red lines show the fitted beta densities.

suffer from bias and subjectivity. In this paper, we demonstrate that the accuracy of manual estimates of the most likely respiratory patterns can be further improved by combining the results from multiple scorers using an EM algorithm. Using EM permits a specified accuracy to be achieved with many fewer scorers than for MV. Thus simulations demonstrated that using EM with 5 scorers resulted in estimates with the same accuracy and less variability than using MV with 25 scorers. These factors greatly improve the feasibility of respiratory pattern studies because the need for fewer scorers reduces the analysis costs and time.

EM post-processing, combined with the tools we described in [10], provides a significant improvement in the repeatability of manual analysis, and also provides a comprehensive description of the respiratory pattern as a function of time. The result is an objective, comprehensive "gold standard" with high accuracy and consistency, compared to the more limited, highly variable CMS.

This comprehensive "gold standard" that better documents the respiratory patterns will be useful for the study of POA, and any other study related to respiratory patterns such as, the prediction of extubation readiness in preterm infants [120], sleep apnea, asthma, opioid effects, etc. The repeatable, reliable analysis made possible with these methods opens the door for large multi-institutional and longitudinal studies, where the volume of data to be analyzed is too large to be performed by a single scorer, and so multiple scorers are a necessity. With CMS using multiple scorers will add noise due to high intra- and inter-scorer variability. This will reduce the statistical power of the study making it necessary to use larger sample sizes. EM enables the development of these large studies because it yields accurate and consistent analyses of the respiratory patterns while minimizing the number of scorers.

The availability of this objective "gold standard" will also help efforts to automate the analysis of respiratory behavior. It can provide the reliable reference analysis to which automated methods need to be compared, since it minimizes the effects of scorer bias and subjectivity. It also represents a reliable, unbiased source for training supervised learning algorithms, which perform better when the reference is accurate.

5.9. Conclusion

We presented a method based on expectation-maximization (EM) to combine analyses from multiple scorers of respiratory patterns, and compared it to individual scorers (IS), and the majority vote (MV) approach. EM estimated the most likely respiratory patterns with higher accuracy and consistency than IS and MV. Moreover, this improvement came at only minimal computational cost and no additional manual effort. The EM method represents an improved "gold standard" for the analysis of respiratory data.

6. Automated Off-Line Respiratory Event Detection for the Study of Postoperative Apnea in Infants

6.1. Preface

In the previous Chapters I presented methods to make manual scoring a repeatable, reliable, and comprehensive analysis of infant respiratory patterns. However, these methods still rely on manual analysis, which is labor intensive, time consuming, and expensive.

This Chapter describes AORED, an Automated Off-Line Respiratory Event Detector I developed to automate the analysis of respiratory patterns, with the objective to make it fast and low-cost. AORED first automatically estimates metrics of respiratory behavior related to the amplitude, frequency, and phase of ribcage and abdomen respiratory signals. These metrics are then used as inputs of a set of respiratory pattern detectors that compare the metrics to thresholds to detect events. Thresholds are obtained from Receiver Operating Characteristics (ROC) analysis of the metrics using as reference a sample manual analysis performed by an expert. The outputs of these pattern detectors are then combined using a decision tree to classify the respiratory patterns on a sample-by-sample basis. I carried out a simulation experiment showing that the metrics are robust in high noise conditions, and compared the results from AORED to those from an expert, manual scorer, and found that both analyses agreed well.

This is the first study I carried out during my Ph.D. The results of this Chapter helped identify two important aspects in the analysis of respiratory data that were addressed in this thesis. The first aspect was the need for a comprehensive "gold standard" reference for evaluation of automated methods. In this Chapter we evaluated the performance of AORED in terms of its agreement with a single, manual scorer. This is an important step but it was necessary to establish a more objective, repeatable, reliable reference since a single human scorer is subjective and has inherent variability. This led to the development of the manual analysis tools described in Chapter 4 and 5, which enable a comprehensive, reliable analysis of infant respiratory patterns with very low variability. As part of this process we identified that to

produce a more objective "gold standard" it was necessary to recruit several manual scorers to perform the analysis, and that it was necessary to obtain scientific and ethics approval to secure funding for recruitment of these additional manual scorers. The second aspect that this Chapter helped to identify was the need to make the analysis of respiratory patterns fully automated, removing the need for manual scoring to determine detection thresholds. The result is a method named AUREA (Automated Unsupervised Respiratory Event Analysis), which is presented in Chapter 7.

I used a separate dataset from the one described in Chapter 3 for the development of this Chapter. These data were acquired by Brown et al. before I started my Ph.D. The data were originally reported in [7, 71]. These data were not included in the public data set. The data acquisition methods were generally similar to those described in Chapter 3. Key differences are: (i) the cut-off frequency of the anti-aliasing filter of 15 Hz compared to 10 Hz; the resolution of the analog-to-digital converter was 12 bits in this Chapter and 16 bits in Chapter 3; and recording sessions were not supervised for the present Chapter.

This Chapter was published in IEEE Transactions on Biomedical Engineering [85]:

A. Aoude, R. E. Kearney, K. A. Brown, H. Galiana, and **C. A. Robles-Rubio**, "Automated Off-Line Respiratory Event Detection for the Study of Postoperative Apnea in Infants," *IEEE Trans Biomed Eng*, vol. 58, pp. 1724-1733, 2011. Digital Object Identifier: 10.1109/TBME.2011.2112657, © 2011 IEEE.

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada.

6.2. Abstract

Previously, we presented automated methods for thoraco-abdominal asynchrony estimation and movement artifact detection in respiratory inductance plethysmography (RIP) signals. This paper combines and improves these methods to give a method for the automated, off-line detection of pause, movement artifact and asynchrony. Simulation studies demonstrated the new combined method is accurate and robust in the presence of noise. The new procedure was successfully applied to cardiorespiratory signals acquired postoperatively from infants in the recovery room. A comparison of the events detected with the automated method to those visually scored by an expert clinician, demonstrated a higher agreement ($\kappa = 0.52$) than that amongst several human scorers ($\kappa = 0.31$) in a clinical study [9]. The method provides the following advantages: 1) it is fully automated; 2) it is more efficient than visual scoring; 3) the analysis is repeatable and standardized; 4) it provides greater agreement with an expert scorer compared to the agreement between trained scorers; 5) it is amenable to on-line detection; and 6) it is applicable to uncalibrated RIP signals. Examples of applications include respiratory monitoring of postsurgical patients and sleep studies.

6.3. Introduction

Respiratory inductive plethysmography (RIP) is a widely accepted method for qualitative and quantitative respiratory monitoring [149, 150] that is used commonly in sleep laboratories and at home [151]. RIP offers noninvasive, robust monitoring which is well tolerated by patients and recommended for diagnostic testing. Therefore, we have applied RIP to study respiration in infants who have received anesthesia and are at risk of postoperative apnea [17].

Visual scoring of cardiorespiratory data is the preferred analysis to identify clinically relevant cardiorespiratory events, including central and obstructive apnea, in part because no reliable automated method has been accepted to date [152-154]. The likelihood of human error in visual coding is high and the results can be subjective [155]. Thus, the development of a reliable automated method to detect respiratory events would provide for a more objective and efficient analysis and have potential application for on-line apnea detection.

Automated methods for the analysis of cardiorespiratory data have employed a wide range of approaches including: hidden Markov models [156], artificial neural networks [80, 89], recursive least squares [71], and fuzzy logic [157]. In general, these methods perform well on simulated data, but encounter difficulties when applied to clinical data where movement artifact is prominent. The need to automatically segment cardiorespiratory signals into epochs with and without artifact has been well recognized. Indeed it has been reported that the performance of automated cardiorespiratory monitoring procedures can be improved significantly if signal segments corrupted with artifacts were systematically and reliably identified [70]. Automated signal segmentation procedures would also be useful in the off-line analysis of the long records of respiratory data acquired during sleep.

Automated event detection algorithms used in clinical practice have been developed for polysomnography (PSG) where RIP is widely used. These are for the most part commercially available algorithms that rely on a calibrated RIP system, frequently based on the Qualitative Diagnostic Calibration (QDC) method [77]. They also require a measure of airflow [158, 159] to detect respiratory events. Sensors at the nose and mouth are poorly tolerated in the recovery room in a patient population of infants who are continuously monitored regardless of their behavioral state.

We previously presented methods for detecting movement artifacts [3] and estimating the thoraco-abdominal asynchrony [4] in uncalibrated infant RIP data, which do not require a sensor applied to the face. The current work describes how these methods were improved and combined with a pause detection algorithm, to yield a comprehensive off-line method that identifies pauses, segments corrupted by movement artifact, and asynchrony between the ribcage and abdomen RIP signals. Some aspects of this work have been part of a conference presentation [5].

The paper is organized as follows: Section 6.4 describes each event detector and how their outputs were combined; Section 6.5 presents the results of simulation studies validating the performance of these detectors using artificially manipulated respiration signals; Section 6.6 provides results obtained by applying the new methods to real infant respiration signals and

compares the events detected automatically to those resulting from visual scoring; Section 6.7 discusses the results and provides some concluding remarks.

6.4. Methods

The proposed method uses detectors for pause, movement artifact and asynchrony. Each detector decides whether or not an event is present by comparing a test statistic to a threshold. The following sections describe the initial filtering of the data, detail the operation of each detector, and demonstrate how they are combined to estimate the respiratory state.

6.4.1. Filtering

Offsets and exponential decays observed in real infant data [5], were removed by applying a digital high-pass elliptical filter of order 6 with a cut-off frequency of 0.08 Hz to the clinical data. The peak-to-peak ripple in the pass band was set to 0.1 dB and the minimum attenuation in the stop band was 50 dB.

The movement artifact and asynchrony detectors use the outputs of a bank of digital elliptical filters with low and high cut-off frequencies f_l and f_h respectively defined in Table 6.1. The order, peak-to-peak ripple in the pass band and the minimum attenuation in the stop band were the same as for the high-pass filter. These filters were chosen to span frequencies from 0 to 2 Hz since most power in infant quiet breathing lies in the range 0.4 - 2.0 Hz, while movement artifacts occur primarily at lower frequencies [4]. The sets $I = \{3,4,\ldots,13\}$ and $J = \{1,2\}$ define the filter numbers that span the quiet breathing and movement artifact bands respectively. A filter bandwidth of 0.2 Hz was used to ensure that the breathing frequency was estimated within a narrow band.

It should be noted that the filter bank could have been implemented using the Short-Time Fourier Transform (STFT). We opted to use a filter bank instead because: (1) the objective was to detect events on a sample-by-sample basis which would require the STFT windows to overlap by N-1 samples for a window of length N which would be computationally inefficient; (2) the analysis focuses only on the low frequency components (0 – 2 Hz) and so does not require the greater

Filter Number (i)	f_l (Hz)	f_h (Hz)
1	-	0.20
2	0.15	0.35
3	0.30	0.50
4	0.45	0.65
5	0.60	0.80
6	0.75	0.95
7	0.90	1.10
8	1.05	1.25
9	1.20	1.40
10	1.35	1.55
11	1.50	1.70
12	1.65	1.85
13	1.80	2.00

Table 6.1. Filter Bank

frequency resolution provided by the STFT; (3) the filter implementation gives the selectively filtered signal used by the Asynchrony Detector (Section 6.4.4) with no additional computation.

We define ab and rc as the raw abdominal and ribcage RIP signals respectively, while ab_i and rc_i are the corresponding outputs from the i^{th} filter. We assumed that a quiet breathing segment would have the most power at a frequency equal to the breathing rate. Therefore, the breathing frequency f_{max} was estimated as the central frequency of the filter with the highest power i_{max}. This yielded a sample-by-sample estimate accurate to within 0.2 Hz, calculated over a time equal to the filters' window length. Note that because we used symmetric two-sided filters the breathing frequency estimates have no associated time delay. The value of f_{max} was set to zero if the highest power was found in filters 1 or 2, since these filters correspond to the expected frequencies of movement artifact. In addition, if the abdominal and ribcage signals produced different breathing frequency estimates, the abdominal signal was given precedence since we found that it generally had a better signal-to-noise ratio.

6.4.2. Pause Detection

Pauses are defined by a lack of respiratory effort and so the RIP signals would be expected to have low power in the quiet breathing band. Consequently, we designed the pause test statistics p^{ab} and p^{rc}, to quantify the power of quiet breathing in the abdominal and ribcage signals respectively. The test statistic for the abdomen over a window of length N_P was defined as:

$$p^{ab}\left[n, N_P\right] = \sqrt{\frac{1}{\wp^{ab} N_P} \sum_{k=n-(N_P-1)/2}^{n+(N_P-1)/2} ab_{bp}^2\left[k\right]}, \tag{6.1}$$

Where $\wp^{ab}$ is the median power of all segments of length N_P in ab_{bp}, and ab_{bp} is the band-pass filtered abdominal signal, with frequencies in the quiet breathing band only (i.e., using a band-pass filter with cut-off frequencies at 0.4 Hz and 2.0 Hz). Note that the method was developed for off-line use, so the entire data record is available to determine the median power of all data ($\wp^{ab}$ and $\wp^{rc}$). The abdomen pause detector was defined as:

$$P^{ab}[n,N_P]=\begin{cases}1,\ if\ p^{ab}[n,N_P]\le\gamma_P^{ab}\\0,\ if\ p^{ab}[n,N_P]>\gamma_P^{ab}\end{cases}. \tag{6.2}$$

Thus, a pause candidate is detected in the abdominal signal if the pause test statistic p^{ab} is below the threshold γ_P^{ab}. A similar pause test statistic p^{rc}, threshold γ_P^{rc}, and detector P^{rc}, were defined for the ribcage signal. The overall pause detector was determined by the logical AND of the ribcage and abdominal pause detectors, that is

$$P[n,N_P]=P^{ab}[n,N_P]\,\&\,P^{rc}[n,N_P]. \tag{6.3}$$

6.4.3. Movement Artifact Detection

Movement artifacts were detected using a test statistic that compares the output power from the filter with highest power in the quiet breathing band to that from the filter with highest power in the movement artifact band. The output power for each filter output was defined as:

$$\wp_i^{ab}[n,N_M]=\frac{1}{N_M}\sum_{k=n-(N_M-1)/2}^{n+(N_M-1)/2}ab_i^2[k],\ i=1,\ldots,13. \tag{6.4}$$

Next, the movement test statistic for the abdomen m^{ab} was defined as:

$$m^{ab}[n,N_M]=\frac{\max_i\left\{\wp_i^{ab}[n,N_M]\right\}_{i\in I}-\max_i\left\{\wp_i^{ab}[n,N_M]\right\}_{i\in J}}{\max_i\left\{\wp_i^{ab}[n,N_M]\right\}_{i\in I}+\max_i\left\{\wp_i^{ab}[n,N_M]\right\}_{i\in J}}. \tag{6.5}$$

Using the output signal with maximum power ensures that the breathing energy is isolated from that of any other source in the breathing band (such as electronic noise).

The movement test statistic m^{ab} is compared to the threshold γ_M^{ab} to decide whether or not a movement artifact is present in the abdominal signal. This detector was defined as:

$$M^{ab}[n,N_M]=\begin{cases}1, & if\ m^{ab}[n,N_M]\le\gamma_M^{ab}\\ 0, & if\ m^{ab}[n,N_M]>\gamma_M^{ab}\end{cases}. \tag{6.6}$$

A similar test statistic m^{rc}, threshold γ_M^{rc}, and detector M^{rc} were defined for the ribcage signal. Since movement is expected to cause artifacts in both abdomen and ribcage signals, the overall movement detector for the subject was determined as:

$$M[n,N_M]=M^{ab}[n,N_M]\&M^{rc}[n,N_M]. \tag{6.7}$$

6.4.4. Asynchrony Detection

In [4] we described an automated phase estimation algorithm that has the advantage of working with uncalibrated RIP measurements to provide quantitative phase estimates (ϕ) in the range $[0,1]$, corresponding to $[0,180]$ degrees.

For the present work, we modified this phase estimator to enhance its performance as an asynchrony test statistic. We improved the signal-to-noise ratio of the input by using selectively filtered RIP signals, rather than the band-pass filtered RIP signals used in [4]. These selectively filtered signals were obtained using the breathing frequency estimate f_{max} at each sample n, by making $ab_S[n]=ab_{i_{max}}[n]$ and $rc_S[n]=rc_{i_{max}}[n]$. Fig. 6.1 shows an example of this signal.

Asynchrony is detected ($A=1$) if the test statistic ϕ is above the threshold γ_A:

$$A[n]=\begin{cases}1, & if\ \phi[n]\ge\gamma_A\\ 0, & if\ \phi[n]<\gamma_A\end{cases}. \tag{6.8}$$

6.4.5. Combining the Detectors

The respiratory state at each sample n, was determined by combining the three outputs of the detectors logically as shown in Fig. 6.2 to ensure that only one event is detected at each time.

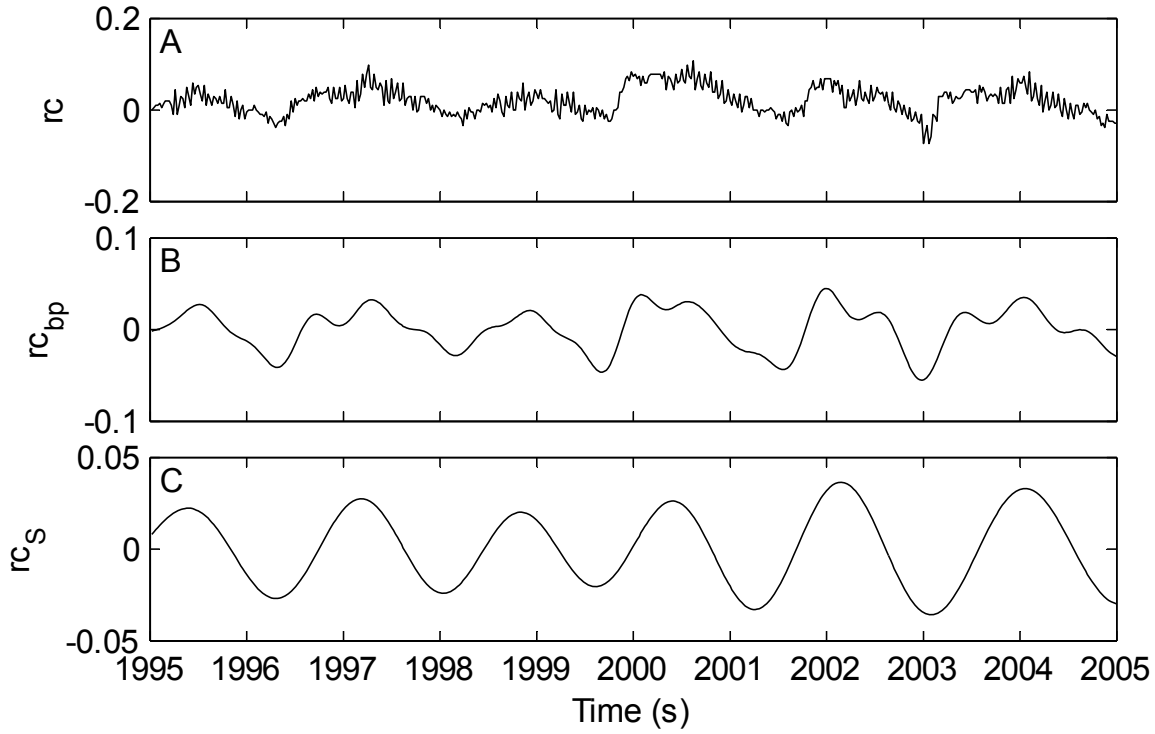

Fig. 6.1. Typical RIP signal from a 47 week old infant: A. Raw original Ribcage RIP signal (rc), B. Band-pass filtered Ribcage RIP signal (rc_{bp}), C. Selectively filtered version of the Ribcage RIP signal (rc_S).

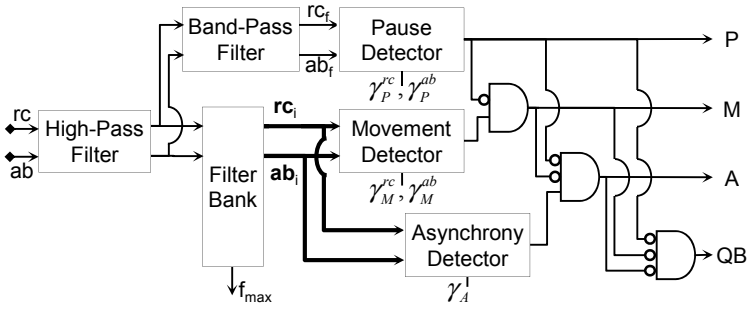

Fig. 6.2. Block diagram of detector combination. P=Pause, M=Movement Artifact, A=Asynchrony, QB=Quiet Breathing.

Pause detection is assigned the highest precedence so that when a pause is detected the other states are forced to zero. The rationale for this is that during pauses there is, by definition, very little power in the quiet breathing frequency band. Consequently, even a low power signal in the movement artifact band would trigger the movement detector leading to false positives. Similarly, the asynchrony detector output will be noisy at low power levels leading to false alarms. Movement detection is assigned the second level of precedence with the output of the asynchrony and quiet breathing states forced to zero when movement is detected. The rationale for this is that movement artifacts may have components in the quiet breathing frequency band causing asynchrony detection to be unreliable. Asynchrony detection has the third level of precedence. Samples not assigned to any of the three previous categories are scored as quiet breathing.

6.5. Method Validation: Simulation Results

6.5.1. Simulated Data

We evaluated the performance of the method using simulated data sets with known properties. To this end, we isolated a representative segment comprising 30 s of infant quiet breathing sampled at $F_s = 50Hz$. As Fig. 6.3 shows, the segment contained synchronous thoraco-abdominal oscillations and no significant movement artifact. The data from this segment were manipulated to simulate different respiratory conditions as follows:

(i) Pause was simulated by attenuating the quiet breathing signals.

(ii) Low frequency movement artifact was modeled as a stochastic diffusion process called *mean reverting Ito process* (see [160]). This was implemented with the stochastic differential equation $d\tilde{m}(t) = \mu(\tilde{m},t)dt + \sigma dW(t)$, where $\tilde{m}(t)$ is a random process that fluctuates randomly, but tends to revert to $\tilde{\mu} = 0$, $W(t)$ is a standard Wiener process, $\mu(\tilde{m},t) = \tilde{c}(\tilde{\mu} - \tilde{m}(t))$ is the drift function, $\sigma = 0.5$ is the short term variance, and $\tilde{c} = 0.1$ is the speed of the reversion; all defined as in our previous work [4].

(iii) Asynchrony between the abdominal and ribcage signal was simulated by shifting one signal with respect to the other.

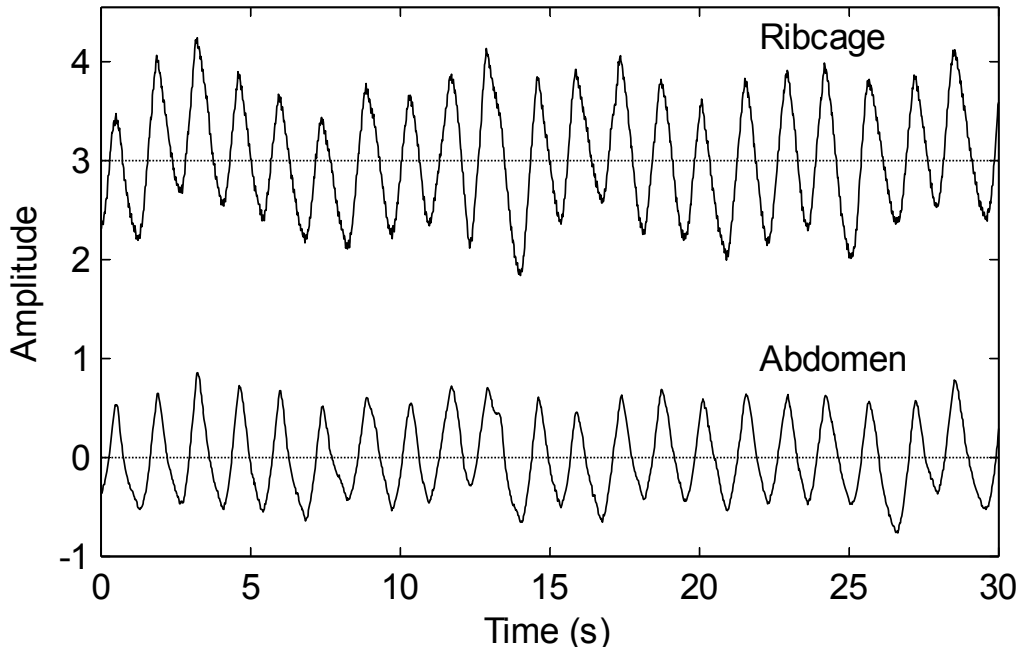

Fig. 6.3. Isolated quiet breathing segment from an infant used in the simulated data. Ribcage and Abdomen are in arbitrary units and offset for clarity.

(iv) Electronic and sensor noise were simulated by adding Gaussian white noise.

Hence, the simulated thoracic and abdominal RIP signals were defined as follows:

$$\begin{aligned} RC[n] &= \rho_{11} rc[n] + \rho_{12} g_1[n] + \rho_{13} \tilde{m}_1[n] \\ AB[n] &= \rho_{21} \{ ab[n] w[n] + ab[n+k](1 - w[n]) \} + \rho_{22} g_2[n] + \rho_{23} \tilde{m}_2[n], \end{aligned} \tag{6.9}$$

where $w[n]$ is the transition window for the start of an asynchrony, defined as

$$w[n] = \begin{cases} 1, & if\ n < n_0 - W_L \\ 0.5/W_L\,(n_0 + W_L - n), & if\ n_0 - W_L \le n \le n_0 + W_L \\ 0, & if\ n > n_0 + W_L \end{cases}$$

and n_0 is the time of asynchrony, $2W_L + 1$ is the length of the transition window for a smooth asynchrony simulation, k is the degree of asynchrony (e.g., if k equals half the period ($T_n/2$), then it represents a shift of 180^o), ab and rc are the experimental RIP signals recorded from the abdomen and ribcage, g_1 and g_2 are the processes for sensor and electronic noise, $\tilde{m}_1$ and $\tilde{m}_2$ are *mean reverting Ito processes* representing low frequency movement artifact, and ρ_{11}, ρ_{12}, ρ_{13}, ρ_{21}, ρ_{22}, ρ_{23} are positive constants. All random numbers used in the simulations were generated using the pseudo-normally distributed random number generator in Matlab 7.10 (MathWorks, Inc, Natick, MA).

