Predicting sepsis on the Neonatal Intensive Care Unit

Anna Mank

13/03/2017

Predicting sepsis on the Neonatal Intensive Care Unit

By

A.B. Mank

Technical University Delft
Department of Biomedical Engineering
4310691

Supervisors:

Ir. T.G. Goos

Technical University Delft