The Signal-to-Noise Ratio (SNR), was defined as the ratio between the power in the quiet breathing signal, i.e. ab and rc, and the power in the added noise and movement signals ($g + \tilde{m}$). To obtain the different values of SNR we varied the values of ρ_{11} and ρ_{21}, while keeping the other four parameters ρ_{12}, ρ_{13}, ρ_{22}, ρ_{23} constant.

We evaluated two aspects of the performance of each detector: the static behavior, to assess the robustness of the algorithms in the presence of noise; and the detection delay, to examine how long each detector takes to identify the corresponding event. For the static performance we varied the SNR on the signals while keeping a fixed start time. In contrast, the SNR was fixed and the start time varied when evaluating the detection delay. The start time was fixed when

evaluating the static performance of each detector to isolate the effect of noise from that of start time.

6.5.2. Pause Detector Performance

Fig. 6.4A shows a sample simulated epoch with a pause starting at $t = 20s$, and $SNR = 5$ before the event. Fig. 6.4C shows the value of the pause test statistic p^{rc} for this epoch; its value falls quickly after the start of the pause.

To examine the static performance of the pause test statistic, we generated 500 realizations of RC and AB at SNRs in the range $[0.01, 0.35]$ during the pause. Each realization had a pause start time T_P of 20 s, ρ_{11} and ρ_{21} were varied from 1.5 to 22 in the quiet breathing region ($t < 20s$); pauses were simulated by scaling ρ_{11} and ρ_{21} by a random coefficient uniformly distributed in $[0, 0.1]$ so that they varied from 0 to 2.2 in the pause region ($t > 20s$) to give the desired range of SNR values in this last section. The remaining parameters were defined as follows: ρ_{12} and ρ_{22} were set to 0.25, ρ_{13} and ρ_{23} were set to 0 since pauses are not expected to occur during movement artifacts, and there was no asynchrony ($k = 0$). The detector window was set to $N_P = 51$ (1 s at $F_S = 50Hz$).

Fig. 6.4B shows the mean and standard deviation of p^{rc} for each SNR. It is evident that even at low SNR values (i.e., $SNR \geq 0.1$), the pause test statistic was a good indicator since the mean value of p^{rc} was close to 0.1, the theoretical limit from the pause simulation. It also shows a threshold of $\gamma_P^{rc} = \gamma_P^{ab} = 0.3$ would detect most pauses where the respiration power was less than 10% of quiet breathing.

To evaluate pause detection delay, we simulated 5,000 epochs starting at random times T_P uniformly distributed in $[18, 28]$. This range of onset times was selected so that most of the signal had a normal quiet breathing power, to mimic clinical data, where short pauses occur between long sections of quiet breathing. The SNR was 5 in the quiet breathing region and pauses were simulated by scaling ρ_{11} and ρ_{21} by a random coefficient uniformly distributed in [0,0.1]. There

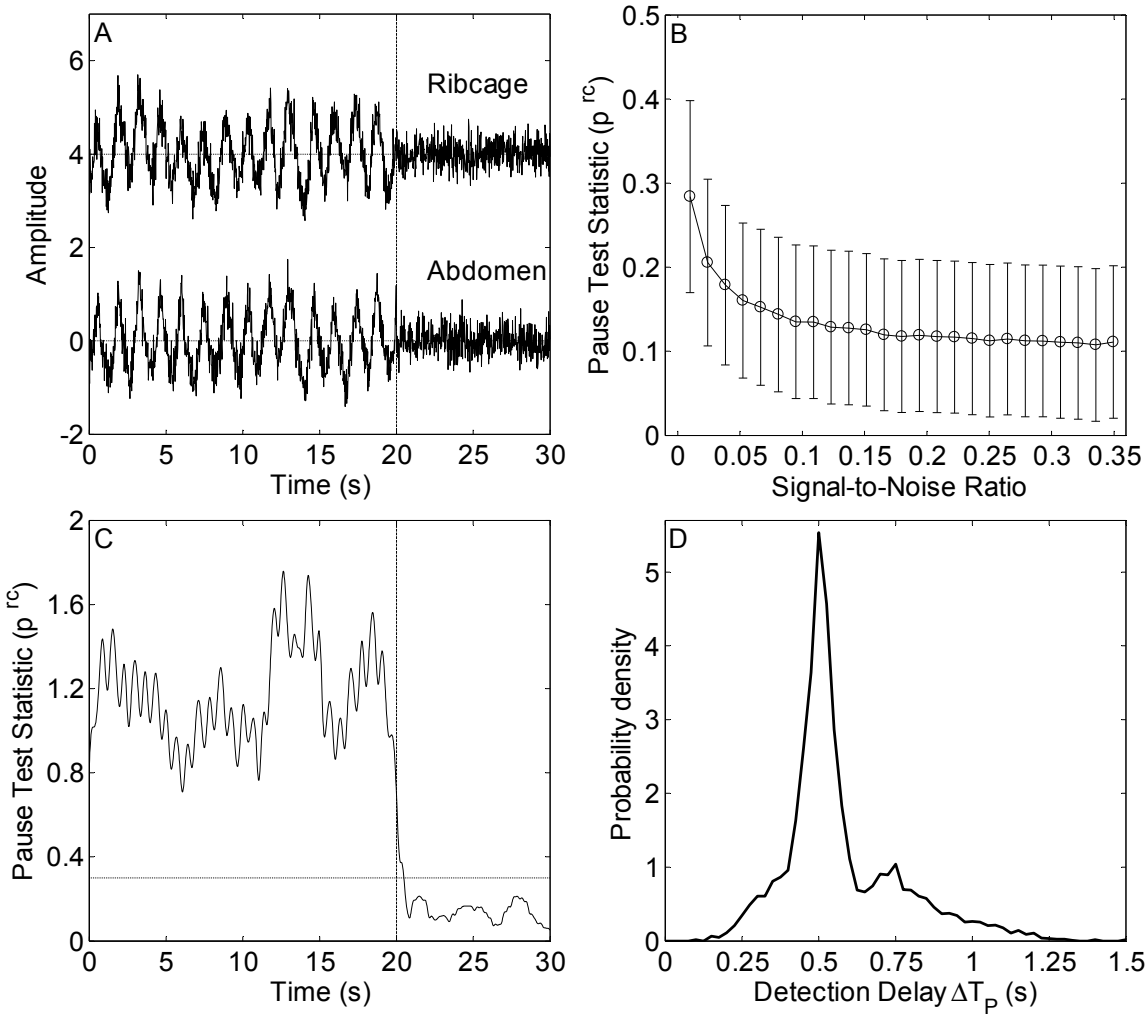

Fig. 6.4. Pause detector performance: A. Simulated epoch with pause starting at $t = 20s$, Ribcage and Abdomen are in arbitrary units and have been offset for clarity, B. Mean and standard deviation of the pause test statistic p^{rc} as a function of SNR, C. p^{rc} for the simulated epoch, D. Probability density of the pause detection delay (ΔT_P).

was no movement artifact or asynchrony simulated. The detector window size was $N_P = 51$. The pause detection delay was defined as $\Delta T_P = T_{PD} - T_P$, where T_{PD} is the detection time defined as the first time when $P = 1$. Fig. 6.4D shows the probability density of ΔT_P. It is clear that the pause detector has a short reaction time, ranging from 0.2 to 1.2 seconds. The detector was also very efficient; it identified pauses in 99% of the epochs simulated (5 misses in 5,005 realizations).

6.5.3. Movement Detector Performance

Figure 6.5A shows an epoch containing a simulated movement artifact starting at $t = 10s$. Fig. 6.5C shows that the movement detection statistic, m^{rc}, drops as soon as the simulated artifact starts.

The static performance of m^{rc} was examined simulating epochs of quiet breathing with different relative amplitudes of movement artifact. The values of ρ_{11} and ρ_{21} were varied from 0 to 2.6, while holding the other simulation parameters constant (i.e., $\rho_{12} = \rho_{22} = 0.25$, $\rho_{13} = \rho_{23} = 1$, $k = 0$) so that the SNR during the movement artifact was ranged from 0 to 3.5. We simulated 500 epochs of 30 s for each SNR using a detector window size of $N_M = 251$. Fig. 6.5B shows the mean and standard deviation of m^{rc} as a function of the SNR; the mean value of the test statistic increases monotonically with the SNR demonstrating that it provides a good estimate of relative magnitude of the movement artifact. The standard deviation was relatively constant across the SNR range.

Next, we investigated the detection delay of the movement detector by simulating 5,000 realizations with movement artifacts starting at a random time T_M within the epoch. We set the SNR to 0.2 during the movement, corresponding to the large artifact contributions observed in the clinical data presented in Section 6.6. The window length was set to $N_M = 251$, and the thresholds to $\gamma_M^{rc} = \gamma_M^{ab} = 0.2$. Values of m^{rc} and/or m^{ab} less than zero indicate that there is more movement power than breathing power, which is the case for $SNR = 0.2$.

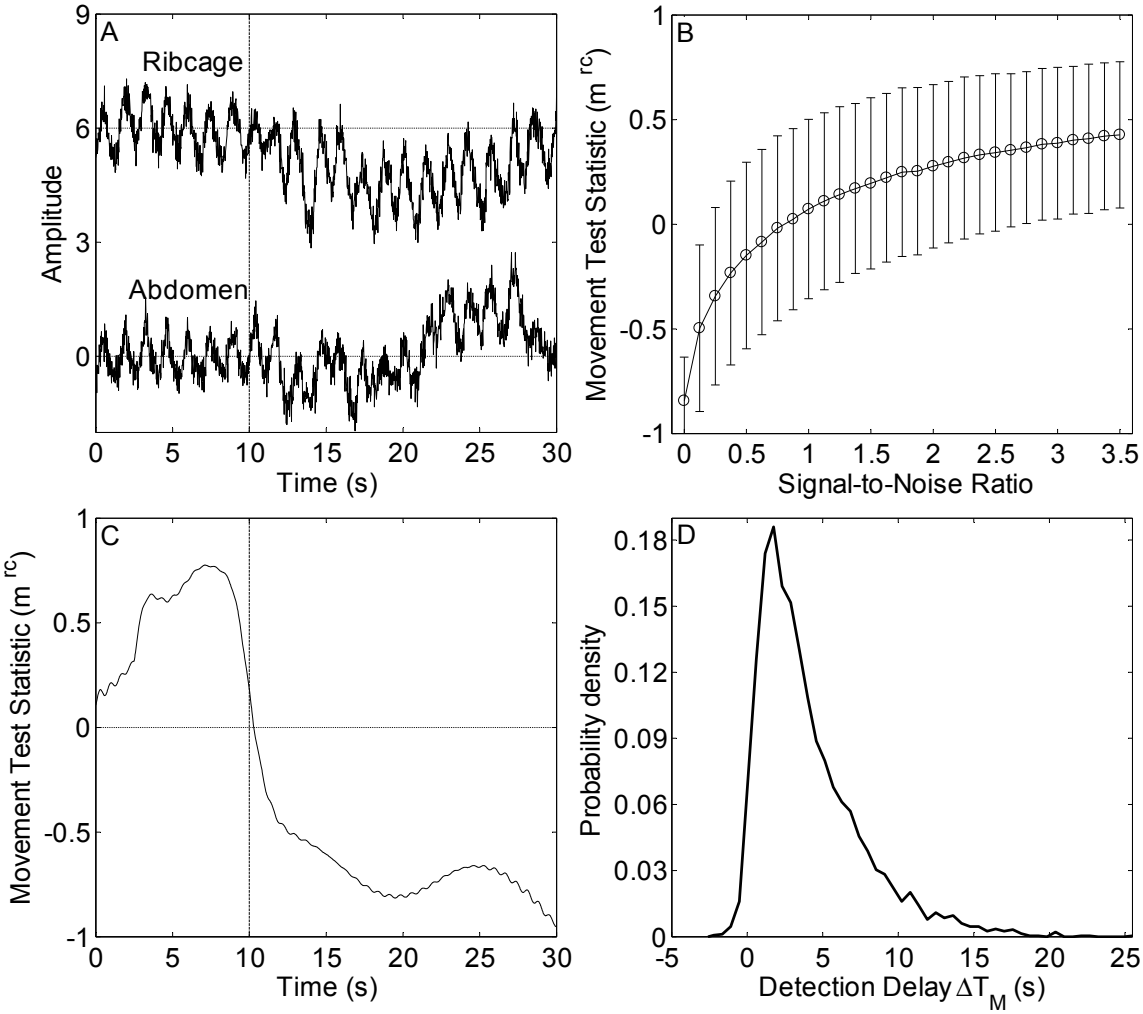

Fig. 6.5. Movement artifact detector performance: A. Simulated epoch with movement starting at $t = 10s$ ($SNR = 0.2$), Ribcage and Abdomen are in arbitrary units and have been offset for clarity, B. Mean and standard deviation of the movement test statistic m^{rc} as a function of SNR, C. m^{rc} for the simulated epoch with movement starting at $t = 10s$, D. Probability density of the movement artifact detection delay (ΔT_M).

We computed m^{rc} and m^{ab} from RC and AB respectively, and combined them as in equation (6.7). Movement artifact was detected in all but 3 of the 5003 realizations simulated. The detection delay was defined as $\Delta T_M = T_{MD} - T_M$, where T_{MD} is the time in which M was set to 1 (i.e., movement as detected). Fig. 6.5D presents the pdf of ΔT_M; it is evident that artifacts have a high probability (0.94) of being detected within the first 10 seconds; that is, within twice the detector window length. Moreover, the probability of detecting an event between -2.5 and 5s was approximately 0.71, meaning that the majority of the events were detected within a delay less than the window length. These relatively long latencies arose due to the low frequency content of the artifact.

6.5.4. Asynchrony Detector Performance

Fig. 6.6A shows a sample simulated epoch with an asynchrony of 180^o ($k = T_N/2$) starting at $t = 10s$. Fig. 6.6C presents the corresponding asynchrony test statistic ϕ calculated with a window of size $N_A = 251$. It is evident that ϕ begins to change at $N_A/2$ samples before the onset of the event, and it reaches its maximum value after approximately N_A samples. This predictive effect is because the window used to obtain ϕ is symmetric about the origin.

To evaluate the static performance of the asynchrony test statistic, we tested its capacity to estimate the actual asynchrony ϕ_{true} for different values of SNR, in the range $[0,3.5]$. For this, we set the simulation parameters to $\rho_{12} = \rho_{22} = 0.25$, $\rho_{13} = \rho_{23} = 0$ (i.e., no movement artifact), ρ_{11} and ρ_{21} were varied from 0 to 3.5, W_L samples for a smooth asynchrony transition, $n_0 = 0$ (i.e., asynchrony started at the beginning of the epoch), and k was selected from a random uniform distribution with limits $[0, T_N/2]$ to span the 0^o-180^o range. Fig. 6.6B shows the mean and standard deviation of the asynchrony estimation error $\phi_E = \phi - \phi_{true}$ for 500 realizations at each SNR value. This demonstrates that the accuracy of the asynchrony test statistic increases with the SNR, as expected. Nevertheless, the asynchrony test statistic produces accurate estimates of phase between the RIP signals for SNRs as low as 1. This makes it valid for clinical applications, since the asynchrony test statistic is only considered in quiet breathing segments, which have high SNR.

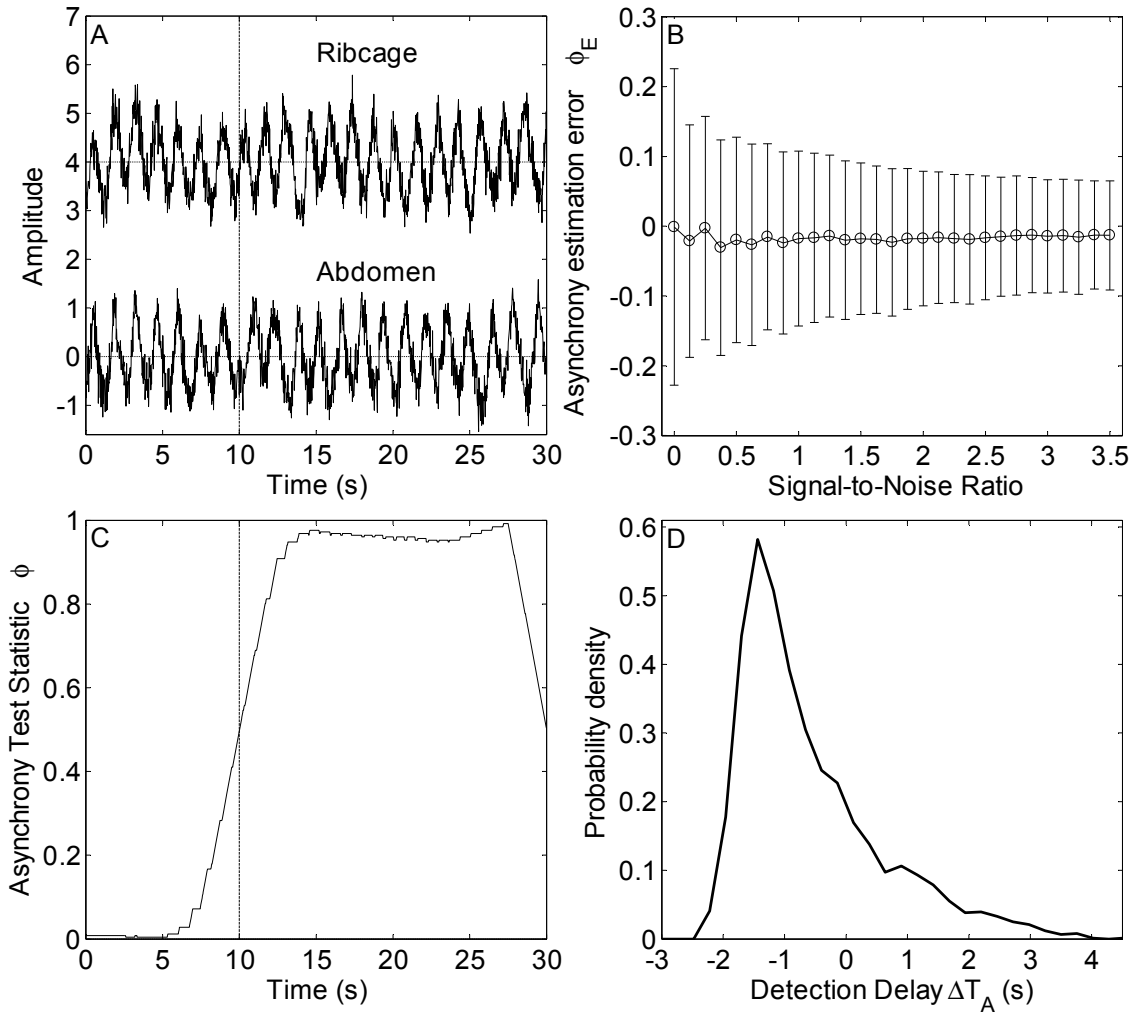

Fig. 6.6. Asynchrony detector performance: A. Simulated epoch with asynchrony starting at $t = 10s$, Ribcage and Abdomen are in arbitrary units and have been offset for clarity, B. Mean and standard deviation of the asynchrony estimation error (ϕ_E) as a function of SNR, C. Asynchrony test statistic ϕ for the epoch shown in A, D. Probability density of the asynchrony detection delay (ΔT_A).

To explore the detection delay the window length was set to $N_A = 251$, the transition parameter was $W_L = 10$, and we considered asynchrony as a phase between the RIP signals greater than 35° ($\gamma_A = 35/180$). This value is based on the results presented in [122], which reports asynchronies between abdomen and ribcage from 35° to 160°. We generated 5,000 realizations of equation (6.9), with $SNR = 5$ and $\rho_{13} = \rho_{23} = 0$, simulating quiet breathing without movement artifact. Each realization had a random asynchrony k, uniformly distributed between 35° and 180°, starting at a random time $T_A = n_0 / F_s$ uniformly distributed along the epoch. The detector performed very well; the overall probability of detection was 0.99 (40 misses in 5,040 realizations).

Fig. 6.6D shows the pdf of $\Delta T_A = T_{AD} - T_A$, the asynchrony detection delay as the difference between the time of detection T_{AD} and the true time of the asynchrony T_A. It demonstrates a very rapid detector response, within $-2.5 \le \Delta T_A \le 2.5$ seconds. This means that it takes at most $N_A/2$ samples, half the window length, to respond while most detections occurred much earlier.

6.6. Application to Infant Data

The simulation results presented above were very encouraging but we felt important to assess the performance of the method when applied to real data acquired from infants using RIP bands.

6.6.1. Description of Infant Data

The data comprised records acquired from 19 infants aged 44 ± 5 weeks (postconceptional age), weighing 4.0 ± 1.5 kg in the postoperative period after elective herniorrhaphy with general and caudal anesthesia. Written informed parental consent was obtained and the procedures were approved by the Institutional Ethics Review Board. These data were previously reported by Brown *et al.* [7, 71].

The ribcage and abdominal signals (Non-Invasive Monitoring Systems, Inc., Respitrace Plus, North Bay Village, Florida), were amplified, low-pass filtered at 15 Hz with 8-pole Bessel filters (Frequency Devices, Haverhill, MA), and sampled at $F_s = 50Hz$ with a 12-bit analog-to-digital converter (Data Translation, Marlborough, MA). Blood oxygen saturation and finger

plethysmograph signals where acquired with an oximeter (Nellcor N-200, Nellcor Inc., Hayward, CA). Data were stored on a computer using LABDAT^{TM} data acquisition software (RHT-InfoDat, Montreal, QC, Canada). After initial setup, the investigators did not supervise the recording session. No attempt was made to calibrate the signals in absolute terms.

6.6.2. Visual Scoring Analysis of Infant Data

In view of the time required for visual analysis, only a subset of 500 epochs for each of the 19 infant data sets was visually scored. This yielded a total of 9,500 epochs, which provided a "gold" standard for comparison with automated results. This was performed by one of the investigators (KAB) using an interactive, graphical visual scoring tool to mark the start and end times of segments comprising:

(i) Pause: Little or no respiratory movement in both the ribcage and abdomen signals,
(ii) Movement Artifact: Non-sinusoidal irregular signals,
(iii) Asynchrony: Asynchronous movement between the ribcage and abdominal signals,
(iv) Quiet Breathing: Quasi-sinusoidal breathing patterns in both the ribcage and abdominal signals.

Fig. 6.7 shows representative data for each category.

The data were scored in epochs of 30 s. The scorer was required to identify the start and end points of all events within the epoch. Thus, if an epoch contained 10 s of quiet breathing, a pause of 6 s, a movement period of 6 s and a final section of quiet breathing of 8 s, the scorer would identify 4 events. Events that spanned more than one epoch were concatenated into a single event. Table 6.2 shows the number of visually scored events in all data sets; there were a total of 10,928 events identified in 9,500 epochs.

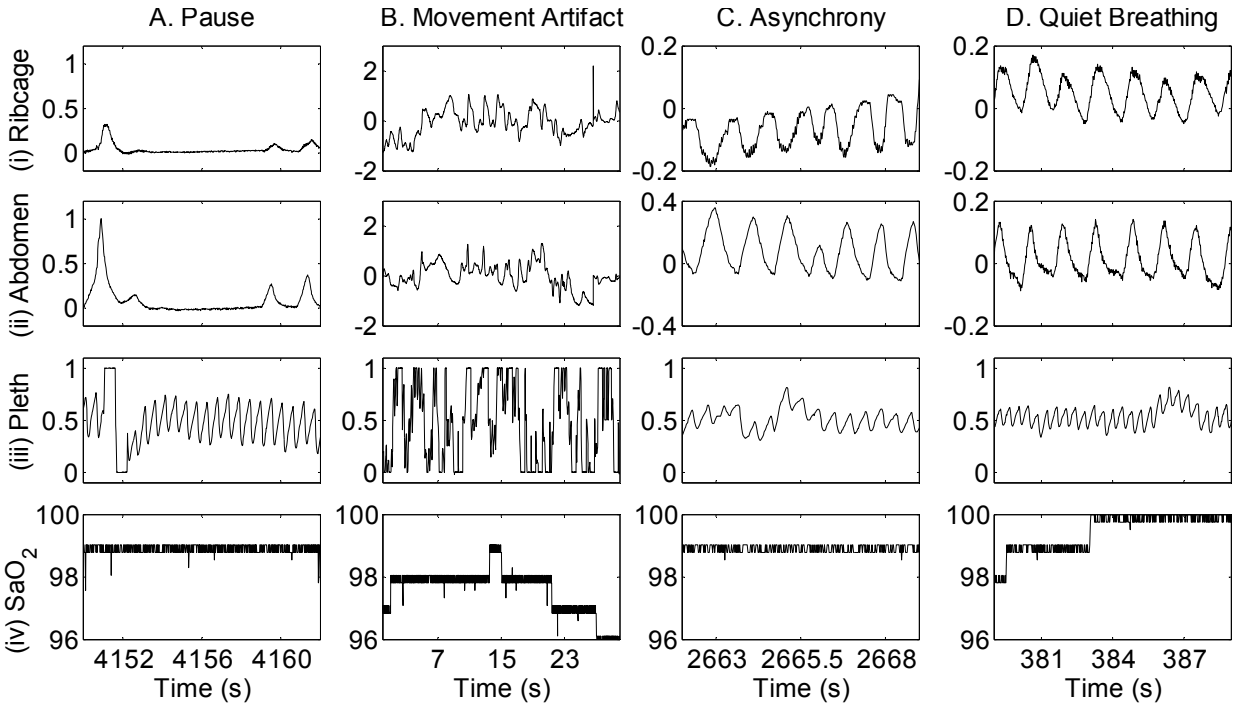

Fig. 6.7. Representative data for: A. Pause, B. Movement Artifact, C. Asynchrony, D. Quiet Breathing. From top to bottom the signals are: ribcage, abdomen, finger plethysmograph in arbitrary units, and blood oxygen saturation (in %).

Event	Number of Events
Quiet Breathing	4,539
Movement Artifact	2,753
Pause	2,480
Asynchrony	1,156

Table 6.2. Visually scored events.

6.6.3. Automated Scoring

We examined the relation between events visually annotated and those identified automatically by our detectors by performing a sample-by-sample comparison between the two results. To do so we estimated two probability density functions (pdf) for each test statistic; one for samples visually scored as containing the event and one for data scored as quiet breathing.

These probability densities were then used to generate the receiver operating characteristics (ROC) summarizing the performance of each test statistic as a function of the threshold. P_{FA} denotes the probability of false alarm, and P_D denotes the probability of detection for each test statistic. For example, from the pdf of the test statistic used to detect movement in the abdominal signal (m^{ab}) shown in Fig. 6.9A, P_D and P_{FA} for a threshold γ_M^{ab} can be found by solving:

$$P_D = \int_{-1}^{\gamma_M^{ab}} dF(m^{ab}, H_1^m),\ P_{FA} = \int_{-1}^{\gamma_M^{ab}} dF(m^{ab}, H_0^m), \tag{6.10}$$

Where H_1^m and H_0^m are the hypotheses of movement present or absent respectively. Each P_D and P_{FA} pair in the ROC plot corresponds to a unique threshold value. Thus, the choice of threshold determines the tradeoff between P_D and P_{FA}. For comparison purposes, we chose the threshold γ_{opt} for each detector that represented the point with the largest distance from the chance line ($P_D = P_{FA}$), as the best tradeoff between P_D and P_{FA}. This is shown in Figs. 6.8B, 6.9B and 6.10B, for pause, movement and asynchrony respectively. The area under the ROC curve (AUC) was also computed as a measure of detector's performance [84]. It is defined as:

$$AUC = \int_0^1 P_D\left(P_{FA}\right) dP_{FA}, \tag{6.11}$$

where $\{P_{FA}, P_D(P_{FA})\}$ defines the ROC curve. A value of $AUC = 1$ indicates perfect detection, while $AUC = 0.5$ represents the performance expected by chance.

6.6.4. Pause Detection

The pause test statistics (p^{ab} and p^{rc}) were calculated using a window size of $N_P = 51$. Fig. 6.8A shows the pdfs of p^{ab} for segments visually identified as quiet breathing and pause for all data. The ROC of p^{ab}, shown in Fig. 6.8B, had an $AUC = 0.82$ with an optimum threshold of $\gamma^{ab}_{P_{opt}} = 0.53$. For the ribcage signal the values were $AUC = 0.74$ and $\gamma^{rc}_{P_{opt}} = 0.49$.

6.6.5. Movement Artifact Detection

The movement artifact test statistics (m^{ab} and m^{rc}) were calculated using a window size of $N_M = 251$. Fig. 6.9A shows the pdfs of m^{ab} for segments visually identified as quiet breathing and movement artifact. The two pdfs are widely separated indicating that m^{ab} is a useful statistic to identify movement artifact in real infant data. The ROC plot for m^{ab} (in Fig. 6.9B) had an $AUC = 0.92$ and optimum threshold $\gamma^{ab}_{M_{opt}} = 0.22$. The values for the ribcage signal were $AUC = 0.85$ and $\gamma^{rc}_{M_{opt}} = 0.13$.

6.6.6. Asynchrony Detection

The asynchrony test statistic (ϕ) was calculated using a window size of $N_A = 251$. Fig. 6.10A shows the pdfs of ϕ for segments visually identified as quiet breathing and asynchrony. The pdfs of the asynchrony test statistic were not as well separated as those for the movement statistic. This is not surprising since the visual identification of asynchrony involves the subjective estimation of the phase difference between the RIP signals, a noisy process. This difference is reflected in the ROC curve in Fig. 6.10B, where $AUC = 0.88$ and the optimum threshold is $\gamma_{A_{opt}} = 0.32(58°)$.

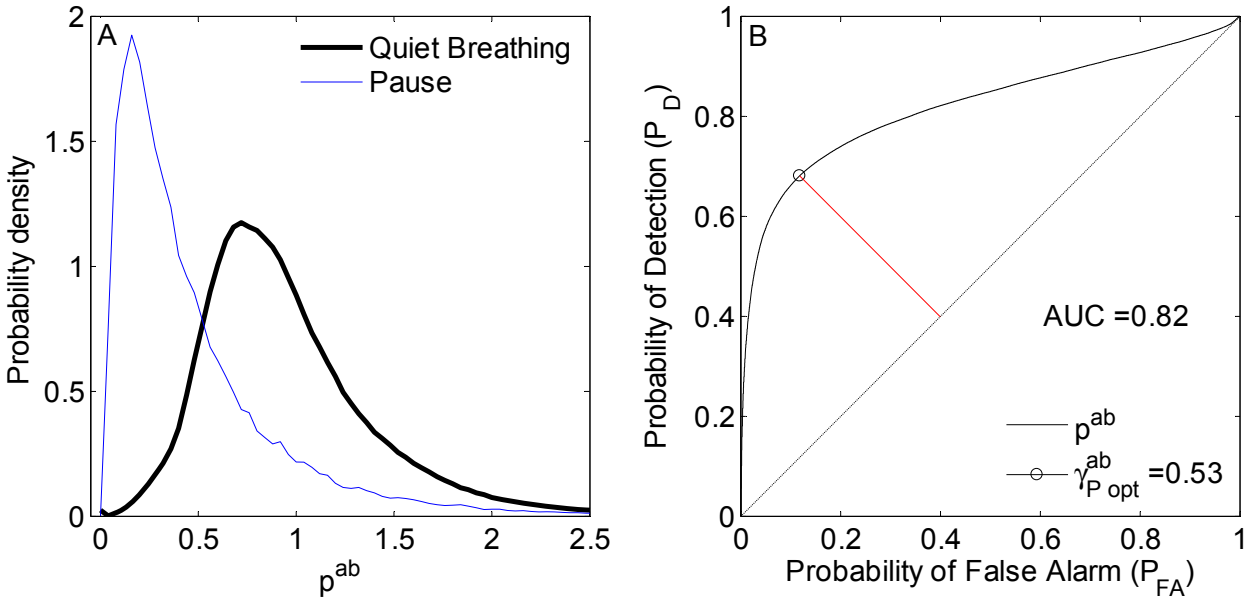

Fig. 6.8. Pause detection: A. Probability density of the pause test statistic for the abdomen (p^{ab}) for all patient data visually identified as quiet breathing and pause, B. Receiver Operating Characteristics (ROC) of p^{ab}, showing the optimum threshold ($\gamma^{ab}_{P_{opt}}$) and the Area Under the ROC curve (AUC).

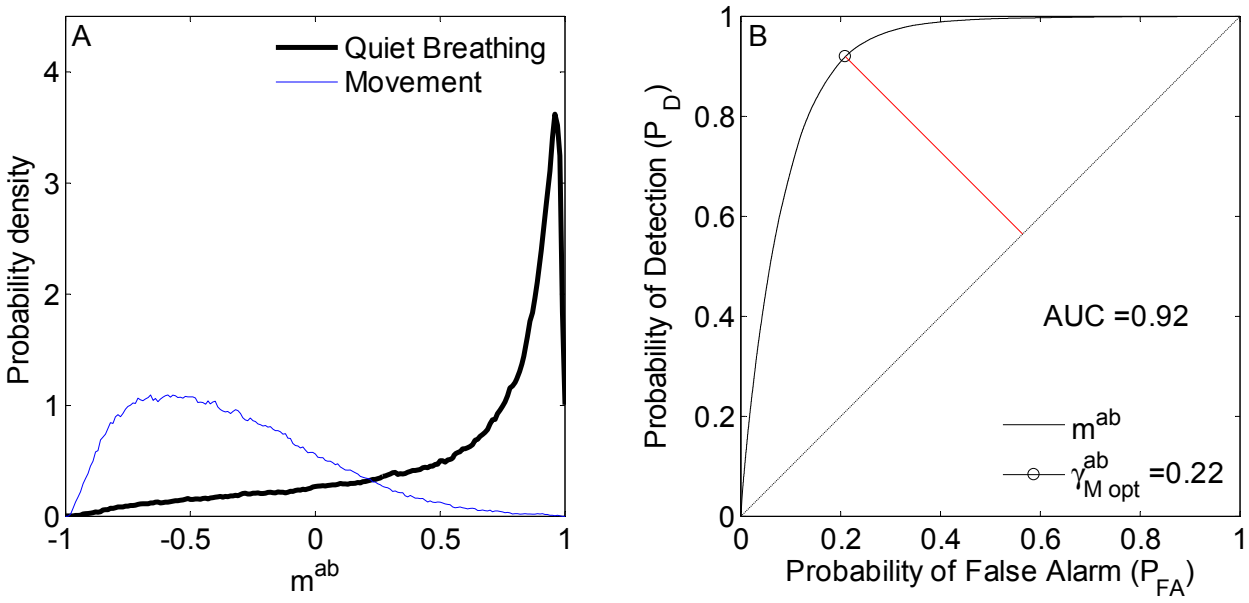

Fig. 6.9. Movement artifact detection: A. Probability density of the movement artifact test statistic for the abdomen (m^{ab}) for all patient data visually identified as quiet breathing and movement artifact, B. Receiver Operating Characteristics (ROC) of m^{ab}, showing the optimum threshold ($\gamma^{ab}_{M_{opt}}$) and the Area Under the ROC curve (AUC).

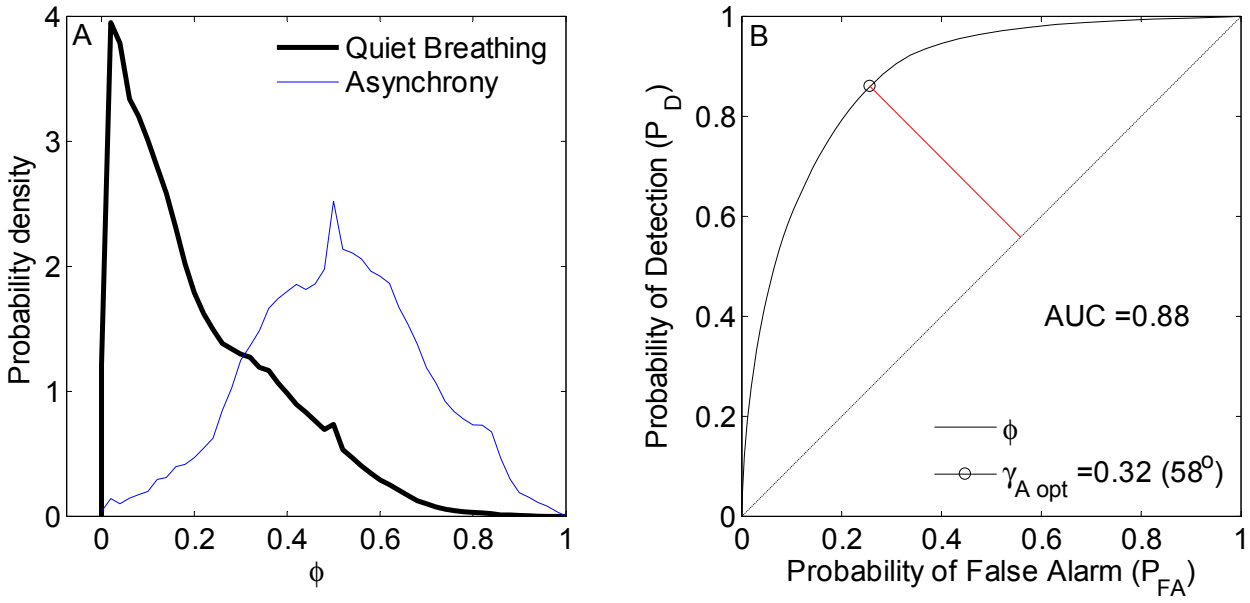

Fig. 6.10. Asynchrony detection: A. Probability density of the asynchrony test statistic (ϕ) for all patient data visually identified as quiet breathing and asynchrony, B. Receiver Operating Characteristics (ROC) of ϕ , showing the optimum threshold ($\gamma_{A_{opt}}$) and the Area Under the ROC curve (AUC).

6.6.7. Overall Performance

To evaluate the overall performance of our system, we verified the agreement between the state estimated by the automated scoring system and that scored visually, using the Fleiss' Kappa (κ) statistic [133] for inter-rater reliability. A value of $\kappa = 1$ indicates perfect agreement, while $\kappa = 0$ reflects the performance expected by chance.

We assessed the agreement in the scores given to each sample that had a single score by the clinician (i.e., scored as only one event type) in the 19 infant data sets. We obtained the overall κ value, and also the category specific agreement for each of the four main classes: Pause, Movement, Asynchrony and Quiet Breathing (QB).

For the detectors we selected the window sizes used before: Pause $N_P = 51$, Movement $N_M = 251$, Asynchrony $N_A = 251$. The thresholds were set to the optimum values determined in the previous sections.

The overall detectors for pause and movement defined in equations (6.3) and (6.7) respectively, combine the results obtained for the ribcage and abdominal signals using a logical AND. This is because both pauses and movement artifacts were expected to be manifested in both RIP signals. To test this assumption we evaluated four different methods of combining the ribcage and the abdomen detector outputs:

A: $P = P^{rc}$ and $M = M^{rc}$, where only rc is used to detect pause and movement;

B: $P = P^{ab}$ and $M = M^{ab}$, where only ab is used to detect pause and movement;

C: $P = \mathrm{OR}\left(P^{rc}, P^{ab}\right)$ and $M = \mathrm{OR}\left(M^{rc}, M^{ab}\right)$, where either rc or ab is used to detect pause and movement (logical OR);

D: $P = \mathrm{AND}\left(P^{rc}, P^{ab}\right)$ and $M = \mathrm{AND}\left(M^{rc}, M^{ab}\right)$, where rc and ab are used simultaneously to detect pause and movement (logical AND).

Table 6.3 shows the results. The best performance was obtained in scenario D, where the ribcage and the abdomen candidate detections were combined with a logical AND to define the pause and movement overall detectors, as in equations (6.3) and (6.7) respectively. Note that agreement

Scenario	Pause	Mvt	Asynch	QB	Overall
A	0.13	0.48	0.37	0.42	0.39
B	0.22	0.58	0.45	0.43	0.45
C	0.06	0.48	0.36	0.31	0.33
D	0.42	0.59	0.44	0.53	0.52

Table 6.3. Agreement (κ) Between Automated and Expert Scorer

for the asynchrony class changed with the method even though the asynchrony detector uses the ribcage and abdominal signals in all 4 methods. This is because both pause and movement have higher precedence as defined in Section 6.4.5.

6.7. Discussion

This paper describes and validates an automated method for detecting respiratory events in uncalibrated RIP data obtained from infants after surgery by combining detectors for pauses, movement artifacts, and asynchrony. Simulation studies demonstrated that the detector for each event distinguished them from normal breathing. Thus they detected most (> 99 %) of all simulated events and were robust in the presence of noise. Moreover, detection was timely: the pause detector identified most of the events within the first $1.2N_P$ samples (1.2 s); the asynchrony detector took from $-N_A/2$ to $N_A/2$ samples (i.e., from -2.5 s to 2.5 s) to detect the events; and the movement artifact detector needed twice the window length ($2N_M = 10\,\text{s}$) to identify most of the events. This longer latency is expected since movement artifact involves low frequencies.

To be clinically useful, any method for the automatic segmentation of respiratory data must distinguish between multiple states (e.g., pause, movement, asynchrony and quiet breathing). This is a more complex situation than the binary choice examined in our simulations. Therefore, we assessed the ability of our combined method to estimate the respiratory state for clinical data by evaluating the agreement between the visual scores provided by an expert clinician, and the respiratory state determined by our method. The overall agreement was $\kappa = 0.52$. This compares favorably to the agreement between human scorers. Thus, a study on scoring variability between expert technologists in sleep laboratories [9], found a κ of 0.31 among 11 scorers for a respiratory index, consisting in the total number of apneas plus hypopneas (pauses). A study in progress at our laboratory of postoperative apnea in infants, found the average agreement in respiratory event classification between two clinician scorers was $\kappa = 0.50$, while the overall agreement among three scorers was lower $\kappa = 0.47$. Moreover, the agreement between the three raters for asynchrony was $\kappa = 0.02$ (unpublished observation), while for the automated system was $\kappa = 0.44$. This is relevant for the differentiation between central and obstructive apneas.

The results we obtained by applying our method to clinical data are very promising, especially since we consider that the performance evaluation methodology was very conservative. Thus, we performed a sample-by-sample analysis which can be expected to exhibit lower probabilities of detection and overall agreement than would an event-by-event evaluation. For instance, if the clinician identified a segment as being 10 s long while the automated method detected an event of 5 s within that segment, then with a sample-by-sample analysis P_D would be equal to 0.5, while with an event-by-event evaluation P_D would be 1, since the event was indeed detected.

It is important to highlight that the data recording was unattended. Even though this source of uncertainty can impact the performance of the method, the results were satisfactory. This is an indication of robustness in data acquisition quality. Further work is required to assess the effect of supervised data recording on the overall agreement between our automated method and an expert scorer.

It should be noted that the scorer had access to the finger plethysmograph and the blood oxygen saturation signals which were not used by the automatic system. This might have an influence in the visual scores, which could have led to a lower agreement. Future studies are necessary to evaluate the relevance of using these two additional signals for the automated scoring of cardiorespiratory data.

As presented, the method is well suited for the off-line study of long data records such as studies of sleep disordered breathing. The detection parameters can be tuned with the scores from an expert clinician, eliminating the need to train additional human scorers. We believe that the automated detectors have the potential to provide more consistent and less subjective results than visual scoring. Thus, when visually scoring, the rater must judge if an event was present. This judgment can differ from scorer to scorer or indeed, from epoch to epoch with the same scorer. Further work, using multiple scorers to annotate data, rather than the single visual scorer in this study, will be required to confirm this.

The method presented here is easily customized. It allows the user to define threshold values for pause, movement artifact and asynchrony to his/her preference. The event detectors can then be

used in combination to identify apnea events to the user's preferred definition. For example, central apnea events might be defined as pauses of a given length. Hypopneas and apneas could be distinguished by using two threshold values for the pause detector, one for medium power (hypopnea), and one for low power (apnea). Obstructive apnea events might be defined as periods with some degree of asynchrony and low power RIP signals for a given period of time. Future studies to explore the method's effectiveness in detecting apneas of specific definitions and lengths, as well as differentiating between central and obstructive apneas are indicated. Note also that even though the reported results are very satisfactory, further work is required to determine the optimal detector window length and filter bank frequency resolution for a given application.

It should also be possible to use the methods in an on-line near real-time use. The only parameter that depends on the complete recording is the term $\wp^{ab}$ from equation (6.1) which defines the power associated with quiet breathing. For real-time use, an initial estimate could be obtained from a short training segment (e.g., 10 min), and then adaptively updated thereafter.

Postoperative apnea (POA) events are rare and so large data sets are required to determine how they relate to other respiratory events. The acquisition of such a large database has not been feasible because visual scoring is labor intensive, expensive, and complicated by low inter-scorer agreement. Thus, we regard one of the primary contributions of the present work as the development of an automated, reliable and repeatable means of analyzing data sets to provide high quality estimates of breathing pauses and asynchronies, while identifying segments corrupted by movement artifacts. This will make it possible to evaluate simultaneously the clinical implications of these respiratory events on infants at risk of POA.

Our methods should also help resolve the clinical significance of the degree of asynchrony. The notion of a continuum from synchrony (0^o) to complete paradox (180^o) was introduced in [117]. The optimum threshold for asynchrony from our ROC analysis, $\gamma_{A_{opt}} = 0.32(58°)$, is consistent with this notion. In [161] they reported little asynchrony ($9^o \pm 3^o$) in healthy term infants, in contrast to preterm infants whose value was $38^o \pm 9^o$. Inspiratory loading in the preterm infant increased the asynchrony to $56^o \pm 7^o$ and was associated with a decrease in respiratory frequency.

Therefore, even small asynchrony increments may be of clinical importance, possibly indicating maturational changes in the respiratory system and potentially linked to respiratory compromise.

Our methods could also be used for the objective definition of cardiorespiratory events. We suggest building a consensus of scores, based on the threshold values defined by expert visual scorers from different laboratories for the pause, movement artifact and asynchrony detectors. For the specific case of asynchrony detection, the use of an automated method as a "gold" standard should be considered, given the lack of agreement reported among scorers, compared to the results presented in this study.

In summary, we have presented a method for the automatic detection of pause, movement artifact corruption and paradoxical respiratory movement in infant RIP data. The main advantages of this method are: (i) it provides full automation and simple implementation; (ii) it is more efficient than visual scoring; (iii) the analysis is repeatable and standardized; (iv) it produces greater agreement with an expert scorer than that between trained scorers; (v) it is amenable to on-line detection; and (vi) it is applicable to uncalibrated RIP signals. This last point is of great importance since the Qualitative Diagnostic Calibration method for RIP has been shown to be limited by changes in measurement conditions and breathing patterns [28].

Symbol	Description
rc , ab	Raw RIP signals from Ribcage and Abdomen
rc_i , ab_i	Output from the ith filter in the bank for rc and ab
rc_{bp} , ab_{bp}	Band-pass filtered rc and ab
rc_S , ab_S	Selectively filtered version of rc and ab
RC , AB	Simulated RIP signals for Ribcage and Abdomen
$f_{\max}$	Breathing frequency estimate
$i_{\max}$	Filter from the bank with the output with highest power
P , M , A	Overall detectors for Pause, Movement and Asynchrony
P^{rc} , P^{ab}	Pause detectors used with rc and ab
M^{rc} , M^{ab}	Movement detectors used with rc and ab
γ_P^{rc} , γ_P^{ab}	Thresholds used with P^{rc} and P^{ab}
γ_M^{rc} , γ_M^{ab}	Thresholds used with M^{rc} and M^{ab}
γ_A	Threshold used with A
$\gamma_{P_{opt}}^{rc}$, $\gamma_{P_{opt}}^{ab}$	Optimum thresholds for P^{rc} and P^{ab}
$\gamma_{M_{opt}}^{rc}$, $\gamma_{M_{opt}}^{ab}$	Optimum thresholds for M^{rc} and M^{ab}
$\gamma_{A_{opt}}$	Optimum threshold for A
p^{rc} , p^{ab}	Pause test statistics for rc and ab
m^{rc} , m^{ab}	Movement test statistics for rc and ab
ϕ	Asynchrony test statistic
N_P , N_M , N_A	Window lengths used with p^{rc} , p^{ab} , m^{rc} , m^{ab} , and ϕ
P_D , P_{FA}	Probabilities of detection and false alarm
$\wp^{rc}$, $\wp^{ab}$	Median power of all segments of length N_P in rc_f and ab_f
$\wp_i^{rc}$, $\wp_i^{ab}$	Power of a segment of length N_M in rc_i and ab_i
F_s	Sampling Frequency
T_n	Simulated signals period
$\tilde{m}_1$, $\tilde{m}_2$	Simulated movement process in RC and AB
g_1 , g_2	Simulated electronic noise in RC and AB
w	Transition window for simulation of asynchrony
W_L	Length parameter for w
n_0	Sample where simulation of asynchrony begins
k	Degree of simulated asynchrony
ρ_{11} , ρ_{12} , ρ_{13}	Scaling parameters used to construct RC
ρ_{21} , ρ_{22} , ρ_{23}	Scaling parameters used to construct AB
T_P , T_M , T_A	Start time for simulated pause, movement and asynchrony
T_{PD} , T_{MD} , T_{AD}	Actual detection time of pause, movement and asynchrony
ΔT_P , ΔT_M , ΔT_A	Detection delay for pause, movement and asynchrony
ϕ_{true}	Actual simulated asynchrony
ϕ_E	Asynchrony estimation error

Table 6.4. Symbols

7. Automated Unsupervised Analysis of Infant Respiratory Patterns

7.1. Preface

Chapter 6 presented AORED, an Automated Off-Line Respiratory Event Detector that performs repeatable, reliable, comprehensive, fast, and low-cost analyses of infant respiratory patterns. However, it is a supervised classifier that must be trained with the results of the manual analysis of a representative data set. Thus, although it addresses many of the limitations of manual scoring, its application still requires substantial manual analysis that is labor intensive, time consuming, and prone to scorer variability. Moreover, the results are still subjective because they reflect the particular biases of the scorers who carried out the manual analysis used for training.

This Chapter describes my development of a fully automated method to analyze infant respiratory patterns that is based on unsupervised classification and so requires no human intervention. This Automated Unsupervised Respiratory Event Analysis (AUREA) system estimates several metrics of respiratory behavior from infant data, and uses K-means clustering to classify them into the respiratory patterns. AUREA's results agree very well with those of the "gold standard" manual analysis described in Chapter 5, and are more accurate than those of AORED. AUREA provides a fully automated, reliable, and repeatable analysis of the respiratory patterns that is completely objective, low-cost, fast, and involves no human judgments. I believe that AUREA will be of great value as a research tool; moreover, AUREA can be readily implemented for real-time use and so could also be used clinically.

This Chapter is a manuscript to be submitted for publication as:

C. A. Robles-Rubio, K. A. Brown, and R. E. Kearney, "Automated Unsupervised Analysis of Infant Respiratory Patterns," to be submitted to *IEEE Trans Biomed Eng*.

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada. The work of C. A. Robles-Rubio was supported in part by the Mexican National

Council for Science and Technology. C. A. Robles-Rubio and K. A. Brown were supported in part by the Queen Elizabeth Hospital of Montreal Foundation Chair in Pediatric Anesthesia.

7.2. Abstract

We previously presented AORED, an Automated Off-line Respiratory Event Detector of infant respiratory patterns from uncalibrated respiratory inductive plethysmography signals. The application of AORED requires a training set comprising a sample of respiratory patterns classified by expert manual scorers. Manual analysis is labor intensive, time consuming, and involves the subjective judgment from the experts which may bias the automated classifier. To address these problems we developed a novel method for Automated Unsupervised Respiratory Event Analysis (AUREA). This paper describes the algorithm underlying AUREA, and demonstrates its successful application to respiratory signals acquired from infants recovering from general anesthesia by comparing it to a "gold standard" manual scoring. AUREA has the following advantages: (i) it is much faster than the "gold standard" manual scoring; (ii) it is fully automated and requires no human intervention, so it is low-cost and objective; (iii) its output has perfect consistency and agrees substantially with that of the "gold standard" manual analysis; (iv) it is significantly more accurate than AORED and has no detection delay; (v) it assigns a respiratory pattern to every sample of infant respiratory data in a repeatable, reliable fashion; and (vi) it is amenable for real-time classification of respiratory patterns.

7.3. Introduction

Infants with postmenstrual ages (PMA) of 60 weeks or less are at increased risk of life threatening apnea following surgery and anesthesia [15, 17, 39]. The first postoperative apnea (POA) event generally occurs within the first 12 hours after surgery, and POA events may continue to occur up to 72 hr postoperatively [14, 39]. Thus, current clinical guidelines suggest that infants with PMA $\leq$ 60 weeks undergoing surgery and anesthesia should be monitored continuously in hospital for extended periods following surgery [14].

There is evidence suggesting that POA events are related to abnormal postoperative respiratory patterns [14, 19, 24, 39]. However, this relation has not been studied comprehensively because there is no appropriate method to do so systematically. Thus, conventional manual scoring (CMS) is currently the preferred analysis method for respiratory data, which requires an expert

scorer to scroll through the data and visually detect "clinically relevant" events based on guidelines published by the American Academy of Sleep Medicine (AASM) [8]. This method is labor intensive, time consuming, expensive, subjective, and produces results with low intra- and inter scorer repeatability [9].

We recently developed a set of manual scoring tools for the analysis of infant respiratory patterns, whose use substantially improves repeatability amongst scorers compared to CMS [10]. These tools also enable the comprehensive scoring of all samples in a data record, rather than only segments deemed as "clinically relevant". We then developed a method to accurately and consistently estimate a "gold standard" scoring of respiratory patterns, by using Expectation-Maximization (EM) to combine the results from multiple scorers optimally. Use of this EM algorithm significantly improves the accuracy and consistency of scores.

Even though this approach produces results with good repeatability, manual scoring remains very labor intensive, time consuming, and expensive. For this reason, it is highly desirable to automate the analysis. To this end, we developed AORED, an Automated Off-line Respiratory Event Detector to automatically classify respiratory patterns, and tested it on data acquired from infants at risk of POA [85]. AORED's results agreed well with the manual analysis from an expert scorer, but its use still requires a sample of manually analyzed data for classifier training. There are two main problems associated with the manual analysis required for training: (i) even though only a sample of the data needs to be analyzed, manual scoring it is still labor intensive, time consuming, and expensive; and (ii) AORED will incorporate any biases and subjective judgments produced by the reference, training scorer(s).

The objective of this work was to develop an automated method for the analysis of infant respiratory patterns that eliminates the need for manually scored data. Some aspects of this paper have been a part of a conference presentation [135].

This Chapter is organized as follows: Section 7.4 presents the Automated Unsupervised Respiratory Event Analysis system (AUREA) that we developed for fully automated analysis of the respiratory patterns; Section 7.5 describes the clinical dataset and "gold standard" manual

analysis, and presents the methods for evaluation of AUREA; Section 7.6 reports the evaluation results; Section 7.7 discusses the findings; and Section 7.8 provides some concluding remarks.

7.4. Automated Unsupervised Respiratory Event Analysis System (AUREA)

7.4.1. Overview

AUREA was designed to automatically classify infant respiratory patterns observed in respiratory inductive plethysmography (RIP) signals into one of 5 types: respiratory pause (PAU), synchronous-breathing (SYB), asynchronous-breathing (ASB), movement artifact (MVT), and unknown (UNK). These patterns were previously identified in RIP signals from infants at risk of postoperative apnea (POA) in [10, 85]. Fig. 7.1 is a schematic of the two analysis stages AUREA applies to RIP signals to yield this classification. In the first stage, AUREA estimates a set of metrics describing respiratory behavior from the RIP signals. These metrics extract information about amplitude, frequency, and phase of the ribcage (RCG) and abdomen (ABD) respiratory movements on a sample-by-sample basis. In the second stage, these metrics are used by K-means classifiers to yield an estimate of the instantaneous respiratory pattern.

7.4.2. Metrics of Respiratory Behavior

This section describes the metrics of respiratory behavior that AUREA uses to classify the respiratory patterns in RIP signals.

7.4.2.1. Trend Removal

The RIP signals were pre-processed to remove any low frequency trends which are common in RIP signals [5]. This trend is estimated as the mean of RCG_{raw} over a two-sided, sliding window of length N_{DT}, i.e.,

$$RCG_T[n] = \frac{1}{N_{DT}} \sum_{i=n-(N_{DT}-1)/2}^{i=n+(N_{DT}-1)/2} RCG_{raw}[i], \tag{7.1}$$

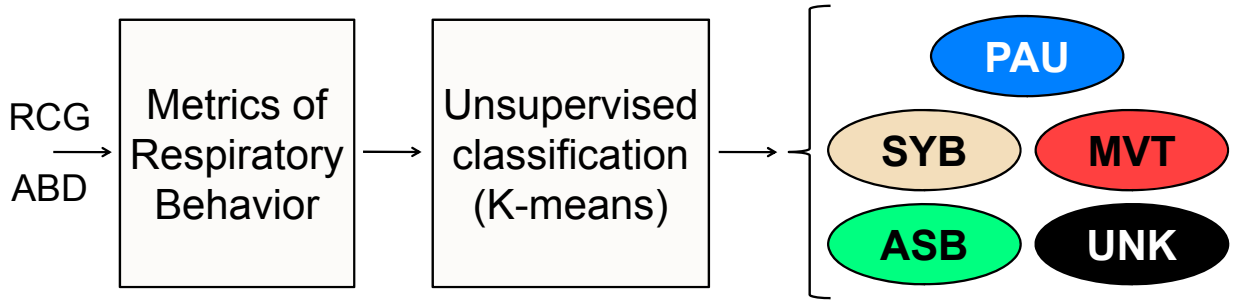

Fig. 7.1. Automated Unsupervised Respiratory Event Analysis (AUREA) overview. Metrics of respiratory behavior are estimated from the ribcage (RCG) and abdomen (ABD) respiratory signals. Then these metrics are used by a series of K-means classifiers to classify the respiratory patterns into the following: respiratory pause (PAU), synchronous-breathing (SYB), asynchronous-breathing (ASB), movement artifact (MVT), or unknown (UNK).

and the de-trended signal is computed as

$$RCG[n] = RCG_{raw}[n] - RCG_T[n]. \tag{7.2}$$

7.4.2.2. Pause Metric

The pause metric is intended to discriminate the PAU pattern, defined as segments where the RCG and ABD signals have amplitudes less that 10 % of those of the preceding normal breath [10]. It is expected that RIP signal variance will be lower during PAU than during any other respiratory pattern. Thus, the pause a metric is based on sample-by-sample variance estimates.

The estimation procedure for the the RCG signal (RCG_{raw}) comprised the following steps, but a similar metric was used for the ABD signal (ABD_{raw}).

(i) The variance of RCG is estimated sample-by-sample, over a two-sided, sliding window of length $N_V \ll N_{DT}$ as

$$v_{RCG}[n] = \frac{1}{N_V} \sum_{i=n-(N_V-1)/2}^{n+(N_V-1)/2} RCG^2[i]. \tag{7.3}$$

(ii) The variance estimate is normalized to $v_{RCG}^{(q)}$, the q^{th} quantile of v_{RCG} in the most recent N_{QV} samples ($N_{QV} \gg N_V$), as:

$$nv_{RCG}[n] = \ln\left(\frac{v_{RCG}[n]}{v_{RCG}^{(q)}[n]}\right). \tag{7.4}$$

This normalization is intended to compensate for nonstationarities in the amplitude of RCG due to postural changes and/or slight displacement of the respibands.

7.4.2.3. Movement Artifact Metric

The second metric is designed to detect the MVT pattern, defined as segments where RCG_{raw} and ABD_{raw} display a chaotic, non-sinusoidal, low frequency motion associated with active or passive movement of the infant [10]. This is a non-periodic power metric that we first used to detect

MVT in photoplethysmography (PPG) signals [138]. This metric is based on two observations: (i) the artifact-free signal has a quasi-periodic waveform; and (ii) MVT comprises stochastic, low frequency noise whose amplitude is larger than that of the artifact-free signal [3, 5]. Since both these assumptions should hold for RIP signals we felt that the same MVT metric could be used. The MVT metric is estimated using the following steps.

(i) Moving-Average, Notch Filtering: The de-trended RCG signal (RCG) is filtered by a moving average, low-pass filter of length N_{MA}. This filter has deep nulls at integer multiples of f_s/N_{MA}, with f_s being the sampling frequency. N_{MA} is chosen so that these nulls occur at the respiratory frequency and its harmonics. Consequently this filter will attenuate periodic components related to respiration, pass other lower frequencies, and attenuate high frequency noise.

(ii) The root mean square (RMS) of the moving-average filtered signal RCG_{MA}, is computed over a two-sided, sliding window of length N_{RMS} as

$$rms_{RCG}[n] = \sqrt{\frac{1}{N_{RMS}} \sum_{i=n-(N_{RMS}-1)/2}^{n+(N_{RMS}-1)/2} RCG_{MA}^2[i]}\,. \tag{7.5}$$

(iii) The RMS is normalized in a similar fashion to the PAU metric. Thus, the non-periodic power MVT metric is

$$npp_{RCG}[n] = \ln\left(\frac{rms_{RCG}[n]}{rms_{RCG}^{(q)}[n]}\right), \tag{7.6}$$

where $rms_{RCG}^{(q)}$ is the q^{th} quantile of rms_{RCG} in the most recent N_{QRMS} samples ($N_{QRMS} \gg N_{RMS}$).

Similar ABD_{MA}, rms_{ABD}, $rms_{ABD}^{(q)}$, and npp_{ABD} are obtained for ABD.

7.4.2.4. Synchronous and Asynchronous-Breathing Metrics

The last metrics are designed to detect SYB and ASB. Breathing is defined as quasi-sinusoidal patterns in both RCG and ABD [10]. During SYB, RCG and ABD movements are in phase (<

90^o), and during ASB these movements are out of phase ($\geq 90^o$). The metrics to detect SYB and ASB are described in detail in a previous paper [137]. Briefly, they are estimated as follows.

(i) The de-trended RCG signal (RCG) is first smoothed to reduce additive noise using a two-sided, sliding window of length $N_{SMO} \ll N_{DT}$ as

$$RCG_{SMO}[n] = \frac{1}{N_{SMO}} \sum_{i=n-(N_{SMO}-1)/2}^{n+(N_{SMO}-1)/2} RCG[i], \tag{7.7}$$

and then converted to binary signal as

$$RCG_B[n] = \begin{cases} 1 & if \quad RCG_{SMO}[n] > RCG_T[n] \\ 0 & otherwise \end{cases}, \tag{7.8}$$

where RCG_T is the low-frequency trend from the raw RCG signal (RCG_{raw}) estimated using equation (7.1). In other words, RCG_B is set to 1 if the value of RCG_{SMO} is above the trend, and to 0 otherwise. The binary ABD signal (ABD_B) is also estimated similarly.

(ii) The sum and difference of the binary signals is computed as

$$SUM[n] = (RCG_B + ABD_B)/2,$$

and

$$DIF[n] = (RCG_B - ABD_B)/2.$$

When breathing is completely synchronous, SUM will oscillate between 0 and 1 at around the respiratory frequency while DIF will stay constant at 0 [137]. Conversely, when breathing is asynchronous, SUM will remain constant at 0.5 while DIF will oscillate between -0.5 and 0.5 at around the respiratory frequency [137].

(iii) SUM and DIF are high-pass filtered at a cut-off frequency of 0.5 Hz to extract the power associated with breathing (for details see [137]). In infants, most respiratory-related power

lies in the frequency band 0.4 Hz to 2.0 Hz [4, 85]. In contrast, PAU and MVT occur at the lower band 0 Hz to 0.4 Hz [5, 85].

(iv) The power of the high-pass filtered sum signal, SUM_{HP}, is estimated over a two-sided, sliding window of length N_B as

$$b^{+}[n] = \frac{1}{N_B} \sum_{i=n-(N_B-1)/2}^{n+(N_B-1)/2} SUM_{HP}^{2}[i]. \tag{7.9}$$

The values of the SYB metric, b^{+}, will be large for SYB and tend to zero otherwise [137].

The ASB metric, b^{-}, is the power of the high-pass filtered difference signal, DIF_{HP}. The values of b^{-} tend to zero during patterns other than ASB, and higher during ASB [137].

7.4.3. Unsupervised Classification of Respiratory Patterns

AUREA applies K-means clustering [162] to these metrics to automatically classify each sample into one of 5 categories: PAU, MVT, SYB, ASB, and UNK.

7.4.3.1. Sample Unbalance and Decision Boundary Adjustment

Initially, K-means with Euclidean distance was used to classify our data into one of 4 patterns (PAU, MVT, SYB, and ASB) applying all metrics as inputs. This provided acceptable results when the number of samples for each pattern was similar (i.e., the data set was balanced). However, infant respiratory patterns are heavily unbalanced since PAU is rare (< 5 % of the data) while SYB is common (> 60 %) [10]. We observed that for such an unbalanced data set the clusters produced by K-means misclassified a significant number of the samples belonging to the most prevalent category.

Fig. 7.2 illustrates this showing the decision boundary obtained by K-means (solid, red line) when discriminating between samples with "gold standard" classification of PAU against non-PAU. Using this decision boundary 11 % of the PAU samples (located to the top and right of the solid, red line in Fig. 7.2B), and 11 % of the non-PAU samples (from the red line to the bottom and left in Fig. 7.2C) were misclassified. However, in this example PAU represented only 15 %

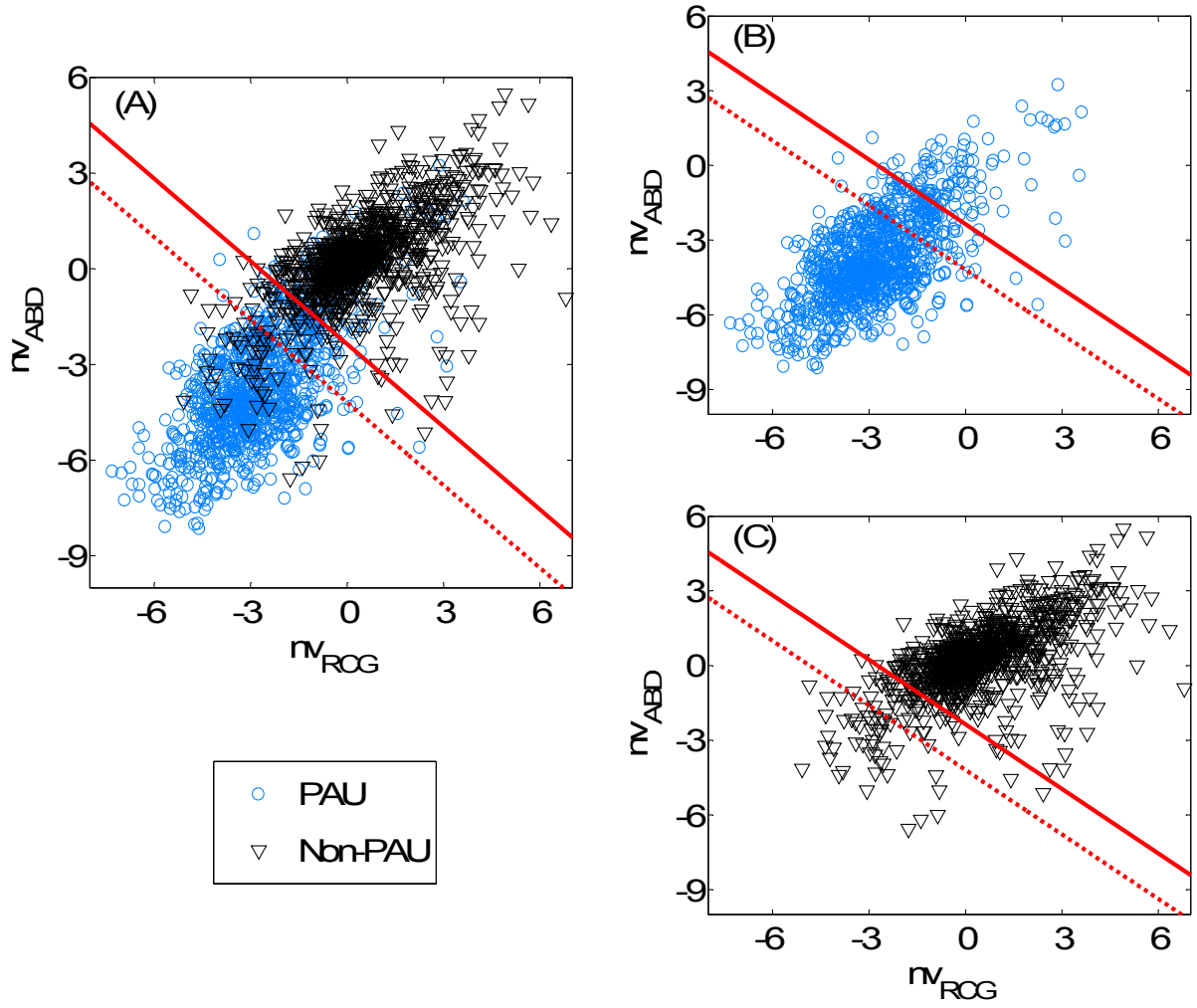

Fig 7.2. (A) Representative samples of infant data with "gold standard" classification of pause (PAU), or non-PAU, scattered as a function of the normalized variance of ribcage nv_{RCG}, and abdomen nv_{RCG}. (B) Detail of only PAU samples from panel (A). (C) Detail of only non-PAU samples from panel (A). The solid, red line corresponds to the original decision boundary determined by K-means, and the dotted, red line to the boundary adjusted for unbalanced sampling.

of the total data, while non-PAU was the remaining 85 %. This meant that misclassified PAU samples corresponded to only 1.65 % of the total data, while misclassified non-PAU samples corresponded to a much higher 9.35 %, which resulted in a substantial loss of overall classification accuracy.

Consequently, we adjusted the K-means decision boundaries to mitigate the effects of the unbalanced data.

We identified that in a dataset with P input metrics, the K-means decision boundary between two clusters, C_j and C_m, forms a hyperplane containing the point $\boldsymbol{\gamma}_{jm} \in \mathbb{R}^P$ with normal vector $\mathbf{v}_{jm} \in \mathbb{R}^P$,

$$\begin{aligned} \mathbf{v}_{jm} &= \mathbf{c}_m - \mathbf{c}_j, \\ \boldsymbol{\gamma}_{jm} &= w_{jm}\mathbf{v}_{jm} + \mathbf{c}_j, \end{aligned} \tag{7.10}$$

where $c_j \in \mathbb{R}^P$ and $c_m \in \mathbb{R}^P$ are the centroids of C_j and C_m respectively, and $w_{jm} = 0.5$ is the decision boundary weighting factor that determines the proportion of the Euclidean space covered by each cluster. The assignment of sample $\mathbf{x}[n] \in \mathbb{R}^P$ to cluster C_j, termed $L\{\mathbf{x}[n]\}$, is determined as

$$L\{\mathbf{x}[n]\} = C_j \leftrightarrow \mathbf{v}_{jm} \cdot (\mathbf{x}[n] - \boldsymbol{\gamma}_{jm}) < 0, \forall m \neq j\,. \tag{7.11}$$

To adjust for the sample unbalance, we modified decision boundary weighting factor to reflect the relative proportion between C_j and C_m as

$$w_{jm} = \frac{w_j}{w_j + w_m}, \tag{7.12}$$

where w_j and w_m were the proportion of samples belonging to C_j and C_m respectively once K-means had converged. This re-weighting effectively shifted the decision boundary allowing the cluster with more samples to cover more space, mitigating the effect of sample unbalance.

Fig. 7.2 illustrates the original and adjusted decision boundaries using the example from above. Using the adjustment resulted in a net decrease of total misclassifications (from 11 % to 7 %), which was produced by reducing the amount of misclassified non-PAU samples (from 11 % to 4 %), at the expense of increasing the amount of misclassified PAU samples (from 11 % to 24 %).

7.4.3.2. Training

AUREA is trained using K-means to determine the classification parameters automatically following the next steps.

(i) Metric Outliers Detection: The metrics may attain unreasonable values during UNK segments where the signal quality is bad. For example, an absent signal may have a very low variance, which is much lower than that of PAU. K-means tends to create very small clusters for such outlying samples, rather than focusing on the data of interest. To avoid this, samples corresponding to metric outliers are removed from the training data. Thus, all samples with metric values below the $\alpha_{metric}/2$ quantile, and above the $1-\alpha_{metric}/2$ quantile are excluded. In this work this parameter was set to $\alpha_{metric}=0.001$, since we expected only a small fraction of our data to be corrupted since data acquisition was continuously attended to ensure high quality in the recordings.

(ii) Input metrics with large scale or great variability strongly affect the results of K-means [163]. To control for this, it is necessary to standardize the metrics. To this end, the median and interquartile range (IQR) of each metric are estimated from the training data. Then, each metric is standardized by subtracting its median, and dividing by its IQR.

(iii) Determination of Classification Parameters: Shifting the decision boundaries for all clusters simultaneously to adjust for unbalanced sampling will generate an uncertainty region in the input space, where instances will not be assigned to any cluster [135]. To avoid this, AUREA uses 4 binary K-means classifiers instead of a single, multi-class K-means implementation. These classifiers are implemented in series, as illustrated in Fig. 7.3, and each decision boundary is adjusted separately. The detailed steps of this implementation are:

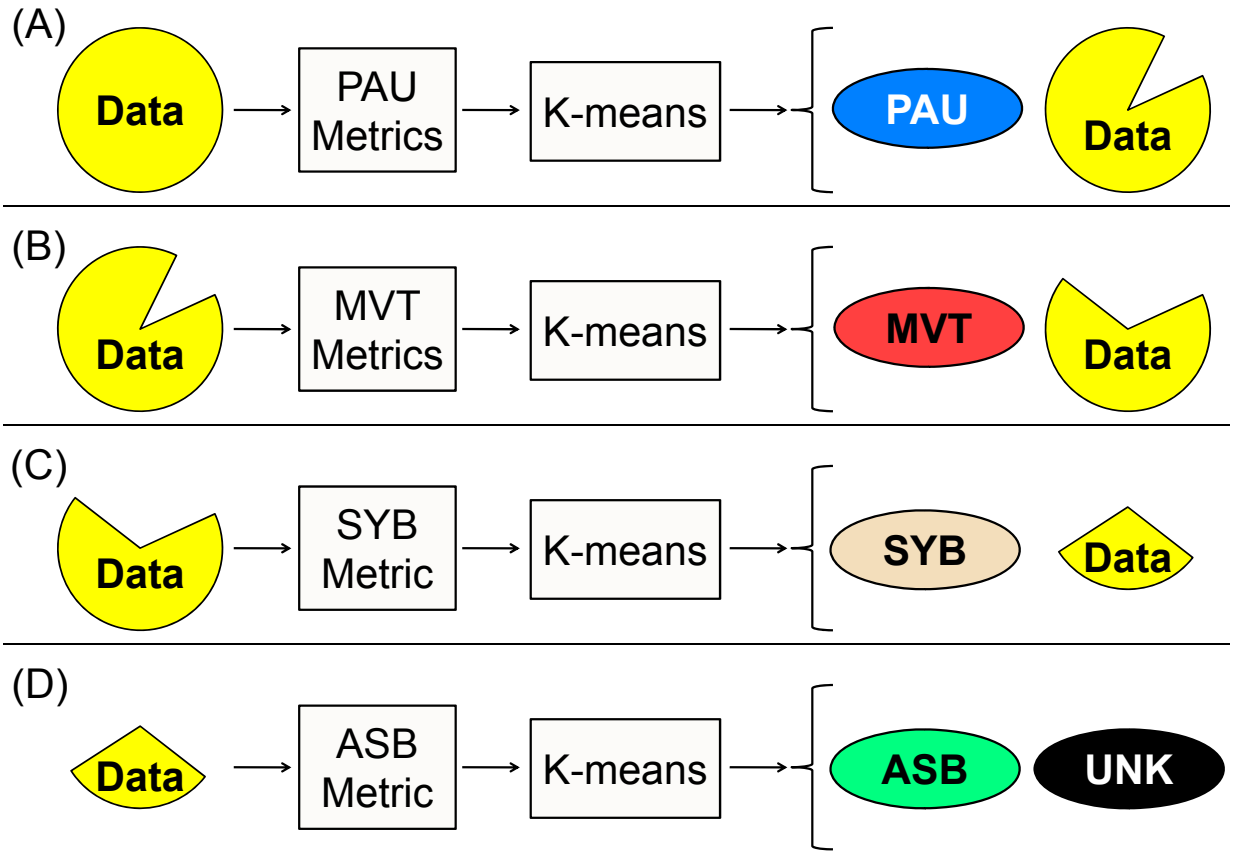

Fig. 7.3. Automated Unsupervised Respiratory Event Analysis (AUREA) training. (A) The pause (PAU) metrics from the training data are input to K-means to detect PAU samples. (B) The movement artifact (MVT) metrics from the remaining data are input to K-means to detect MVT samples. (C) Then, the synchronous-breathing (SYB) metric is used with K-means to detect SYB samples in the remaining data. (D) Finally, the asynchronous-breathing (ASB) metrics is input to K-means to discriminate between ASB and unknown (UNK) samples. Classification parameters of the 4 K-means classifiers are stored to use with new data.

(A) Identify PAU samples. A K-means with 2 clusters is applied to the PAU metrics to discriminate between PAU and all other patterns. The starting point for each cluster is listed in Table 7.1. The decision boundary is adjusted using equations (7.12) and (7.10). Samples belonging to the adjusted PAU cluster are removed from the training data set.

(B) Identify MVT samples. K-means with 2 clusters is applied to the MVT metrics to discriminate between MVT and the remaining patterns. Samples classified as MVT are removed.

(C) Identify SYB samples. Apply K-means with 2 clusters to the SYB metric to discriminate between SYB and the remaining patterns. Adjust the decision boundary, and remove samples classified as SYB.

(D) Identify ASB samples. Apply K-means with 2 clusters to the ASB metric to discriminate between ASB and UNK. Remove samples classified as ASB.

(E) Classify all remaining samples as UNK.

(F) Store the metric standardization parameters (i.e., median and IQR of each metric), and the classification parameters for each of the 4 K-means classifiers to use with new data. These classification parameters comprise the point on the adjusted decision boundary $(\boldsymbol{\gamma}_{jm})$ and the vector normal to this boundary $(\mathbf{v}_{jm})$.

The precedence of patterns was determined based on our previous work [85]. PAU was classified ahead of MVT because PAU data could be erroneously classified as MVT. This is because there is very little respiratory power during PAU, and even a low-power, low-frequency signal could trigger the MVT classifier. MVT was classified second because MVT may have components in the respiratory frequency band, which could cause false positive classification of SYB or ASB. SYB was classified third because the SYB metric showed better detection performance than the ASB metric in our previous work [137].

7.4.3.3. Classification

Once training is completed, new data can be classified using the following steps to yield a signal representing the respiratory pattern at each time.

Pattern of Interest	Metrics Used	Cluster Starting Points		Justification of Starting Points
		Pattern of Interest	*All Others*	
PAU	nv_{RCG} and nv_{ABD}	$\begin{bmatrix}\min(nv_{RCG})\\ \min(nv_{ABD})\end{bmatrix}$	$\begin{bmatrix}\max(nv_{RCG})\\ \max(nv_{ABD})\end{bmatrix}$	Metrics tend to lower values during PAU, and higher values for other patterns.
MVT	npp_{RCG} and npp_{ABD}	$\begin{bmatrix}\max(npp_{RCG})\\ \max(npp_{ABD})\end{bmatrix}$	$\begin{bmatrix}\min(npp_{RCG})\\ \min(npp_{ABD})\end{bmatrix}$	Metrics tend to higher values during MVT, and lower during other patterns.
SYB	b^{+}	$\max(b^{+})$	$\min(b^{+})$	Metric tends to higher values during SYB, and to zero otherwise.
ASB	b^{-}	$\max(b^{-})$	$\min(b^{-})$	Metric tends to higher values during ASB, and to zero otherwise.

Table 7.1. Clustering parameters. Starting points were selected based on the properties of each metric. PAU = Respiratory pause, MVT = movement artifact, SYB = synchronous-breathing, ASB = asynchronous-breathing.

(i) Estimate the PAU, MVT, SYB, and ASB metrics from the RCG and ABD signals.

(ii) Standardize each metric using their median and IQR values obtained during training. To do this, subtract the metric median and divide by its IQR.

(iii) Classify samples using the procedure in Fig. 7.3 and the classification parameters determined during training.

(iv) Identify outliers in each adjusted cluster and reclassify them as UNK. Samples that are outliers in a cluster have characteristics different from other samples in the cluster, so they should be classified as UNK. To implement this, the adjusted PAU, MVT, SYB, and ASB clusters are represented as functions of the 6 input metrics. The centroid of each cluster is estimated as the median of the metrics for all samples within each cluster, and the Euclidean distance of each sample to its cluster centroid is computed. Samples whose distance is greater than the $1-\alpha_{cluster}$ quantile are deemed to be outliers and classified as UNK. The cluster outlier detection parameter was set to $\alpha_{cluster} = 0.001$.

(v) Patterns are assigned to all samples to generate a continuous signal. The minimum expected length of an infant breath is 0.5 s, since infant respiratory frequencies are limited to the band 0.4 Hz to 2.0 Hz [85]. Therefore, assigned pattern segments less than 0.5 s long are removed to eliminate multiple segments of very short length and reduce fragmentation. This operation results in short segments with no pattern assigned, and so interpolation is necessary to reclassify their samples. To do this, the first half of the segment is assigned the pattern of the preceding segment, and the second half is assigned the pattern of the following segment.

7.5. Performance Evaluation

This section describes the methods we used to evaluate the performance of AUREA, including the clinical dataset, "gold standard" analysis, a classifier from the literature [85] for comparison, the cross-validation setup, and a number of performance parameters. Note that clinical data and manual scoring analyses used to evaluate performance are described only briefly. They have been described in detail previously [10], and are freely available from the Dryad Digital Repository (doi:10.5061/dryad.72dk5) [11].

7.5.1. Infant Data and Manual Analysis

Data were acquired from 21 infants at risk of postoperative apnea (16 male, birth age 31 ± 4 weeks, postmenstrual age 43 ± 2 weeks, weight 3.6 ± 1.0 kg) immediately after surgery in the postanesthesia care unit of the Montreal Children's Hospital. The study was approved by the Institutional Review Board of the McGill University Health Centre / Montreal Children's Hospital. Written, informed parental consent was obtained for each infant recruited.

The signals acquired were ribcage (RCG) and abdomen (ABD) respiratory inductive plethysmography (RIP), as well as photoplethysmography (PPG) and blood oxygen saturation (SAT) from an oximeter taped to a digit. Signals were low-pass filtered at 10 Hz, sampled at $f_s = 50$ Hz, and stored. No attempt was made to calibrate the RIP signals.

Three scorers with varied backgrounds were recruited and trained to analyze the data. Each data record was truncated to a maximum of 20,000 s, and each scorer analyzed the 21 truncated data sets twice in independent, randomly ordered instances [10]. Thus, there were 6 sequences describing the respiratory patterns of each infant at each time.

During manual analysis, scorers assigned data samples to one of 6 unique, mutually exclusive respiratory patterns [10]: synchronous-breathing (SYB), asynchronous-breathing (ASB), sigh (SIH), respiratory pause (PAU), movement artifact (MVT), or unknown (UNK).

7.5.2. "Gold Standard"

The 6 manual scoring sequences were combined using Expectation-Maximization (EM) as described in Chapter 5 to yield a "gold standard" classification of the respiratory patterns. Briefly, an initial estimate of the "gold standard" was obtained using the majority vote, i.e., each sample was scored as the pattern assigned/voted by the majority. The confusion matrix of each scoring sequence was then estimated using the initial "gold standard" as reference. After that, the "gold standard" was re-estimated by weighting the votes in each scoring sequence by its estimated confusion matrix. The final estimate of the "gold standard" was obtained by iterating this process until convergence.

7.5.3. Supervised Classifier

The performance of AUREA was compared to that of the Automated Off-line Respiratory Event Detector (AORED) presented in [85]. AORED compared metrics of respiratory behavior to thresholds to individually detect PAU, MVT, or ASB patterns. These metrics were similar in concept to those used in AUREA, although implemented quite differently [85].

Training. To determine AORED thresholds, the "gold standard", manual classification was used to estimate two nonparametric probability density functions (PDF) for each metric: one for samples classified as the pattern of interest (e.g., PAU, MVT, or ASB), and one for samples classified as SYB. Regarding the metrics as pattern detectors, the PDFs were used to generate the Receiver Operating Characteristics (ROC) curves relating the probability of detection (P_D) to the probability of false alarm (P_{FA}) as a function of the threshold. The value of the threshold for each metric was selected to provide the best tradeoff between P_D and P_{FA}, defined as the point in the ROC curve farthest from the chance line (this line is where $P_D = P_{FA}$, and corresponds to the performance of a coin-toss classifier).

Classification. The respiratory pattern was determined by combining the output of the individual pattern detectors with the following precedence: PAU had the highest priority so when PAU was detected the other pattern detectors were forced to zero. MVT was assigned the second level of precedence with the output of the ASB detector forced to zero when MVT was detected. ASB had the third level of precedence. Samples with no pattern assigned were scored as SYB. This method was not designed to classify UNK samples.

7.5.4. Cross-validation

Leave-one-patient-out cross-validation was used to evaluate the generalization ability and repeatability of AUREA and AORED. Fig. 7.4 illustrates the cross-validation procedure. The dataset was split in three disjoint parts: (i) 1 record was used for testing, (ii) 19 records were kept for training, and (iii) 1 record was excluded. Thus, for a given testing record, it was possible to have 20 different training sets by keeping 19 of the remaining 20 records in the training data and excluding 1 at a time. This allowed the evaluation of the classifiers under varied training

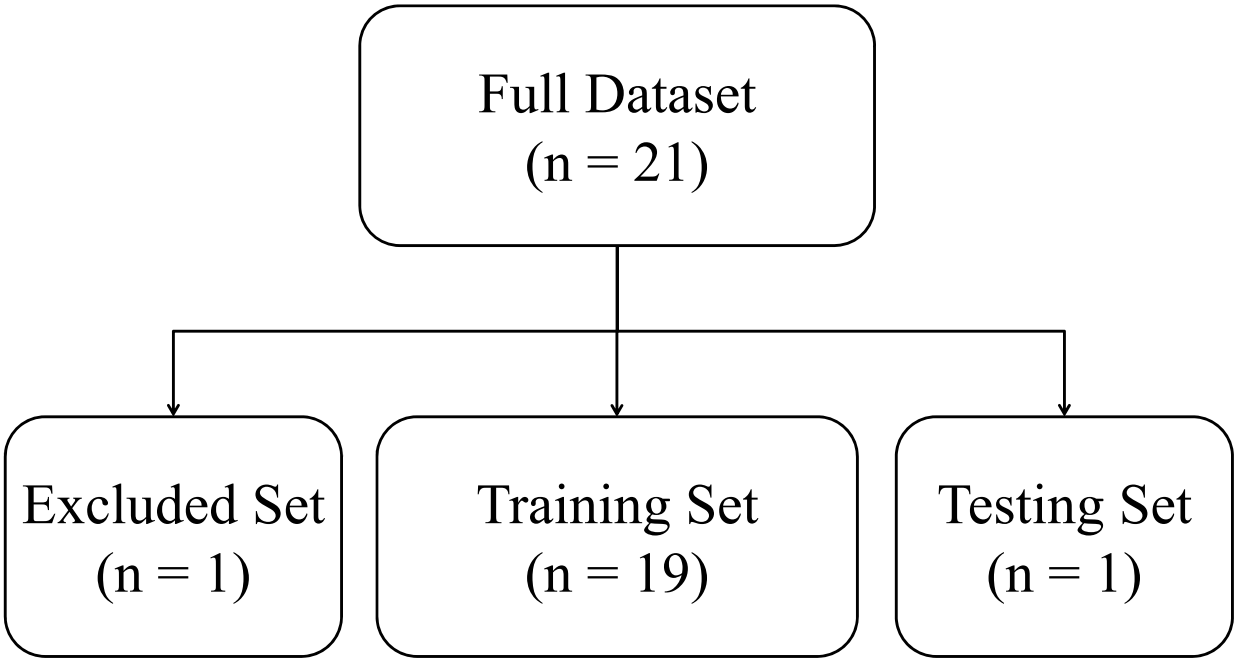

Fig 7.4. Cross-validation setup. The full dataset was split into three sets: (i) training, with 19 data records, (ii) testing, with 1 record, and (iii) excluded, with 1 record.

conditions, while still maximizing the amount of data used for training. Classification parameters for both AUREA and AORED were obtained using the training set, and then these parameters were used to classify the testing set. The process was repeated for all possible combinations of data records in the testing and excluded sets, giving 20 different classifier instances for AUREA and 20 for AORED to be used with each testing record.

7.5.5. Performance Evaluation Parameters

We compared AUREA to AORED in terms of the ability to reproduce the "gold standard" classification. The following sections describe the evaluated parameters. All results were obtained from the testing performance obtained with cross-validation.

7.5.5.1. Preliminary Considerations

Segments scored as MVT in the "gold standard" are data corrupted by non-respiratory movements. These data should be excluded in studies using the analysis of respiratory patterns to obtain physiological information. A similar argument applies to segments scored as UNK, since they have high noise, bad signal quality, or ambiguous patterns. For this reason, we joined MVT and UNK to form a single pattern category termed "bad data" (BAD).

Additionally, AUREA does not support classification of SIH, so all samples scored as SIH were excluded from all analyses. This amounted to < 2 % of the total dataset.

The values of the parameters used to estimate AUREA's respiratory metrics are described in Table 7.2. Parameters for AORED were set as indicated in [85].

7.5.5.2. Statistical Analysis

Significant differences in median values were assessed using the Wilcoxon rank-sum test [141] with a p-value < 0.01 considered to be statistically significant.

7.5.5.3. Detection Performance

AUREA and AORED were used to classify all samples in the test set and the probabilities of detection (P_D) and false alarm (P_{FA}) for each pattern were determined. $P_D = 1$ indicates that all samples were correctly classified, and $P_D = 0$ indicates that no samples were correctly classified.

Metric Type	Parameter	Value	Rationale
PAU	q	0.5	In [10] we observed that infants have a SYB pattern more than 60 % of the time. Thus, we considered that a normalization quantile equal to the median should provide an appropriate reference to normal SYB values on a normalization window 2 min long.
	N_{QV}	2 min	
	N_V	1 s	To allow for a fast response and the detection of short pauses.
	N_{DT}	5 s	Based on previous work [85, 138].
MVT	N_{MA}	1.42 s	The length of the moving average filter was selected such that the filter nulls were at harmonics of the most frequent respiratory frequency, defined by the mode of the respiratory frequency histogram computed from the entire clinical data set. This frequency was 0.7 Hz (i.e., 42 breaths per minute).
	q	0.01	Based on previous work [85, 138]
	N_{QNPP}	10 min	
	N_{RMS}	5 s	
	N_{DT}	5 s	
SYB and ASB	N_{SMO}	0.42 s	Based on previous work [137]
	N_B	2 s	
	N_{DT}	2 s	

Table 7.2. Parameter Selection for Metrics of Respiratory Behavior

$P_{FA} = 0$ indicates that no samples were incorrectly classified, while $P_{FA} = 1$ shows that all negative samples were false positives. Thus, the ideal detection performance corresponds to $P_D = 1$ and $P_{FA} = 0$.

The P_D for PAU was estimated as

$$P_D^{PAU} = \frac{TP^{PAU}}{P^{PAU}}, \tag{7.13}$$

where TP^{PAU} was the number of samples correctly classified by the automated method as PAU, and P^{PAU} was the total number of samples "gold standard" scored as PAU (i.e., positives). Similar P_D^{SYB}, P_D^{ASB}, and P_D^{BAD} were estimated for SYB, ASB, and BAD respectively.

The P_{FA} for PAU was estimated as

$$P_{FA}^{PAU} = \frac{FP^{PAU}}{N^{PAU}}, \tag{7.14}$$

where FP^{PAU} was the number of samples incorrectly classified as PAU (i.e., false positives), and N^{PAU} was the number of samples with "gold standard" score not equal to PAU (i.e., negatives). Similar P_{FA}^{SYB}, P_{FA}^{ASB}, and P_{FA}^{BAD} were estimated for SYB, ASB, and BAD respectively.

D-values, $d = P_D - P_{FA}$, were estimated for each classifier to provide a single performance parameter. The *d*-value measures the normalized distance of any point on the ROC curve from the chance line [139]. A value of $d = 0$ corresponds to a P_D and P_{FA} combination that lies on the chance line, while $d = 1$ indicates perfect classification.

7.5.5.4. Accuracy and Consistency

The cross-validation procedure generated 20 different classification sets for each of the 21 test data records, yielding a total of 420 classification sets. Accuracy was measured as the agreement between the classifier and the "gold standard" for each classification set. Consistency was measured as the agreement among the 20 classification sets produced for the same test data record.

Agreement was assessed on a sample-by-sample basis using the Fleiss' κ statistic [133] for inter-scorer agreement. This κ implementation generalizes the traditional Cohen's statistic [136] to evaluate agreement between multiple scorers when classifying observations into two or more categories. Values of κ were interpreted according to the intervals proposed in [140]: $\kappa < 0$: poor, $0 \leq \kappa \leq 0.2$: slight, $0.2 < \kappa \leq 0.4$: fair, $0.4 < \kappa \leq 0.6$: moderate, $0.6 < \kappa \leq 0.8$: substantial, and $0.8 < \kappa \leq 1$: almost perfect. Results are listed as: Median [1st Quartile, 3rd Quartile].

7.5.5.5. Detection Delay

To assess the delay with which AUREA and AORED detected each pattern, we identified all "gold standard" classified segments, defined as sets of contiguous samples with the same assigned pattern for a minimum length of 0.5 s. Each segment was described by 3 properties: (i) pattern type, (ii) start time (T_S), and (iii) end time (T_E). Fig. 7.5A illustrates these T_S and T_E with an example.

A "gold standard" segment was considered to have been detected by a classifier if more than 50 % of its samples were assigned the correct pattern.

The detection start $\left(\hat{T}_S\right)$ and end $\left(\hat{T}_E\right)$ times of AUREA and AORED were estimated for each "gold standard" segment. Fig. 7.5B illustrates $\hat{T}_S$ and $\hat{T}_E$ with examples. $\hat{T}_S$ was defined by the first sample that was correctly classified, and either it was part of the "gold standard" segment, or it occurred before the "gold standard" segment but was contiguous to a detected segment. $\hat{T}_E$ was determined by the last sample that was classified correctly, and either it was part of the "gold standard" segment, or it occurred after the "gold standard" segment but was contiguous to a detected segment.

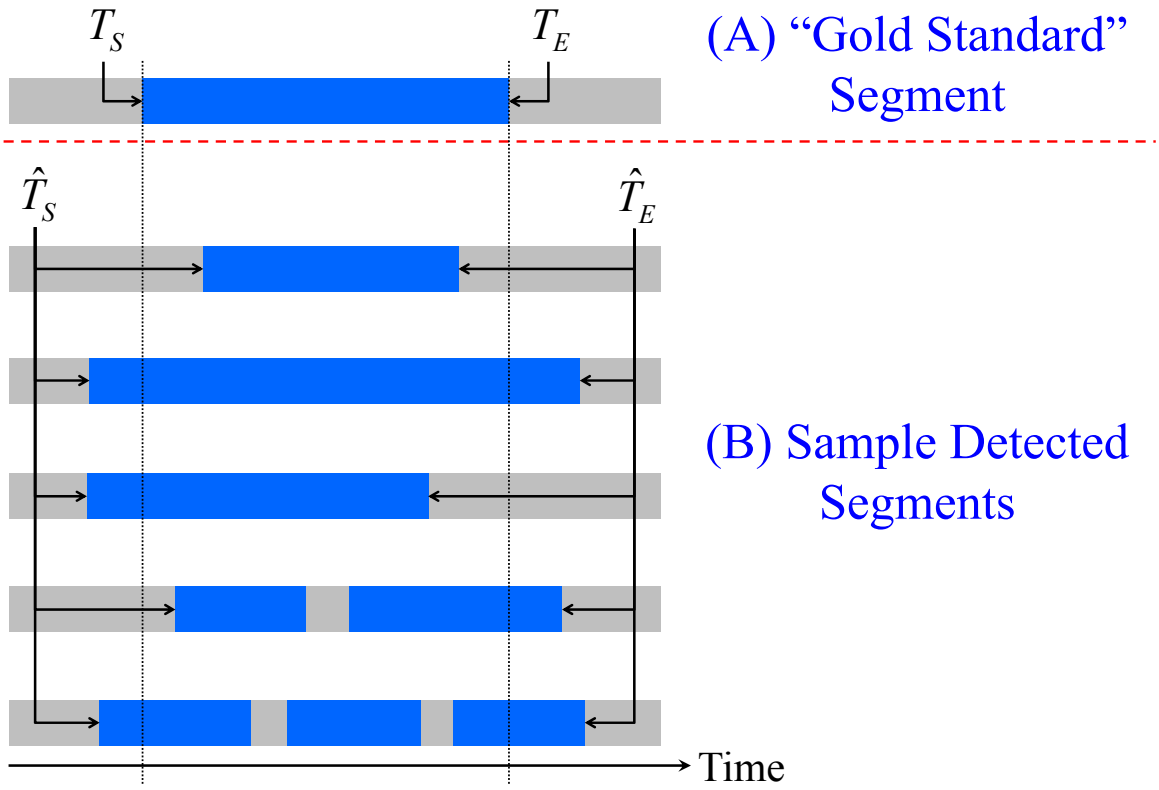

Fig. 7.5. Segment properties. (A) "Gold standard" segment. (B) Sample segments detected by an automated classifier. T_S = start time, T_E = end time, $\hat{T}_S$ = detection start time, and $\hat{T}_E$ = detection end time.

The start delay was estimated as $\Delta T_S = \hat{T}_S - T_S$; a positive ΔT_S indicated that the segment was detected late, and a negative ΔT_S meant that the segment was detected ahead of time. The end delay was estimated similarly as $\Delta T_E = \hat{T}_E - T_E$; a positive delay indicated that the segment was terminated late, while a negative value showed that the segment was terminated prematurely.

7.6. Results

This section reports the results of the performance evaluation, comparing the ability of AUREA and AORED to replicate the classification provided by the "gold standard".

7.6.1. Detection Performance

Fig. 7.6 shows the *d*-values for AUREA and AORED. AUREA performed better than AORED for all patterns. The difference was greatest for PAU, where AUREA had a $d = 0.73$ and AORED a $d = 0.53$. The second greatest difference was in SYB, with a $d = 0.78$ for AUREA and $d = 0.70$ for AORED. The differences in ASB and BAD were ≤ 0.03.

Table 7.3 shows the P_D and P_{FA} values of AUREA and AORED for all respiratory patterns. All AUREA P_D values were greater than 0.7. P_D^{PAU} and P_D^{SYB} were significantly higher for AUREA than AORED. P_D^{BAD} reached statistical significance, but the values were very similar as evidenced by the equal medians and similar quartiles. In the case of P_{FA}, differences between AUREA and AORED reached statistical significance, but these differences were small.

7.6.2. Accuracy and Consistency

Fig. 7.7 shows the accuracy and consistency results. AUREA had substantial accuracy (κ = 0.68 [0.64, 0.69]), and this accuracy was significantly higher (*p*-value < 0.01) than that of AORED (κ = 0.6 [0.56, 0.62]), as expected from the differences in *d*-values. The two classifiers had almost perfect consistency, with κ values of 0.98 or higher.

7.6.3. Detection Delay

Table 7.4 shows the start and end delays for AUREA and AORED. Overall, AUREA was always as good as or better than AORED. Both AUREA and AORED had negligible (i.e., $|\Delta T_S|$ and

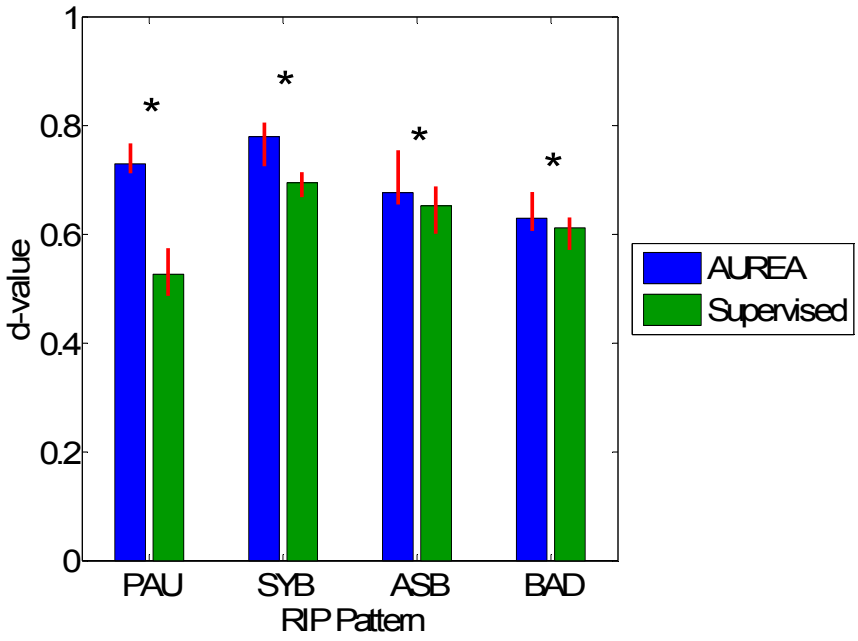

Fig. 7.6. Comparison of AUREA (blue) and AORED (green) in terms of their *d*-values (i.e., the probability of detection minus probability of false alarm). Bars indicate the median, and the red error bars span the interquartile range. An '*' indicates a p-value < 0.01. PAU = respiratory pause, SYB = synchronous-breathing, ASB = asynchronous-breathing, BAD = bad data.

Parameter	AUREA	AORED	p-value
P_D^{PAU}	0.77 [0.74, 0.81]	0.55 [0.51, 0.62]	< 0.01
P_D^{SYB}	0.86 [0.83, 0.88]	0.78 [0.75, 0.83]	< 0.01
P_D^{ASB}	0.71 [0.69, 0.80]	0.71 [0.65, 0.76]	0.05
P_D^{BAD}	0.76 [0.72, 0.79]	0.76 [0.72, 0.80]	< 0.01
P_{FA}^{PAU}	0.04 [0.03, 0.04]	0.02 [0.02, 0.03]	< 0.01
P_{FA}^{SYB}	0.07 [0.06, 0.09]	0.08 [0.07, 0.11]	< 0.01
P_{FA}^{ASB}	0.03 [0.03, 0.05]	0.05 [0.04, 0.08]	< 0.01
P_{FA}^{BAD}	0.11 [0.11, 0.13]	0.15 [0.11, 0.20]	< 0.01

Table 7.3. Probabilities of Detection (P_D) and False Alarm (P_{FA}) of the automated classifiers of respiratory patterns. Results presented as: Median [1st Quartile, 3rd Quartile]. PAU = respiratory pause, SYB = synchronous-breathing, ASB = asynchronous-breathing, BAD = bad data.

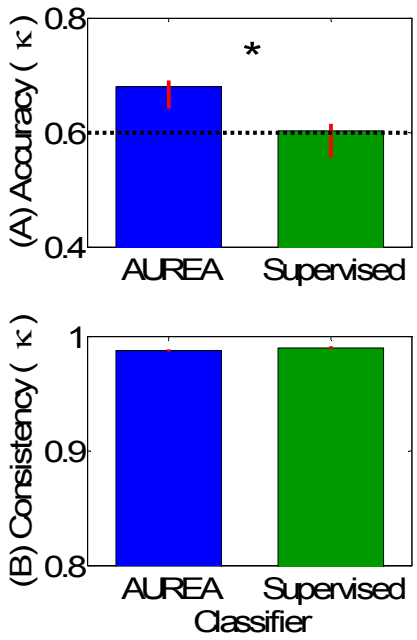

Fig. 7.7. (A) Accuracy and (B) consistency of AUREA (blue), and AORED (green). Bars indicate the median, and the red error bars span the interquartile range. An '*' indicates a p-value < 0.01. The black-dotted line in (A) indicates the limit between moderate and substantial accuracy.

Parameter	Pattern	AUREA	AORED	p-value
ΔT_S (s)	*Overall*	-0.08 [-0.78, 0.36]	-0.22 [-2.14, 0.42]	< 0.01
	PAU	0.00 [-0.26, 0.20]	0.06 [-0.18, 0.34]	< 0.01
	SYB	0.12 [-0.42, 1.00]	0.04 [-0.80, 1.48]	< 0.01
	ASB	-0.16 [-0.88, 0.28]	-0.58 [-3.10, 0.22]	< 0.01
	BAD	-0.84 [-2.22, -0.02]	-1.88 [-4.58, -0.40]	< 0.01
ΔT_E (s)	*Overall*	-0.04 [-0.46, 0.64]	0.18 [-0.44, 2.20]	< 0.01
	PAU	-0.12 [-0.32, 0.14]	-0.12 [-0.36, 0.10]	< 0.01
	SYB	-0.20 [-1.02, 0.28]	-0.12 [-1.50, 0.68]	< 0.01
	ASB	0.10 [-0.42, 0.82]	0.62 [-0.22, 3.04]	< 0.01
	BAD	0.64 [-0.16, 2.22]	1.96 [0.38, 4.80]	< 0.01

Table 7.4. Start (ΔT_S) and end (ΔT_E) detection delays of the automated classifiers of respiratory patterns. Results presented as: Median [1st Quartile, 3rd Quartile]. PAU = respiratory pause, SYB = synchronous-breathing, ASB = asynchronous-breathing, BAD = bad data.

$|\Delta T_E|$ < 0.5 s) overall median start and end delays. This indicates that there was no systematic error for either classifier when considering all respiratory patterns combined. The main difference between the classifiers was in the interquartile range (IQR), defined as the difference of the 3rd quartile minus the 1st quartile. A shorter IQR indicates higher precision, a desirable property. AUREA overall delays were very close to the median with IQR values of around 1.1 s, less than half those of AORED, which were around 2.5 s.

In the pattern-specific evaluation on Table 7.4, AUREA showed negligible median start and end delays for PAU, SYB and ASB. AORED behaved similarly for PAU and SYB only, but showed significantly higher delays for ASB; it started segments 0.58 s ahead of time, and terminated them 0.62 s late. AUREA detected BAD segments 0.84 s early, and terminated them 0.64 s late. This represented a median lengthening of 1.48 s per BAD segment. This was significantly shorter than that estimated by AORED, which started BAD segments 1.88 s early, and terminated them 1.96 s late, representing a total median extension of 3.84 s.

AUREA was more precise than AORED for SYB, ASB, and BAD. The IQR values of AUREA were between 1.6 (ΔT_S of SYB) and 2.9 (ΔT_S of ASB) times lower than those of AORED. The precision of both classifiers was similar for PAU (IQR values were around 0.5 s).

7.7. Discussion

This work presented AUREA, a novel, completely automated method to classify respiratory patterns from respiratory inductive plethysmography (RIP) signals. AUREA classifies respiratory patterns sample-by-sample into one of 5 types: respiratory pause (PAU), synchronous-breathing (SYB), asynchronous-breathing (ASB), movement artifact (MVT), and unknown (UNK). AUREA has the following advantages:

(i) it is much faster than the "gold standard" manual scoring;
(ii) it is fully automated requiring no human intervention, so it is low-cost and objective;
(iii) it comprehensively classifies the respiratory patterns on a sample-by-sample basis;
(iv) it performs significantly better than AORED, a previous Automated Off-line Respiratory Event Detector [85];

(v) it has perfect consistency, and substantial accuracy when compared to the "gold standard", which makes it a repeatable, reliable means to comprehensively analyze infant respiratory patterns; and

(vi) it is amenable for real-time classification of respiratory patterns.

7.7.1. Interpretation of Results

AUREA had substantial accuracy when compared to the "gold standard", as well as almost perfect consistency, which makes it a reliable and repeatable method. Additionally, it had negligible detection delays.

AUREA classified respiratory patterns with substantial accuracy, contrasted to AORED, whose accuracy was only borderline substantial, and about 50 % of the times it was only moderate. AUREA had better detection performance than AORED in all respiratory patterns. This was especially marked for PAU, where the probability of detection of AUREA was $P_D = 0.77$, compared to a $P_D = 0.55$ of AORED. Moreover, AUREA was able to achieve high P_D values, while keeping low probabilities of false alarm (P_{FA}). In fact, when we combined P_D and P_{FA} to yield the *d*-value, a measure of overall detection performance [139], AUREA was significantly better than AORED in all patterns.

AUREA was also better than AORED at detecting the start and end times of pattern segments. AUREA estimated these times with negligible bias (< 0.1 s), and high precision (IQR ≈ 1.1 s). In contrast, AORED had a small bias (≈ 0.2 s) and lower precision (IQR ≈ 2.5 s). Thus, AUREA was better at estimating the length of a segment, a property that is very important for the study of apnea, since apneic events are defined by PAU length.

7.7.2. Training Considerations

AUREA assumes that all respiratory patterns exist in the training data. The validity of this assumption increases with the amount of training data, so it is recommended to train with the most possible data. A large training set from several subjects can also help to minimize over-

fitting, since its use would result in a more generalized classifier that would be more representative of the population.

AUREA was designed with the intention to be implemented in real-time. For this reason, training and classification were split as two independent stages; training occurred first, and classification of an independent record was performed once the classification parameters had been obtained. However, there may be situations where the analysis can be performed off-line, or there is few data available for training. In these cases, it may be better to use all available data to train AUREA, and then use the classification parameters to classify them. Such an off-line implementation would maximize the amount of training data, which increases the chances of having all patterns represented.

7.7.3. Comparison to Other Methods

AUREA performs a comprehensive, sample-by-sample classification of the respiratory patterns. AORED was used throughout the paper as a comparison to AUREA because it is the only available method in the literature that performs a similar analysis. However, other methods have been developed to analyze some aspect(s) of the respiratory patterns. This section compares several characteristics of these methods to those of AUREA.

7.7.3.1. Trend Removal

Respiratory signals often contain a low frequency trend [85]. The method in [91] de-trends the signal by using the second difference with respect to time. This process may be problematic since it tends to amplify high-frequency noise, especially when signal-to-noise ratio is low. In contrast, AUREA estimated the low frequency trend by passing the signal through a moving-average filter. This averaging process attenuated additive white noise, removed high frequency noise, and yielded a smooth estimate of the trend [138]. The de-trended signal was obtained by subtracting the trend from the raw signal.

7.7.3.2. Unsupervised Classification

Previous detectors of respiratory patterns have required the selection of thresholds to perform the classification. The approach has been to either select an arbitrary threshold [3, 5, 86, 87, 92], or

determine the threshold based on data manually analyzed by experts [85]. The first strategy is subjective, and may not yield optimal results due to poor threshold selection. The second strategy can find the threshold that optimizes the relation between P_D and P_{FA}, but still requires a sample of data manually analyzed by expert scorers.

AUREA automatically determines the classification parameters with no human intervention by using clustering, an unsupervised machine learning approach. This made AUREA a fully automated, completely objective method, which was fast to implement given that no manual analysis was required.

7.7.3.3. Comprehensive Classification of Respiratory Patterns

Most of the methods for the study of respiratory patterns have focused on detecting a single pattern of interest, e.g., PAU or apnea [86, 87, 89-91, 93], ASB [4, 71, 116], or MVT [3, 5]. However, it may be possible that simultaneous classification of multiple patterns would significantly improve the overall classification performance, as recognized by De Groote et al. [80].

AUREA discriminates between multiple respiratory patterns simultaneously by combining several metrics of respiratory behavior. By doing this, AUREA yielded a comprehensive classification describing the full sequence of respiratory patterns at each time, with high accuracy and consistency, and almost no delay.

This analysis produced by AUREA describes the occurrence of patterns in time, as well as the sequencing of these patterns. These properties could be used to determine the relationship between different postoperative respiratory patterns and postoperative apnea (POA), and study other respiratory conditions like extubation readiness [120], bronchopulmonary dysplasia, and others.

7.7.3.4. Sample-by-sample Analysis

The manual scoring guidelines provided by the American Academy of Sleep Medicine (AASM) require scorers to classify only segments with apneic events [8]. Thus, much of the data is not classified, and so cannot be used for further analysis. In contrast, AUREA assigns a respiratory

pattern to each sample. This comprehensive, continuous classification enables the application of signals and systems, and time-series analysis to the data, broadening the spectrum of tools that may be used to study the respiratory patterns.

AUREA analyzes data on a sample-by-sample basis, so we thought it was important analyze its performance on the same sample-by-sample scale. This type of evaluation is the most accurate because it takes into account every sample, but it is also the strictest, because every misclassified sample reflects negatively on the performance. AUREA performed very well under this scheme. In contrast, most other studies involving automated detectors of respiratory patterns have opted for more lenient evaluations.

The most popular approach is an epoch-by-epoch scheme, where data are split into epochs of 30 s, or epochs of 15 s [80], and up to 1 min [94]. However, with this approach if an epoch contains the pattern of interest (e.g., a PAU), the full epoch is deemed to be correctly classified even if only a short portion of the pattern of interest was actually identified. Thus it will overestimate the detection probabilities.

Another scheme is the segment-by-segment, or event-by-event evaluation, in which an automatically detected segment is deemed to be detected if at least part of it overlaps a "gold standard" scored segment [92]. This scheme will also overestimate performance, since segments are counted as correct, even if only a fraction of the data is classified correctly.

7.7.4. Possible Limitations and Future Work

7.7.4.1. Management of Sample Unbalance

We used K-means clustering to classify respiratory patterns in an unsupervised fashion. However, K-means required an adjustment to the decision boundaries due to sampling unbalance. Future work should attempt to improve classification performance by implementing a different sample re-balancing strategy [164], where training is performed on data pre-processed such that all respiratory patterns are evenly represented. Sample re-balancing could be done by reducing the samples from the pattern that is oversampled, and obtaining additional pseudo-samples with a procedure like the Synthetic Minority Over-sampling Technique (SMOTE) [165].

7.7.4.2. Training with Limited Data

When there is few training data it may be possible that one or more respiratory patterns are not represented. This is especially important for PAU and ASB, since each is present less than 5 % of the time in infants at risk of POA [10]. It is possible that the K-means classifiers would not find clusters for such patterns, but this hypothesis was not evaluated. Future work should evaluate this scenario in an attempt to determine the minimum amount of data required for appropriate training. This could be achieved by training AUREA with subsets of the clinical data where the number of samples with "gold standard" classification of PAU and/or ASB is gradually reduced until they are completely removed, and assessing the accuracy of AUREA under these circumstances.

7.7.4.3. Sigh Classification

The work in [10] defined 6 unique, mutually exclusive respiratory patterns. AUREA was designed to classify 5 of them: PAU, SYB, ASB, MVT, and UNK. The 6th pattern was sigh (SIH), which only represents a small portion of the respiratory patterns (e.g., SIH comprised only < 2 % of the clinical data set used in this work). AUREA assigned samples with "gold standard" classification of SIH to the other 5 patterns with these probabilities: PAU 0.01, SYB 0.11, ASB 0.01, MVT 0.81, and UNK 0.05. Since AUREA assigned most "gold standard" SIH samples to MVT, SIH classification could be accomplished by having a post-processing step to distinguish between SIH and MVT in samples classified as MVT. Future work should attempt this to enable SIH classification in AUREA, since literature reports that in infants SIH and apnea may be linked [25, 153].

7.7.5. Significance

7.7.5.1. Evidence-based Definition of Apnea

AUREA detected PAU segments very well, and estimated their start and end times with no bias and high precision. Thus, AUREA can accurately determine the length of PAU segments. This is a relevant property that is required by studies of apnea, since a key component in the definition of apnea is PAU length.

Apnea definitions often qualify events as clinically relevant if they are accompanied by decreased heart rate (bradycardia) or drops in blood oxygen saturation (desaturation). However, these definitions are arbitrary and are based on physician experience and subjective judgments. AUREA, in conjunction with measures of heart rate and blood oxygen saturation, could provide the means to establish better, evidence-based, objective definitions of apnea, by studying which sequences of respiratory patterns are associated with bradycardia and/or desaturation.

7.7.5.2. Improved Analysis of Respiratory Patterns

AUREA provides an objective, fast, reliable, repeatable, and low-cost means of analysis of respiratory patterns. It had substantial accuracy when compared to the "gold standard", and did not share limitations associated with manual scoring.

AUREA makes it possible to carry out large studies of respiratory patterns, involving multiple institutions, given that it can analyze data quickly and with no added costs. Moreover, AUREA has almost perfect consistency, so results can be easily compared among institutions. This consistency also enables the development of longitudinal studies, where the respiratory patterns of patients need to be assessed in a repeatable way, to be compared at multiple times in life. These types of studies have not been possible with conventional manual (CMS) scoring due to its high variability and cost.

AUREA is amenable for real-time implementation. Thus, it could be implemented at the bedside at hospitals or even at home. This would allow for a comprehensive, real-time monitoring of respiratory patterns of patients, which would provide evidence for improved clinical decision making. This evidence could be used to improve patient management, as well as to better distribute hospital resources.

7.7.5.3. Study of Postoperative Respiratory Patterns

There is evidence that postoperative apnea (POA) events are associated with abnormal respiratory patterns [14, 20, 24]. This suggests that an analysis of the underlying postoperative respiratory patterns could be used to estimate the risk of POA, and the time at which this risk abates. AUREA enables this by providing an objective, and reliable classification of the

respiratory patterns, which could be used to identify relationships between patterns, or sequences of patterns, and the risk of POA.

In fact, Cote et al. [17] recognized that standardized, continuous monitoring was necessary to better study POA. This can be performed with AUREA, which also has the potential to be implemented in real-time at the bedside.

7.7.5.4. Other Studies of Respiration

AUREA was developed for the study of the respiratory patterns of infants at risk of POA. However, it has multiple applications in other fields involving respiration such as: (i) apnea of prematurity, (ii) prediction of extubation readiness [120], (iii) sudden infant dead syndrome, (iv) asthma, (v) bronchopulmonary dysplasia, and (vi) sleep apnea, among others.

7.8. Conclusion

We presented AUREA, a completely automated method for unsupervised classification of respiratory patterns using RIP signals, and successfully applied it to data from infants recovering from surgery and anesthesia. AUREA eliminates the shortcomings of human intervention, while comparing favorably to the "gold standard", and performing substantially better than a previous method based on supervised classification [85].

8. Discussion and Future Work

8.1. Summary

Evidence suggests that a comprehensive analysis of the postoperative respiratory patterns may provide insight about the probability of an infant suffering postoperative apnea (POA) [14, 19, 20, 24, 39]. This could help to identify those infants at risk, and also determine how long after surgery this risk persists.

However, two main limitations have hindered the study of the postoperative respiratory patterns. First, there is no available data representative of infants at risk of POA. Second, existing analysis methods are not adequate. The most accepted method is conventional manual scoring (CMS), which is performed by expert scorers using the guidelines from the American Academy of Sleep Medicine (AASM) [8]. This method has several limitations: it has low intra- and inter-scorer repeatability [9], is labor intensive, time-consuming, and expensive. Automated methods have been developed that address some of these limitations, but as yet there is no comprehensive, reliable alternative to CMS.

The objective of this thesis was to address these limitations by: (i) acquiring a representative dataset of postoperative respiratory patterns from infants at risk of POA, and making these data available to the public; and (ii) improving tools for the analysis of infant respiratory patterns by developing a comprehensive, reliable (i.e., high accuracy), repeatable (i.e., high consistency), fast, and low-cost methodology to classify the respiratory patterns as a function of time.

The work described in this thesis included the acquisition of cardiorespiratory data from infants at risk of POA. Infants recruited to the study had a postmenstrual age (PMA) of 60 weeks or less, since this is the most important clinical risk factor [14, 17, 37, 39, 41, 45, 46]. Data acquisition started immediately after surgery, and continued for up to 12 h [14], since the first POA may occur within the first hours and up to 12 h postoperatively [14-16]. Measurements were made using Respiratory Inductive Plethysmography (RIP) for respiratory movements and an oximeter for blood oxygen saturation (SAT) and photoplethysmography (PPG). These sensors were selected because they are noninvasive, do not cover the face (and so do not interfere with the

infant's breathing, feeding or care), and RIP measurements can detect periods of airway obstruction [8, 55, 56]. This last point was especially important because a significant proportion of POA events has an obstructive component, which leads to larger decreases in SAT than do central POA [19].

With respect to the analysis of these data, it was necessary to develop a method to analyze large datasets with high accuracy and consistency. Given the length of records in the data set (~12 h per recording), it was necessary to automate the analysis to make it low-cost and fast. However, to establish the validity of the automated method, it was essential to evaluate it against an accurate "gold standard" reference. Currently the most accepted analysis of respiratory patterns is CMS based on AASM guidelines [8]. However, CMS cannot produce a reliable analysis due to its low intra- and inter-scorer repeatability [9]. Moreover, AASM guidelines only define "clinically relevant" events (i.e., central, obstructive, and mixed apnea), but do not consider the other respiratory patterns that occur postoperatively (e.g., short pauses, sighs, thoraco-abdominal asynchrony). As a result large sections of data records are not scored, and therefore no conclusions can be drawn from the respiratory patterns in those sections. Moreover, in infants there is no consensus as to the length threshold that separates an apnea from a short pause. For instance, the American Academy of Pediatrics (AAP) defines the threshold as 20 s [166], but other studies frequently use lengths of 15 s [14, 22, 34, 52, 167, 168] or others [169]; moreover these thresholds were chosen based on subjective clinical observations and judgment, rather than objective experimental evidence.

For these reasons, we felt it necessary to develop a set of tools to support the comprehensive, manual analysis of infant respiratory patterns, to use as a "gold standard" reference. The resulting tool set included: definitions for all patterns encountered in infant respiratory data, i.e., synchronous-breathing (SYB), asynchronous-breathing (ASB), sigh (SIH), respiratory pause (PAU), movement artifact (MVT), and unknown (UNK); rules to apply these definitions; and software to facilitate application of the rules to infant data. Use of these tools yields an improved manual analysis that assigns a respiratory pattern type to every sample in the recording, resulting in a continuous signal describing the instantaneous respiratory pattern.

To further reduce intra- and inter-scorer variability of the manual analysis, we developed a method based on Expectation-Maximization (EM) to combine the results from multiple, manual scorers. We showed that this method yields a more accurate and consistent analysis than those of individual, manual scorers, resulting in a high quality "gold standard" reference.

An important aspect not available in the literature was an automated method that could comprehensively analyze the respiratory patterns, and deliver a continuous signal representing the instantaneous respiratory pattern similar to the "gold standard" analysis. Previous methods were designed to detect isolated segments of either PAU or MVT, or to estimate the degree of ASB. Thus, we combined PAU, MVT, and ASB detectors to yield AORED, an Automated Off-Line Respiratory Event Detector for comprehensive classification of the instantaneous respiratory pattern.

In the literature there were two main types of automated PAU detectors, those based on metrics from the amplitude of respiratory signals compared to thresholds [86, 87, 91], and those that estimate several features and input these to specialized classifiers [89, 90]. Our review in Chapter 2 showed that PAU detectors based on amplitude metrics and thresholds performed as well as methods using more complex classifiers. Thus, AORED estimated metrics of respiratory signal amplitude and compared them to thresholds to detect PAU. With respect to ASB, we decided to incorporate the XOR method developed by Motto et al. [4] into AORED since it performs better than any other available method, especially during high noise conditions. Finally, in the case of MVT detection, our review revealed that for infants, the best available detector was that developed by Aoude et al. [5], and so we decided to incorporate this detector as well.

An important limitation of many previous, threshold-based, PAU detectors is that they suffer from high probabilities of false alarm (P_{FA}) [86, 87]. We believe that this likely resulted from inappropriate threshold selection, since these studies used only arbitrary thresholds. In fact, one apnea detector that optimized the threshold based on the manual analysis from expert scorers [91] had a much lower P_{FA} than those with arbitrary thresholds. However, neither the TAA estimator from [4] nor the MVT detector from [5] addressed the problem of how to estimate the

optimum threshold. For AORED, we developed a systematic, evidence-based method for threshold selection using ROC analysis based on a reference set of manually analyzed data.

Later, we explored the use of K-means clustering [162], an unsupervised machine learning technique, to determine the thresholds automatically and classify the instantaneous respiratory pattern without the need of any manual analysis. As a result, we developed AUREA, an Automated Unsupervised Respiratory Event Analysis system that requires no human intervention to reproduce the "gold standard" analysis with high accuracy and consistency.

AUREA estimates metrics of respiratory behavior from RIP signals, and uses them to classify samples into one of 5 types of respiratory patterns: SYB, ASB, PAU, MVT, and UNK. The analysis produced by AUREA agrees substantially with the "gold standard" analysis, but is much faster and does not share the limitations associated with manual scoring. Moreover, AUREA is significantly more accurate than AORED, which makes it the method of choice for analysis of infant respiratory patterns.

8.2. Original Contributions

8.2.1. Library of Infant Data

The first contribution from this thesis is the acquisition of a library of cardiorespiratory data from infants at risk of POA, and its deposition in a public archive [11]. These data represent a valuable collection of cardiorespiratory signals because they: (i) are representative of infants at risk of POA; and (ii) were acquired continuously, starting immediately after surgery and lasting for up to 12 h, and so can be used to study the postoperative respiratory patterns and their relation to POA. We made these data fully available to the public, without restriction [11].

This is a unique set of data since there is no other equivalent data available. The most similar dataset is from the Collaborative Home Infant Monitoring Evaluation (CHIME) study [118], which is a collection of overnight, cardiorespiratory signals from more than 1,000 infants. However, the CHIME dataset is not from infants at risk of POA, and it only comprises a few seconds before and after automatically-detected periods of slow heart rate or apnea, so it is not

possible to study respiratory patterns as a continuous sequence of events. This library will enable the development of new methods to analyze infant cardiorespiratory data, and will also help advance the clinical understanding of POA.

8.2.2. Comprehensive Classification of Infant Respiratory Patterns

The second contribution of this work is a methodology for the comprehensive classification of infant respiratory patterns. We identified 6 unique, mutually exclusive patterns that may be observed in Respiratory Inductive Plethysmography (RIP) signals from infants recovering from surgery and anesthesia. Furthermore, we developed explicit, concise definitions for these patterns, established manual scoring rules to assign the patterns to signal segments, and made these definitions and scoring rules publicly available [10].

Previous CMS analyses only produced a list of "clinically relevant" events (e.g., apnea) and the time of their occurrence, but failed to account for patterns not considered "clinically relevant" by the AASM guidelines [8]. However, these discarded patterns may provide information about future POA events, so it is important to include them in the analysis. Our definitions and scoring rules addressed this by including these patterns into the analysis.

8.2.3. Manual Scoring Tools

We developed a set of tools that support the efficient, manual scoring of cardiorespiratory signals according to these rules. These tools include: (i) RIPScore, a software to visualize cardiorespiratory signals and apply the rules to these data; (ii) a curated library of segments representative of the 6 respiratory patterns; (iii) a fully automated training protocol, which is incorporated into RIPScore; and (iv) a quality control method to monitor scorer performance over time.

The tools allow to comprehensively analyze infant respiratory patterns in a continuous, sample-by-sample fashion, while also allow to establish and maintain high intra- and inter-scorer repeatability. This provides significant analysis improvements compared to CMS, which is limited by low intra- and inter-scorer repeatability [9]. We made these tools fully available to the

public [10, 12] so software is readily available at no cost, unlike commercial software that requires licensing fees.

Previous studies of infant respiration using CMS required manual scorers to be certified by the AASM in an attempt to maintain the analysis quality. However, this approach did not resolve the low intra- and inter-scorer repeatability of CMS [9]. We addressed this with our automated training protocol and quality control method. These tools make sure that trainees transition to trained scorers once they have reached an excellent level of accuracy and consistency. Then, when scorers analyze data for a given study, the quality control tool monitors their performance session-by-session, making investigators aware of scorer performance. Using this tool, investigators can ensure quality by taking prompt, corrective actions if a given scorer shows evidence of diminished performance. For example, the scorer could be re-trained, or if low performance is a recurrent problem, a new scorer could be recruited.

Other benefits of the automated training are its efficiency and cost-effectiveness. This is because trainees need only to install RIPScore on their personal computers, and carry out the training protocol guided by the software. Trainees do not require continuous supervision and feedback from senior scorers, which is an advantage over CMS. Thus, automated training permits to better allocate the effort of senior scorers that would otherwise be spent in coaching trainees.

8.2.4. "Gold Standard" Analysis of Respiratory Patterns

This thesis also presents a method to combine the analyses from multiple, manual scorers using Expectation-Maximization (EM) to reduce the inter-scorer variability and yield a "gold standard" analysis. This post-processing method has excellent accuracy and consistency, and its performance is significantly higher than that of individual, manual scorers. The "gold standard" produced by this method represents a comprehensive, highly accurate and consistent option, which is superior to that produced by CMS.

Another approach to find a consensus between multiple, manual scorers is the majority vote (MV), where samples are assigned the pattern with the most votes [10, 143]. However, when there is much disagreement among scorers the majority may not be absolute, and so the final

pattern would be determined by a minority of votes. Also, votes from all scorers are weighted equally regardless of their performance. Our method based on EM was able to significantly increase the classification accuracy from that of MV by taking into account the individual performance of each scorer and weighting their votes accordingly. We showed that the EM method requires many fewer scorers to reach a given level of performance, as evidenced by the simulation study in Chapter 5.

8.2.5. Supervised Classification of Respiratory Patterns

We developed AORED, an Automated Off-Line Respiratory Event Detector that automated the analysis of infant respiratory patterns. AORED combines individual detectors of pause, movement artifact [5], and asynchronous-breathing [4], to automatically classify the respiratory patterns by comparing metrics of respiratory behavior to thresholds. AORED classifies the respiratory patterns on a sample-by-sample basis, is repeatable, standardized, robust in high noise conditions, and amenable for real-time implementation.

AORED classifies samples into one of several respiratory patterns in a manner similar to the "gold standard". This improves over previous automated methods aimed to detect a single pattern of interest [4, 5, 86, 87, 89-92], which failed to comprehensively describe the multiple respiratory patterns.

AORED provides a strategy to select optimum threshold values based on representative manual analysis results. This is an improvement with respect to previous methods that selected thresholds arbitrarily [5, 6, 86, 87]. This threshold selection is based on Receiver Operating Characteristics (ROC) analysis to determine the optimum threshold values. ROC curves are generated for each respiratory pattern using a representative sample of manually analyzed data, and the threshold is selected as the optimal point in the curve, i.e., the point that is furthest from the diagonal "chance" line.

8.2.6. Fully Automated Analysis of Respiratory Patterns

The final contribution of this work is AUREA, an Automated Unsupervised Respiratory Event Analysis system that comprehensively analyzes infant respiratory patterns in an accurate,

consistent, fast and low-cost fashion. AUREA makes use of K-means clustering [162], an unsupervised learning technique, to automatically classify respiratory patterns. Because of this, AUREA requires no human intervention to work and makes the analysis fully objective, which is a significant improvement from AORED that requires manually scored data to determine classification thresholds.

The full automation of AUREA reduces the analysis time and costs, and at the same time delivers an analysis with near-perfect consistency and substantial accuracy. In other words, AUREA addresses the subjective, variable, costly, time-consuming nature of CMS that has limited the study of infant respiratory patterns and POA.

Additionally, AUREA makes it possible to accurately analyze large datasets from multiple institutions with almost perfect repeatability, enabling the development of large, multi-institutional studies of infant respiration. AUREA also opens the possibility for real-time monitoring of respiratory patterns, which could provide important information with potential to improve patient care.

8.3. Implications of the Results

8.3.1. Advance the Study of Postoperative Apnea

The contributions described in this thesis address two aspects that have limited the study of POA and its relation to the postoperative respiratory patterns. First, we acquired a representative dataset from infants at risk of POA, and made it available publicly. Second, we developed a set of tools to analyze the respiratory patterns in a fully automated, comprehensive, high quality manner. These contributions will help advance the study of POA in a variety of important ways:

(i) The dataset is a valuable collection of cardiorespiratory signals from the immediate postoperative period, which can be used to explore possible hypothesis related to POA and the respiratory patterns. These will allow investigators to focus their efforts in exploring new approaches to the analysis of respiratory patterns, without designing, obtaining approval, and carrying out a data acquisition protocol.

(ii) AUREA can be used to characterize the respiratory patterns in a dataset, and the resulting information used to study the relation between these patterns and POA. This could result in a predictor of POA risk able to determine when such risk no longer exists, so that infants could be promptly released from the recovery room. The ability to identify infants at risk for POA would help to appropriately distribute clinical resources: subjects that require more attention would be able to receive it, while those at low risk would reduce the demand of resources. This would yield an improvement in health care service.

(iii) AUREA produces a comprehensive analysis of the respiratory patterns in a fully automated manner; it has high accuracy and consistency, so it can be used to analyze large datasets. This opens the door to multi-institutional and/or longitudinal studies, which were not previously feasible because CMS, the only available analysis method, is very labor intensive, expensive, and has high variability.

(iv) AUREA has the potential for real-time implementation. Thus, it could be used to monitor infants at the bedside to provide more detailed, instantaneous information about the respiratory patterns compared to conventional clinical monitors.

8.3.2. Definition of Postoperative Apnea

An important limitation in previous studies of POA is that investigators have used a variety of definitions for apnea, which has made it difficult to compare the results, and generalize the conclusions. The literature has defined two main types of apneas based on the perceived clinical relevance: (i) prolonged, life-threatening apnea, and (ii) short, brief apnea. Tables 8.1 and 8.2 show a comprehensive list of these definitions.

These definitions of POA have 3 problems. First, the term “apnea” is used indiscriminately to refer to either brief or prolonged events, even though their consequences may be very different clinically. This leads to confusion and misinterpretation of the results. Second, the duration threshold that defines an apnea varies widely; the most common threshold is 15 s, but there are definitions with thresholds of 6 s, 10 s, and 20 s. This makes it impossible to compare the results and conclusions from different POA studies. Third, life-threatening POA has two definitions: in one case it is defined as a respiratory pause accompanied by a decrease in heart rate or blood oxygen saturation, while the second definition is based only on the duration of the respiratory

Manifestation	Duration (s)	As defined in
Respiratory pause	≥ 20	[53]
	> 20	[37, 169]
	≥ 15	[17, 21, 22, 40, 41]
	> 15	[14, 39]
	> 10	[45]
Respiratory pause that may be accompanied by slow heart rate	≥ 15	[20]
Cessation of breathing associated with slow heart rate, cyanosis, or pallor	NS	[37, 169]
Respiratory pause accompanied by slow heart rate	NS	[14, 17, 21, 22, 40, 41]
Respiratory pause accompanied by slow heart rate, or blood oxygen desaturation	> 10	[16]
Cessation of airflow	≥ 15	[50]
	> 6	[19]
Cessation of respiratory movement	>15	[52]

Table 8.1. Definitions of prolonged, life-threatening postoperative apnea. NS = Not specified.

Manifestation	Duration (s)	As defined in
Respiratory pause	< 15	[20]
	> 6 but < 15	[14]
Respiratory pause not associated with slow heart rate	< 15	[21, 22, 40, 41]
Cessation of respiratory movement	11 to 15	[52]
Cessation of airflow	> 5 but < 15	[50]

Table 8.2. Definitions of short, brief postoperative apnea.

pause. These two definitions are treated as equivalent, although they may correspond to two different types of events. Thus, the first definition concerns a clinically significant apnea that results in hypoxia and/or heart rate instability. In contrast, the second definition could be a respiratory pause that does not pose an immediate clinical concern, but should be closely monitored.

It is therefore necessary to standardize the definition of apnea to advance in the study of POA. The dataset and methods presented in this thesis could assist in the development of these definitions. AUREA can analyze the respiratory signals to obtain the respiratory patterns; the SAT signal can be used to detect periods of desaturation, and the photoplethysmography signal to estimate the heart rate. These data could then be mined to identify different groups of respiratory pauses (e.g., life-threatening apnea, long pause without desaturation or slow heart rate, and short pause), and these groups could define different types of apnea. Indeed, a pilot study of this idea found that the threshold between short and long respiratory pauses is 14.6 s [139], which is very close to the most common value used to define POA in previous studies.

8.3.3. Other Studies of Infant Respiration

The tools for analysis of infant respiratory patterns from this thesis were developed to study POA. However, they have multiple applications in other studies related to infant respiration. For example, we have used them to study the respiratory patterns of intubated preterm infants with the intention to predict extubation readiness [120, 170]. AUREA is also amenable for use in studies of sleep and breathing, where CMS performed by sleep laboratory technicians is still considered the preferred method of analysis in spite of its multiple limitations.

8.4. Future Work

8.4.1. Robustness of AUREA in High Noise

We have previously shown that the MVT metric from AUREA is robust in the presence of high wideband, and low-frequency noise [138]. However, a similar study has not been performed for the remaining AUREA metrics. These metrics are estimated from RIP signals which may exhibit low frequency trends [5], so it is especially important to verify the robustness of AUREA in low

frequency noise. Moreover, recordings might have low signal-to-noise ratio (SNR) due to incorrect sensor placement, electromagnetic interference, or faulty wires. Thus, future studies should assess the robustness of these metrics using a procedure similar to our previous work [85, 138].

This robustness assessment is especially important if AUREA is to be used to analyze data from studies with long, unattended data acquisition sessions. During unattended data acquisition the sensors may be displaced due to patient handling, and this might yield noisier recordings. In our data, we observed that unattended recordings (i.e., data used in Chapter 6) had lower SNR than continuously attended recordings (i.e., the data library described in Chapter 3).

8.4.2. Outlier Detection

AUREA requires detection of outliers in the pre- and post-processing stages. Outliers were detected as those points situated beyond a pre-defined quantile threshold. This is efficient and simple to implement, but may exclude samples that should be included, or vice versa. Outlier detection is particularly important for the training step in AUREA. This is because AUREA uses K-means clustering [162] to determine the classification parameters, and K-means is very sensitive to outliers. Thus, having outliers in the training data would result in inaccurate classification boundaries biased towards the outliers, and this would yield an invalid classification of the respiratory patterns. Thus, it is important that all outliers are excluded from the training data. Future work should implement an automated detector to find optimum outlier thresholds.

8.4.3. Real-time Implementation

AUREA is amenable for real-time, or near real-time use, because only training needs to be performed off-line. After training, data can be classified as soon as the input metrics are available. Most of the input metrics used by AUREA were implemented using two-sided, finite impulse response (FIR) filters, so they can be estimated in near real-time with a delay of only half the length of the filter windows. Only the SYB and ASB metrics use zero-phase, forward-backward, band-pass infinite impulse response filter [137]. Future work should redesign this filter as a FIR filter to enable real-time implementation of AUREA.

Real-time usage would allow clinicians to use AUREA for comprehensive, bedside monitoring of respiration, which could provide important, prompt information about the state of the patient. This could help inform clinical decision making in a better capacity than current clinical monitors based on impedance pneumography, which have a slow response, miss apneic events, and have many false alarms [30].

AUREA employs a training dataset to determine the classification parameters based on the respiratory patterns from the population, and then these parameters are used to classify separate test data from new individuals. Classification could be improved if the parameters were adjusted to match the individual variations of each patient's respiratory patterns. Future experiments could evaluate this possibility by making the classification parameters patient-specific, adapting them based on newly observed test data.

8.4.4. Application to Adult Data

The band-pass filters in AUREA's metrics were designed to pass frequencies in the infant respiratory band (i.e., 0.4 Hz and 2.0 Hz [85]). To implement AUREA in adult data, it is necessary to modify the cut-off frequencies to span the adult respiratory band. Future studies could implement this modification, and test AUREA in adult data from sleep studies.

Many adults suffer from Obstructive Sleep Apnea Syndrome (OSAS) [123], and the diagnosis requires CMS analysis of overnight recordings [8]. AUREA could significantly improve reliability of the analysis, while also dramatically reducing the time required to analyze the data. This would make the diagnosis of OSAS more efficient, which would contribute to alleviate the long waiting times that exist in sleep laboratories [123]. Moreover, systematic implementation of AUREA would help to establish a very high consistency among different institutions.

8.5. Conclusion

The study of POA and its relation with the postoperative respiratory patterns has been hindered by the lack of two main factors: (i) appropriate data from infants at risk, and (ii) a method to analyze the respiratory patterns in a comprehensive, reliable (i.e., high accuracy), repeatable (i.e., high consistency), fast, and low-cost fashion.

In this thesis we addressed these limitations by: (i) acquiring cardiorespiratory data from infants recovering from surgery and anesthesia, and making these data publicly available; (ii) establishing definitions and scoring rules to comprehensively analyze infant respiratory patterns; (iii) developing a set of tools to enable the application of these rules to infant data, and making these tools publicly available; (iv) providing a method to consolidate several manual scoring results into a highly accurate and consistent "gold standard" manual analysis; (v) developing AORED, an automated, supervised detector able to replicate the analysis by learning from the "gold standard" reference; and (vi) developing AUREA, a fully automated, unsupervised classifier that requires no human intervention to replicate the "gold standard" analysis, and that performs better than AORED. With AUREA it is possible to produce a comprehensive, reliable, repeatable, fast, and low-cost analysis that can be used to study respiratory patterns from infants at risk of POA.

9. References

[1] S. Semienchuk, "A Portable Monitor for Automated, On-line Cardiorespiratory State Classification," M. Eng., Biomedical Engineering, McGill University, Montreal, QC, Canada, 2006.

[2] S. M. Semienchuk, A. L. Motto, H. L. Galiana, K. A. Brown, and R. E. Kearney, "A Portable, PC-Based Monitor for Automated, On-line Cardiorespiratory State Classification," in *Conf Proc 27th IEEE Eng Med Biol Soc*, Shanghai, China, 2005, pp. 4420-4423.

[3] A. L. Motto, H. L. Galiana, K. A. Brown, and R. E. Kearney, "Detection of movement artifacts in respiratory inductance plethysmography: performance analysis of a Neyman-Pearson energy-based detector," *Conf Proc 26th IEEE Eng Med Biol Soc,* vol. 1, pp. 49-52, 1-5 Sept. 2004 2004.

[4] A. L. Motto, H. L. Galiana, K. A. Brown, and R. E. Kearney, "Automated estimation of the phase between thoracic and abdominal movement signals," *IEEE Trans Biomed Eng,* vol. 52, pp. 614-621, Apr 2005.

[5] A. A. Aoude, A. L. Motto, H. L. Galiana, K. A. Brown, and R. E. Kearney, "Power-Based Segmentation of Respiratory Signals Using Forward-Backward Bank Filtering," in *Conf Proc 28th IEEE Eng Med Biol Soc*, 2006, pp. 4631-4634.

[6] A. Aoude, "Automated off-line cardiorespiratory event detection and validation," M. Eng., Biomedical Engineering, McGill University, Montreal, QC, Canada, 2006.

[7] K. A. Brown, A. A. Aoude, H. L. Galiana, and R. E. Kearney, "Automated respiratory inductive plethysmography to evaluate breathing in infants at risk for postoperative apnea," *Can J Anaesth,* vol. 55, pp. 739-747, Nov 2008.

[8] R. B. Berry, R. Budhiraja, D. J. Gottlieb, D. Gozal, C. Iber, V. K. Kapur, C. L. Marcus, R. Mehra, S. Parthasarathy, S. F. Quan, S. Redline, K. P. Strohl, S. L. Davidson Ward, M. M. Tangredi, and M. American Academy of Sleep, "Rules for scoring respiratory

events in sleep: update of the 2007 AASM Manual for the Scoring of Sleep and Associated Events. Deliberations of the Sleep Apnea Definitions Task Force of the American Academy of Sleep Medicine," *J Clin Sleep Med,* vol. 8, pp. 597-619, Oct 15 2012.

[9] N. A. Collop, "Scoring variability between polysomnography technologists in different sleep laboratories," *Sleep Med,* vol. 3, pp. 43-47, Jan 2002.

[10] C. A. Robles-Rubio, G. Bertolizio, K. A. Brown, and R. E. Kearney, "Scoring Tools for the Analysis of Infant Respiratory Inductive Plethysmography Signals," *PLoS ONE,* vol. 10, p. e0134182, 2015.

[11] C. A. Robles-Rubio, G. Bertolizio, K. A. Brown, and R. E. Kearney, "Data from: Scoring Tools for the Analysis of Infant Respiratory Inductive Plethysmography Signals.," ed: Dryad Data Repository. doi: 10.5061/dryad.72dk5, 2015.

[12] C. A. Robles-Rubio, K. A. Brown, and R. E. Kearney, "McGill CardioRespiratory Infant Behavior Software (McCRIBS): Initial Release," ed: Zenodo, DOI: 10.5281/zenodo.31441, 2015.

[13] I. Hrynaszkiewicz, M. L. Norton, A. J. Vickers, and D. G. Altman, "Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers," *BMJ,* vol. 340, 2010.

[14] C. D. Kurth, A. R. Spitzer, A. M. Broennle, and J. J. Downes, "Postoperative Apnea in Preterm Infants," *Anesthesiology,* vol. 66, pp. 483-488, Apr 1987.

[15] D. J. Steward, "Preterm Infants are More Prone to Complications Following Minor Surgery than are Term Infants," *Anesthesiology,* vol. 56, pp. 304-306, 1982.

[16] A. J. Davidson, N. S. Morton, S. J. Arnup, J. C. De Graaff, N. Disma, D. E. Withington, G. Frawley, R. W. Hunt, P. Hardy, M. Khotcholava, B. S. Von Ungern Sternberg, N. Wilton, P. Tuo, I. Salvo, G. Ormond, R. Stargatt, B. G. Locatelli, and M. E. Mccann, "Apnea after Awake Regional and General Anesthesia in Infants. The General Anesthesia

Compared to Spinal Anesthesia Study—Comparing Apnea and Neurodevelopmental Outcomes, A Randomized Controlled Trial," *Anesthesiology,* May 14 2015.

[17] C. J. Cote, A. Zaslavsky, J. J. Downes, C. D. Kurth, L. G. Welborn, L. O. Warner, and S. V. Malviya, "Postoperative Apnea in Former Preterm Infants after Inguinal Herniorrhaphy: A Combined Analysis," *Anesthesiology,* vol. 82, pp. 809-822, Apr 1995.

[18] L. J. Jones, P. D. Craven, A. Lakkundi, J. P. Foster, and N. Badawi, "Regional (spinal, epidural, caudal) versus general anaesthesia in preterm infants undergoing inguinal herniorrhaphy in early infancy," *Cochrane Database Syst Rev,* vol. 6, p. CD003669, 2015.

[19] C. D. Kurth and S. E. Lebard, "Association of Postoperative Apnea, Airway Obstruction, and Hypoxemia in Former Premature Infants," *Anesthesiology,* vol. 75, pp. 22-26, Jul 1991.

[20] L. G. Welborn, N. Ramirez, T. Hee Oh, U. E. Ruttimann, R. Fink, P. Guzzetta, and B. S. Epstein, "Postanesthetic Apnea and Periodic Breathing in Infants," *Anesthesiology,* vol. 65, pp. 658-661, Dec 1986.

[21] L. G. Welborn, R. S. Hannallah, R. Fink, U. E. Ruttimann, and J. M. Hicks, "High-dose Caffeine Suppresses Postoperative Apnea in Former Preterm Infants," *Anesthesiology,* vol. 71, pp. 347-349, Sep 1989.

[22] L. G. Welborn, L. J. Rice, R. S. Hannallah, L. M. Broadman, U. E. Ruttimann, and R. Fink, "Postoperative Apnea in Former Preterm Infants: Prospective Comparison of Spinal and General Anesthesia," *Anesthesiology,* vol. 72, pp. 838-842, May 1990.

[23] C. D. Kurth, "Postoperative arterial oxygen saturation: what to expect," *Anesth Analg,* vol. 80, pp. 1-3, Jan 1995.

[24] C. J. Coté and D. H. Kelly, "Postoperative Apnea in a Full-Term Infant with a Demonstrable Respiratory Pattern Abnormality," *Anesthesiology,* vol. 72, pp. 559-560, Mar 1990.

[25] B. Hoch, M. Bernhard, and A. Hinsch, "Different Patterns of Sighs in Neonates and Young Infants," *Biol Neonate,* vol. 74, pp. 16-21, 1998.

[26] T. Hoppenbrouwers, J. E. Hodgman, K. Arakawa, D. J. Mcginty, J. Mason, R. M. Harper, and M. B. Sterman, "Sleep Apnea as Part of a Sequence of Events: A Comparison of Three Months Old Infants at Low and Increased Risk for Sudden Infant Death Syndrome (SIDS)," *Neuropädiatrie,* vol. 9, pp. 320,337, 18.11.2008 1978.

[27] C. Iber, S. Ancoli-Israel, A. J. Chesson, S. Quan, and A. Academy of Sleep Medicine, *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*, 1 ed. Welchester, IL: American Academy of Sleep Medicine, 2007.

[28] K. Brown, C. Aun, E. Jackson, A. Mackersie, D. Hatch, and J. Stocks, "Validation of respiratory inductive plethysmography using the Qualitative Diagnostic Calibration method in anaesthetized infants," *Eur Respir J,* vol. 12, pp. 935-943, Oct 1998.

[29] R. T. Brouillette, D. Hanson, R. J. David, L. Klemka, A. Szatkowski, S. Fernbach, and C. E. Hunt, "A Diagnostic Approach To Children With Suspected Obstructive Sleep Apnea (Osa)," *Pediatric Research,* vol. 18, p. 228A, 1984.

[30] R. T. Brouillette, A. S. Morrow, D. E. Weese-Mayer, and C. E. Hunt, "Comparison of respiratory inductive plethysmography and thoracic impedance for apnea monitoring," *J Pediatr,* vol. 111, pp. 377-383, Sep 1987.

[31] R. T. Brouillette, S. K. Fernbach, and C. E. Hunt, "Obstructive sleep apnea in infants and children," *J Pediatr,* vol. 100, pp. 31-40, Jan 1982.

[32] S. M. Sale, J. A. Read, P. A. Stoddart, and A. R. Wolf, "Prospective comparison of sevoflurane and desflurane in formerly premature infants undergoing inguinal herniotomy," *Br J Anaesth,* vol. 96, pp. 774-778, June 1, 2006 2006.

[33] F. Al-Sufayan, M. Bamehrez, K. Kwiatkowski, and R. E. Alvaro, "The effects of airway closure in central apneas and obstructed respiratory efforts in mixed apneas in preterm infants," *Pediatr Pulmonol,* vol. 44, pp. 253-259, Mar 2009.

[34] S. M. Sale, "Neonatal apnoea," *Best Pract Res Clin Anaesthesiol,* vol. 24, pp. 323-336, Sep 2010.

[35] M. American Academy of Sleep, *The International Classification of Sleep Disorders, Revised: Diagnostic and Coding Manual*. Chicago, Illinois: American Academy of Sleep Medicine, 2001.

[36] G. A. Gregory and D. J. Steward, "Life-threatening Perioperative Apnea in the Ex-“premie”," *Anesthesiology,* vol. 59, pp. 495-498, Dec 1983.

[37] L. M. P. Liu, C. J. Coté, N. G. Goudsouzian, J. F. Ryan, S. Firestone, D. F. Dedrick, P. L. Liu, and I. D. Todres, "Life-threatening Apnea in Infants Recovering from Anesthesia," *Anesthesiology,* vol. 59, pp. 506-510, Dec 1983.

[38] T. K. Mcintosh, H. L. Bush, M. Palter, J. R. Hay, F. Aun, N. S. Yeston, and R. H. Egdahl, "Prolonged disruption of plasma β-endorphin dynamics following surgery," *J Surg Res,* vol. 38, pp. 210-215, Mar 1985.

[39] C. Bell, R. Dubose, J. Seashore, R. Touloukian, C. Rosen, T. H. Oh, C. W. Hughes, S. Mooney, and T. Z. O'connor, "Infant apnea detection after herniorrhaphy," *J Clin Anesth,* vol. 7, pp. 219-223, May 1995.

[40] L. G. Welborn, H. D. Soto, R. S. Hannallah, R. Fink, U. E. Ruttimann, and R. Boeckx, "The Use of Caffeine in the Control of Post-anesthetic Apnea in Former Premature Infants," *Anesthesiology,* vol. 68, pp. 796-798, May 1988.

[41] L. G. Welborn, R. S. Hannallah, N. L. Luban, R. Fink, and U. E. Ruttimann, "Anemia and Postoperative Apnea in Former Preterm Infants," *Anesthesiology,* vol. 74, pp. 1003-1006, Jun 1991.

[42] L. G. Welborn, "Post-operative apnoea in the former preterm infant: a review," *Pediatric Anesthesia,* vol. 2, pp. 37-44, 1992.

[43] D. A. Dransfield, A. R. Spitzer, and W. W. Fox, "Episodic airway obstruction in premature infants," *Am J Dis Child,* vol. 137, pp. 441-443, May 1983.

[44] A. D. Milner, A. W. Boon, R. A. Saunders, and I. E. Hopkin, "Upper airways obstruction and apnoea in preterm babies," *Arch Dis Child,* vol. 55, pp. 22-25, January 1, 1980 1980.

[45] B. Naylor, J. Radhakrishnan, and D. Mclaughlin, "Postoperative apnea in infants," *J Pediatr Surg,* vol. 27, pp. 955-957, Aug 1992.

[46] A. Davidson, G. P. Frawley, S. Sheppard, R. Hunt, and P. Hardy, "Risk factors for apnea after infant inguinal hernia repair," *Paediatr Anaesth,* vol. 19, pp. 402-403, Apr 2009.

[47] J. E. Tetzlaff, D. W. Annand, M. A. Pudimat, and H. F. Nicodemus, "Postoperative Apnea in a Full-term Infant," *Anesthesiology,* vol. 69, pp. 426-427, Sep 1988.

[48] J. Karayan, L. Lacoste, and J. Fusciardi, "Postoperative Apnea in a Full-term Infant," *Anesthesiology,* vol. 75, pp. 375-375, Aug 1991.

[49] D. B. Andropoulos, M. B. Heard, K. L. Johnson, J. T. Clarke, and R. W. Rowe, "Postanesthetic Apnea in Full-term Infants after Pyloromyotomy," *Anesthesiology,* vol. 80, pp. 216-219, Jan 1994.

[50] J. L. Galinkin, P. J. Davis, F. X. Mcgowan, A. M. Lynn, M. F. Rabb, M. Yaster, L. G. Henson, R. Blum, D. Hechtman, L. Maxwell, P. Szmuk, R. Orr, E. J. Krane, S. Edwards, and C. D. Kurth, "A Randomized Multicenter Study of Remifentanil Compared with Halothane in Neonates and Infants Undergoing Pyloromyotomy. II. Perioperative Breathing Patterns in Neonates and Infants with Pyloric Stenosis," *Anesth Analg,* vol. 93, pp. 1387-1392, Dec 2001.

[51] G. Gollin, C. Bell, R. Dubose, R. J. Touloukian, J. H. Seashore, C. W. Hughes, T. H. Oh, J. Fleming, and T. O'connor, "Predictors of postoperative respiratory complications in

premature infants after inguinal herniorrhaphy," *J Pediatr Surg,* vol. 28, pp. 244-247, Feb 1993.

[52] E. J. Krane, C. M. Haberkern, and L. E. Jacobson, "Postoperative Apnea, Bradycardia, and Oxygen Desaturation in Formerly Premature Infants: Prospective Comparison of Spinal and General Anesthesia," *Anesth Analg,* vol. 80, pp. 7-13, Jan 1995.

[53] K. H. Sartorelli, J. Christian Abajian, J. M. Kreutz, and D. W. Vane, "Improved outcome utilizing spinal anesthesia in high-risk infants," *J Pediatr Surg,* vol. 27, pp. 1022-1025, Aug 1992.

[54] K. Konno and J. Mead, "Measurement of the separate volume changes of rib cage and abdomen during breathing," *J Appl Physiol,* vol. 22, pp. 407-422, Mar 1967.

[55] R. K. Millard, "Inductive plethysmography components analysis and improved non-invasive postoperative apnoea monitoring," *Physiol Meas,* vol. 20, p. 175, May 1999.

[56] R. K. Millard, "Key to better qualitative diagnostic calibrations in respiratory inductive plethysmography," *Physiol Meas,* vol. 23, pp. N1-8, May 2002.

[57] J. H. T. Bates, M. J. Turner, C. J. Lanteri, B. Jonson, and P. D. Sly, "Measurement of Flow and Volume," in *Infant Respiratory Function Testing*, J. Stocks*, et al.*, Eds., ed New York: Wiley-Liss, Inc., 1996, pp. 81-116.

[58] S. Stick, "Measurements During Tidal Breathing," in *Infant Respiratory Function Testing*, J. Stocks*, et al.*, Eds., ed New York: Wiley-Liss, Inc., 1996, pp. 117-138.

[59] J. A. Adams, "Respiratory Inductive Plethysmography," in *Infant Respiratory Function Testing*, J. Stocks*, et al.*, Eds., ed New York: Wiley-Liss, Inc., 1996, pp. 139-164.

[60] C. L. Herry, D. Townsend, G. C. Green, A. Bravi, and A. J. Seely, "Segmentation and classification of capnograms: application in respiratory variability analysis," *Physiol Meas,* vol. 35, pp. 2343-58, Dec 2014.

[61] D. B. Raemer and I. Calalang, "Accuracy of end-tidal carbon dioxide tension analyzers," *J Clin Monit,* vol. 7, pp. 195-208, Apr 1991.

[62] K. Zwerneman, "End-Tidal Carbon Dioxide Monitoring: A VITAL Sign Worth Watching," *Crit Care Nurs Clin North Am,* vol. 18, pp. 217-225, Jun 2006.

[63] E. S. Goldensohn and L. Zablow, "An electrical impedance spirometer," *J Appl Physiol,* vol. 14, pp. 463-464, 1959-05-01 1959.

[64] A. F. Pacela, "Impedance pneumography—A survey of instrumentation techniques," *Medical & Biological Engineering,* vol. 4, pp. 1-15, 1966/01/01 1966.

[65] A. Grenvik, S. Ballou, E. Mcginley, J. E. Millen, W. L. Cooley, and P. Safar, "Impedance pneumography: Comparison between chest impedance changes and respiratory volumes in ii healthy volunteers," *Chest,* vol. 62, pp. 439-443, 1972.

[66] L. A. Geddes, H. E. Hove, D. M. Hickman, and A. G. Moore, "The impedance pneumograph," *Aerospace Med,* vol. 33, pp. 28-33, 1962.

[67] I. M. Stein and D. C. Shannon, "The Pediatric Pneumogram: A New Method for Detecting and Quantitating Apnea in Infants," *Pediatrics,* vol. 55, pp. 599-603, May 1, 1975 1975.

[68] J. A. Adams, I. A. Zabaleta, D. Stroh, and M. A. Sackner, "Measurement of breath amplitudes: Comparison of three noninvasive respiratory monitors to integrated pneumotachograph," *Pediatr Pulmonol,* vol. 16, pp. 254-258, Oct 1993.

[69] H. L. Watson, M. A. Sackner, and F. D. Stott, "Method and apparatus for monitoring respiration," 1982.

[70] D. E. Weese-Mayer, M. J. Corwin, M. R. Peucker, J. M. Di Fiore, D. R. Hufford, L. R. Tinsley, M. R. Neuman, R. J. Martin, L. J. Brooks, S. L. D. Ward, G. Lister, M. Willinger, and C. S. Grp, "Comparison of apnea identified by respiratory inductance plethysmography with that detected by end-tidal CO2 or thermistor," *Am J Resp Crit Care,* vol. 162, pp. 471-480, August 1, 2000.

[71] K. A. Brown, R. Platt, and J. H. T. Bates, "Automated analysis of paradoxical ribcage motion during sleep in infants," *Pediatr Pulmonol,* vol. 33, pp. 38-46, Jan 2002.

[72] G. M. Nixon and R. T. Brouillette, "Diagnostic techniques for obstructive sleep apnoea: is polysomnography necessary?," *Paediatr Respir Rev,* vol. 3, pp. 18-24, 2002.

[73] H. L. Watson, D. A. Poole, and M. A. Sackner, "Accuracy of respiratory inductive plethysmographic cross-sectional areas," *J Appl Physiol,* vol. 65, pp. 306-308, July 1, 1988 1988.

[74] D. Stagg, M. Goldman, and J. N. Davis, "Computer-aided measurement of breath volume and time components using magnetometers," *J Appl Physiol,* vol. 44, pp. 623-633, 1978-04-01 1978.

[75] J. D. Sackner, A. J. Nixon, B. Davis, N. Atkins, and M. A. Sackner, "Non-invasive Measurement of Ventilation During Exercise Using A Respiratory Inductive Plethysmograph," *Am Rev Respir Dis,* vol. 122, pp. 867-871, 1980/12/01 1980.

[76] T. S. Chadha, H. Watson, S. Birch, G. A. Jenouri, A. W. Schneider, M. A. Cohn, and M. A. Sackner, "Validation of Respiratory Inductive Plethysmography Using Different Calibration Procedures," *Am Rev Respir Dis,* vol. 125, pp. 644-649, 1982/06/01 1982.

[77] M. A. Sackner, H. Watson, A. S. Belsito, D. Feinerman, M. Suarez, G. Gonzalez, F. Bizousky, and B. Krieger, "Calibration of respiratory inductive plethysmograph during natural breathing," *J Appl Physiol,* vol. 66, pp. 410-420, January 1, 1989 1989.

[78] P. V. Zimmerman, S. J. Connellan, H. C. Middleton, M. V. Tabona, M. D. Goldman, and N. Pride, "Postural Changes in Rib Cage and Abdominal Volume-Motion Coefficients and Their Effect on the Calibration of a Respiratory Inductance Plethysmograph," *Am Rev Respir Dis,* vol. 127, pp. 209-214, 1983/02/01 1983.

[79] A. De Groote, M. Paiva, and Y. Verbandt, "Mathematical assessment of qualitative diagnostic calibration for respiratory inductive plethysmography," *J Appl Physiol,* vol. 90, pp. 1025-1030, March 1, 2001 2001.

[80] A. De Groote, J. Groswasser, H. Bersini, P. Mathys, and A. Kahn, "Detection of obstructive apnea events in sleeping infants from thoracoabdominal movements," *J Sleep Res,* vol. 11, pp. 161-168, Jun 2002.

[81] D. Bliwise, N. G. Bliwise, H. C. Kraemer, and W. Dement, "Measurement error in visually scored electrophysiological data: respiration during sleep," *J Neurosci Methods,* vol. 12, pp. 49-56, Nov 1984.

[82] C. A. Robles-Rubio, R. E. Kearney, and K. A. Brown, "Automated pause frequency estimation to assess the risk of Postoperative Apnea in infants," presented at the 12th Int. Symp. Sleep Breath., Barcelona, Spain, 2011.

[83] S. M. Kay, *Fundamentals of statistical signal processing* vol. 2. Upper Saddle River, NJ: Prentice Hall PTR, 1998.

[84] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recog.,* vol. 30, pp. 1145-1159, 1997.

[85] A. A. Aoude, R. E. Kearney, K. A. Brown, H. L. Galiana, and C. A. Robles-Rubio, "Automated Off-Line Respiratory Event Detection for the Study of Postoperative Apnea in Infants," *IEEE Trans Biomed Eng,* vol. 58, pp. 1724-1733, Jun 2011.

[86] P. M. Macey, R. P. K. Ford, P. J. Brown, J. Larkin, W. R. Fright, and K. L. Garden, "Apnoea detection: human performance and reliability of a computer algorithm," *Acta Paediatr,* vol. 84, pp. 1103-1107, Oct 1995.

[87] H. Lee, C. G. Rusin, D. E. Lake, M. T. Clark, L. Guin, T. J. Smoot, A. O. Paget-Brown, B. D. Vergales, J. Kattwinkel, J. R. Moorman, and J. B. Delos, "A new algorithm for detecting central apnea in neonates," *Physiol Meas,* vol. 33, pp. 1-17, Jan 2012.

[88] M. Bsoul, H. Minn, and L. Tamil, "Apnea MedAssist: Real-time Sleep Apnea Monitor Using Single-Lead ECG," *IEEE Trans Inf Technol Biomed,* vol. 15, pp. 416-427, May 2011.

[89] P. M. Macey, J. S. J. Li, and R. P. K. Ford, "Expert system for the detection of apnoea," *Eng Appl Artif Intel,* vol. 11, pp. 425-438, Jun 1998.

[90] D. Álvarez-Estévez and V. Moret-Bonillo, "Fuzzy reasoning used to detect apneic events in the sleep apnea-hypopnea syndrome," *Expert Syst Appl,* vol. 36, pp. 7778-7785, 5// 2009.

[91] J. Han, H.-B. Shin, D.-U. Jeong, and K. S. Park, "Detection of apneic events from single channel nasal airflow using 2nd derivative method," *Comput Methods Programs Biomed,* vol. 91, pp. 199-207, Sep 2008.

[92] P. J. Van Houdt, P. P. W. Ossenblok, M. G. Van Erp, K. E. Schreuder, R. J. J. Krijn, P. a. J. M. Boon, and P. J. M. Cluitmans, "Automatic breath-to-breath analysis of nocturnal polysomnographic recordings," *Med Biol Eng Comput,* vol. 49, pp. 819-830, Jul 2011.

[93] P. Varady, T. Micsik, S. Benedek, and Z. Benyo, "A novel method for the detection of apnea and hypopnea events in respiration signals," *IEEE Trans Biomed Eng,* vol. 49, pp. 936-942, Sep 2002.

[94] P. De Chazal, C. Heneghan, E. Sheridan, R. Reilly, P. Nolan, and M. O'malley, "Automated processing of the single-lead electrocardiogram for the detection of obstructive sleep apnoea," *IEEE Trans Biomed Eng,* vol. 50, pp. 686-696, Jun 2003.

[95] A. H. Khandoker, J. Gubbi, and M. Palaniswami, "Automated Scoring of Obstructive Sleep Apnea and Hypopnea Events Using Short-Term Electrocardiogram Recordings," *IEEE Trans Inf Technol Biomed,* vol. 13, pp. 1057-1067, Nov 2009.

[96] B. Xie and H. Minn, "Real-Time Sleep Apnea Detection by Classifier Combination," *IEEE Trans Inf Technol Biomed,* vol. 16, pp. 469-477, May 2012.

[97] P. M. Macey, J. S. J. Li, and R. P. K. Ford, "Deterministic properties of apnoeas in an abdominal breathing signal," *Med Biol Eng Comput,* vol. 37, pp. 335-343, May 1999.

[98] P. M. Macey, "Apnoea detection," Ph.D., Electrical and Electronic Engineering, University of Canterbury, Christchurch, New Zealand, 1998.

[99] C. Zamarrón, F. Gude, J. Barcala, J. R. Rodriguez, and P. V. Romero, "Utility of Oxygen Saturation and Heart Rate Spectral Analysis Obtained From Pulse Oximetric Recordings in the Diagnosis of Sleep Apnea Syndrome*," *Chest,* vol. 123, pp. 1567-1576, 2003.

[100] J. E. Mietus, C. K. Peng, P. C. Ivanov, and A. L. Goldberger, "Detection of obstructive sleep apnea from cardiac interbeat interval time series," in *Computers in Cardiology 2000*, 2000, pp. 753-756.

[101] E. Gil, R. Bailon, J. M. Vergara, and P. Laguna, "PTT Variability for Discrimination of Sleep Apnea Related Decreases in the Amplitude Fluctuations of PPG Signal in Children," *IEEE Trans Biomed Eng,* vol. 57, pp. 1079-1088, May 2010.

[102] J. V. Marcos, R. Hornero, D. Álvarez, F. Del Campo, C. Zamarrón, and M. López, "Utility of multilayer perceptron neural network classifiers in the diagnosis of the obstructive sleep apnoea syndrome from nocturnal oximetry," *Comput Methods Programs Biomed,* vol. 92, pp. 79-89, Oct 2008.

[103] H. M. Al-Angari and A. V. Sahakian, "Automated Recognition of Obstructive Sleep Apnea Syndrome Using Support Vector Machine Classifier," *IEEE Trans Inf Technol Biomed,* vol. 16, pp. 463-468, May 2012.

[104] D. Álvarez, R. Hornero, M. García, F. Del Campo, and C. Zamarrón, "Improving diagnostic ability of blood oxygen saturation from overnight pulse oximetry in obstructive sleep apnea detection by means of central tendency measure," *Artif Intell Med,* vol. 41, pp. 13-24, Sep 2007.

[105] F. Yasuma and J.-I. Hayano, "Respiratory sinus arrhythmia*: Why does the heartbeat synchronize with respiratory rhythm?," *Chest,* vol. 125, pp. 683-690, 2004.

[106] C. Guilleminault, R. Winkle, S. Connolly, K. Melvin, and A. Tilkian, "Cyclical Variation of the Heart-Rate in Sleep-Apnoea Syndrome: Mechanisms, and Usefulness of 24 h Electrocardiography as a Screening Technique," *Lancet,* vol. 323, pp. 126-131, 1/21/ 1984.

[107] L. J. Epstein and G. R. Dorlac, "Cost-effectiveness analysis of nocturnal oximetry as a method of screening for sleep apnea-hypopnea syndrome," *Chest,* vol. 113, pp. 97-103, Jan 1998.

[108] G. B. Moody, R. G. Mark, A. Zoccola, and S. Mantero, "Derivation of respiratory signals from multi-lead ECGs," in *Computers in Cardiology 1985*. vol. 12, ed Washington, DC: IEEE Computer Society Press, 1985, pp. 113-116.

[109] A. Johansson and P. Å. Öberg, "Estimation of respiratory volumes from the photoplethysmographic signal. Part I: experimental results," *Med Biol Eng Comput,* vol. 37, pp. 42-47, 1999/01/01 1999.

[110] Y. Sivan, S. D. Ward, T. Deakers, T. G. Keens, and C. J. L. Newth, "Rib cage to abdominal asynchrony in children undergoing polygraphic sleep studies," *Pediatr Pulmonol,* vol. 11, pp. 141-146, 1991.

[111] D. Andersson, G. Gennser, and P. Johnson, "Phase characteristics of breathing movements in healthy newborns," *J Dev Physiol,* vol. 5, pp. 289-298, Oct 1983.

[112] M. Benameur, M. D. Goldman, C. Ecoffey, and C. Gaultier, "Ventilation and thoracoabdominal asynchrony during halothane anesthesia in infants," *Journal of Applied Physiology,* vol. 74, pp. 1591-1596, April 1, 1993 1993.

[113] K. A. Brown, "Pattern of ventilation during halothane anaesthesia in infants less than two months of age," *Can J Anaesth,* vol. 43, pp. 121-128, Feb 1996.

[114] K. Brown, C. Aun, J. Stocks, E. Jackson, A. Mackersie, and D. Hatch, "A comparison of the respiratory effects of sevoflurane and halothane in infants and young children," *Anesthesiology,* vol. 89, pp. 86-92, Jul 1998.

[115] R. C. Pascucci, M. B. Hershenson, N. F. Sethna, S. H. Loring, and A. R. Stark, "Chest wall motion of infants during spinal anesthesia," *J Appl Physiol,* vol. 68, pp. 2087-2091, May 1990.

[116] G. K. Prisk, J. Hammer, and C. J. L. Newth, "Techniques for measurement of thoracoabdominal asynchrony," *Pediatric Pulmonology,* vol. 34, pp. 462-472, 2002.

[117] J. L. Allen, M. R. Wolfson, K. Mcdowell, and T. H. Shaffer, "Thoracoabdominal asynchrony in infants with airflow obstruction," *Am Rev Respir Dis,* vol. 141, pp. 337-342, Feb 1990.

[118] R. Ramanathan, M. J. Corwin, C. E. Hunt, G. Lister, L. R. Tinsley, T. Baird, J. M. Silvestri, D. H. Crowell, D. Hufford, R. J. Martin, M. R. Neuman, D. E. Weese-Mayer, L. A. Cupples, M. Peucker, M. Willinger, T. G. Keens, and For the Collaborative Home Infant Monitoring Evaluation Study Group, "Cardiorespiratory Events Recorded on Home Monitors: Comparison of Healthy Infants With Those at Increased Risk for SIDS," *J Am Med Assoc,* vol. 285, pp. 2199-2207, May 2, 2001 2001.

[119] D. Precup, C. A. Robles-Rubio, K. A. Brown, L. Kanbar, J. Kaczmarek, S. Chawla, G. M. Sant'anna, and R. E. Kearney, "Prediction of Extubation Readiness in Extreme Preterm Infants Based on Measures of Cardiorespiratory Variability," in *Conf. Proc. 34th IEEE Eng. Med. Biol. Soc.*, San Diego, USA, 2012, pp. 5630-5633.

[120] C. A. Robles-Rubio, J. Kaczmarek, S. Chawla, L. Kovacs, K. A. Brown, R. E. Kearney, and G. M. Sant Anna, "Automated analysis of respiratory behavior in extremely preterm infants and extubation readiness," *Pediatr Pulmonol,* vol. 50, pp. 479-486, May 2015.

[121] J. Dall'ava-Santucci and A. Armanganidis, "Respiratory Inductive Plethysmography," in *Pulmonary Function in Mechanically Ventilated Patients*. vol. 13, S. Benito and A. Net, Eds., ed: Springer Berlin Heidelberg, 1991, pp. 121-142.

[122] J. L. Allen, J. S. Greenspan, K. S. Deoras, E. Keklikian, M. R. Wolfson, and T. H. Shaffer, "Interaction between chest wall motion and lung mechanics in normal infants and infants with bronchopulmonary dysplasia," *Pediatric Pulmonology,* vol. 11, pp. 37-43, 1991.

[123] W. W. Flemons, N. J. Douglas, S. T. Kuna, D. O. Rodenstein, and J. Wheatley, "Access to Diagnosis and Treatment of Patients with Suspected Sleep Apnea," *Am J Respir Crit Care Med,* vol. 169, pp. 668-672, 2004/03/15 2004.

[124] C. Iber, S. Ancoli-Israel, A. L. J. Chesson, and S. F. Quan, "The New Sleep Scoring Manual-The Evidence Behind The Rules," *J Clin Sleep Med,* vol. 3, p. 107, 2007.

[125] T. Penzel, M. Hirshkowitz, J. Harsh, R. D. Chervin, N. Butkov, M. Kryger, B. Malow, M. V. Vitiello, M. H. Silber, C. A. Kushida, and A. L. J. Chesson, "Digital Analysis and Technical Specifications," *J Clin Sleep Med,* vol. 3, pp. 109-120, Mar 15 2007.

[126] M. H. Silber, S. Ancoli-Israel, M. H. Bonnet, S. Chokroverty, M. M. Grigg-Damberger, M. Hirshkowitz, S. Kapen, S. A. Keenan, M. H. Kryger, T. Penzel, M. R. Pressman, and C. Iber, "The Visual Scoring of Sleep in Adults," *J Clin Sleep Med,* vol. 3, p. 22, Mar 15 2007.

[127] M. H. Bonnet, K. Doghramji, T. Roehrs, E. J. Stepanski, S. H. Sheldon, A. S. Walters, M. Wise, and A. L. J. Chesson, "The Scoring of Arousal in Sleep: Reliability, Validity, and Alternatives," *J Clin Sleep Med,* vol. 3, pp. 133-145, Mar 15 2007.

[128] S. M. Caples, C. L. Rosen, W. K. Shen, A. S. Gami, W. Cotts, M. Adams, P. Dorostkar, K. Shivkumar, V. K. Somers, T. I. Morgenthaler, E. J. Stepanski, and C. Iber, "The Scoring of Cardiac Events During Sleep," *J Clin Sleep Med,* vol. 3, pp. 147-154, Mar 15 2007.

[129] A. S. Walters, G. Lavigne, W. Hening, D. L. Picchietti, R. P. Allen, S. Chokroverty, C. A. Kushida, D. L. Bliwise, M. W. Mahowald, C. H. Schenck, and S. Ancoli-Israel, "The Scoring of Movements in Sleep," *J Clin Sleep Med,* vol. 3, pp. 155-167, Mar 15 2007.

[130] S. Redline, R. Budhiraja, V. Kapur, C. L. Marcus, J. H. Mateika, R. Mehra, S. Parthasarthy, V. K. Somers, K. P. Strohl, L. G. Sulit, D. Gozal, M. S. Wise, and S. F. Quan, "The Scoring of Respiratory Events in Sleep: Reliability and Validity," *J Clin Sleep Med,* vol. 3, pp. 169-200, Mar 15 2007.

[131] M. Grigg-Damberger, D. Gozal, C. L. Marcus, S. F. Quan, C. L. Rosen, R. D. Chervin, M. Wise, D. L. Picchietti, S. H. Sheldon, and C. Iber, "The Visual Scoring of Sleep and Arousal in Infants and Children," *J Clin Sleep Med,* vol. 3, pp. 201-240, Mar 15 2007.

[132] C. A. Robles-Rubio, R. E. Kearney, and K. A. Brown, "Inclusion of Lissajous Plot on Scoring Software Improves Classification of Thoracoabdominal Asynchrony," presented at the 13th Int. Symp. Sleep Breath., Montreal, Canada, 2013.

[133] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychol. Bulletin,* vol. 76, pp. 378-382, 1971.

[134] G. Cardillo, "Fleiss'es kappa: compute the Fleiss'es kappa for multiple raters.," ed. MATLAB CENTRAL: The MathWorks, Inc., 2007.

[135] C. A. Robles-Rubio, K. A. Brown, and R. E. Kearney, "Automated Unsupervised Respiratory Event Analysis," in *Conf. Proc. 33rd IEEE Eng. Med. Biol. Soc.*, Boston, USA, 2011, pp. 3201-3204.

[136] J. Cohen, "A coefficient of agreement for nominal scales," *Educ Psychol Meas,* vol. 20, pp. 37-46, 1960.

[137] C. A. Robles-Rubio, K. A. Brown, and R. E. Kearney, "Detection of Breathing Segments in Respiratory Signals," in *Conf. Proc. 34th IEEE Eng. Med. Biol. Soc.*, San Diego, USA, 2012, pp. 6333-6336.

[138] C. A. Robles-Rubio, K. A. Brown, and R. E. Kearney, "A New Movement Artifact Detector for Photoplethysmographic Signals," in *Conf. Proc. 35th IEEE Eng. Med. Biol. Soc.*, Osaka, Japan, 2013, pp. 2295 - 2299.

[139] C. A. Robles-Rubio, K. A. Brown, G. Bertolizio, and R. E. Kearney, "Automated Analysis of Respiratory Behavior for the Prediction of Apnea in Infants following General Anesthesia," in *Conf. Proc. 36th IEEE Eng. Med. Biol. Soc.*, Chicago IL, USA, 2014, pp. 262-265.

[140] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics,* vol. 33, pp. 159-174, 1977.

[141] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bull.,* vol. 1, pp. 80-83, 1945.

[142] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*. New York [etc.]: Chapman & Hall, 1993.

[143] S. Warfield, J. Dengler, J. Zaers, C. R. G. Guttmann, W. M. Wells, G. J. Ettinger, J. Hiller, and R. Kikinis, "Laboratory Investigation:Automatic Identification of Gray Matter Structures from MRI to Improve the Segmentation of White Matter Lesions," *Journal of Image Guided Surgery,* vol. 1, pp. 326-338, Jan 1995.

[144] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning From Crowds," *J. Mach. Learn. Res.,* vol. 11, pp. 1297-1322, 2010.

[145] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological),* vol. 39, pp. 1-38, 1977.

[146] A. P. Dawid and A. M. Skene, "Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm," *J Roy Stat Soc C-App,* vol. 28, pp. 20-28, 1979.

[147] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation," *IEEE Trans Med Imaging,* vol. 23, pp. 903-921, Jul 2004.

[148] C. F. J. Wu, "On the Convergence Properties of the EM Algorithm," *Ann Stat,* vol. 11, pp. 95-103, 1983.

[149] C. L. Marcus, S. England, R. D. Annett, L. J. Brooks, R. T. Brouillette, J. L. Carroll, D. Givan, D. Gozal, J. Kiley, S. Redline, C. L. Rosen, G. Rosen, and D. Tunkel, "Cardiorespiratory Sleep Studies in Children. Establishment of Normative Data and

Polysomnographic Predictors of Morbidity," *Am J Resp Crit Care,* vol. 160, pp. 1381-1387, October 1, 1999.

[150] P. Martinot-Lagarde, R. Sartene, M. Mathieu, and G. Durand, "What does inductance plethysmography really measure?," *J Appl Physiol,* vol. 64, pp. 1749-1756, April 1, 1988.

[151] S. V. Jacob, A. Morielli, M. A. Mograss, F. M. Ducharme, M. D. Schloss, and R. T. Brouillette, "Home testing for pediatric obstructive sleep apnea syndrome secondary to adenotonsillar hypertrophy," *Pediatr Pulmonol,* vol. 20, pp. 241-252, Oct 1995.

[152] G. A. Little, R. A. Ballard, J. G. Brooks, R. T. Brouillette, L. Culpepper, H. B. J. Gray, P. King, M. O. Kolb, A. Neale, M. R. Neuman, D. R. Peterson, S. O. Schwietzer, and H. Weiss, "National Institutes of Health Consensus Development Conference on Infantile Apnea and Home Monitoring, Sept 29 to Oct 1, 1986," *Pediatrics,* vol. 79, pp. 292-299, Feb 1987.

[153] A. Kahn, D. Blum, E. Rebuftat, M. Sottiaux, J. Levitt, A. Bochner, M. Alexander, J. Grosswasser, and M. F. Muller, "Polysomnographic Studies of Infants Who Subsequently Died of Sudden Infant Death Syndrome," *Pediatrics,* vol. 82, pp. 721-727, November 1, 1988.

[154] J. E. Yount, "Technical problems in recognizing and monitoring infant apnea," *Conf. Proc. IEEE Eng. Med. Biol. Soc.,* vol. 11, pp. 325-326, 9-12 Nov 1989 1989.

[155] S. Lord, B. Sawyer, D. Pond, D. O'connell, A. Eyland, A. Mant, M. J. Hensley, and N. A. Saunders, "Interrater reliability of computer-assisted scoring of breathing during sleep," *Sleep,* vol. 12, pp. 550-8, Dec 1989.

[156] T. Al-Ani, Y. Hamam, R. Fodil, F. Lofaso, and D. Isabey, "Using hidden Markov models for sleep disordered breathing identification," *Simul Model Pract Th,* vol. 12, pp. 117-128, May 2004.

[157] F. Steimann and K.-P. Adlassnig, "Clinical monitoring with fuzzy automata," *Fuzzy Set Syst,* vol. 61, pp. 37-42, Jan 10 1994.

[158] V. Kaplan, J. N. Zhang, E. W. Russi, and K. E. Bloch, "Detection of inspiratory flow limitation during sleep by computer assisted respiratory inductive plethysmography," *Eur Respir J,* vol. 15, pp. 570-578, March 1, 2000 2000.

[159] M. P. Villa, S. Piro, A. Dotta, E. Bonci, P. Scola, B. Paggi, M. G. Paglietti, F. Midulla, and R. Ronchetti, "Validation of automated sleep analysis in normal children," *Eur Respir J,* vol. 11, pp. 458-461, February 1, 1998 1998.

[160] A. W. Lo, "Maximum Likelihood Estimation of Generalized Itô Processes with Discretely Sampled Data," *Econometric Theory,* vol. 4, pp. 231-247, 1988.

[161] K. S. Deoras, J. S. Greenspan, M. R. Wolfson, E. N. Keklikian, T. H. Shaffer, and J. L. Allen, "Effects of Inspiratory Resistive Loading on Chest Wall Motion and Ventilation: Differences between Preterm and Full-Term Infants," *Pediatric Research,* vol. 32, pp. 589-594, Nov 1992.

[162] J. Macqueen, "Some methods for classification and analysis of multivariate observations," in *Proc. Fifth Berkeley Symp. on Math, Statistics, and Probability*, Berkeley, CA, USA, 1967, pp. 281-297.

[163] K. Tanioka and H. Yadohisa, "Effect of Data Standardization on the Result of k-Means Clustering," in *Challenges at the Interface of Data Analysis, Computer Science, and Optimization*, W. A. Gaul*, et al.*, Eds., ed: Springer Berlin Heidelberg, 2012, pp. 59-67.

[164] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: special issue on learning from imbalanced data sets," *SIGKDD Explor Newsl,* vol. 6, pp. 1-6, 2004.

[165] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J Artif Intell Res,* vol. 16, pp. 321-357, 2002.

[166] F. Committee On, Newborn, and P. American Academy Of, "Apnea, Sudden Infant Death Syndrome, and Home Monitoring," *Pediatrics,* vol. 111, pp. 914-917, April 1, 2003 2003.

[167] D. J. Henderson-Smart and P. A. Steer, "Prophylactic caffeine to prevent postoperative apnoea following general anaesthesia in preterm infants," *Cochrane Database of Systematic Reviews,* 2001.

[168] J. M. Williams, P. A. Stoddart, S. a. R. Williams, and A. R. Wolf, "Post-operative recovery after inguinal herniotomy in ex-premature infants: comparison between sevoflurane and spinal anaesthesia," *Br J Anaesth,* vol. 86, pp. 366-371, March 1, 2001 2001.

[169] P. D. Craven, N. Badawi, D. J. Henderson-Smart, and M. O'brien, "Regional (spinal, epidural, caudal) versus general anaesthesia in preterm infants undergoing inguinal herniorrhaphy in early infancy," *Cochrane Database Syst Rev,* p. CD003669, 2003.

[170] L. Kanbar, W. Shalish, C. A. Robles-Rubio, D. Precup, K. Brown, G. M. Sant'anna, and R. E. Kearney, "Correlation of Clinical Parameters with Cardiorespiratory Behavior in Successfully Extubated Extremely Preterm Infants," in *Conf. Proc. 37th IEEE Eng. Med. Biol. Soc.*, Milan, Italy, 2015, pp. 4431-4434.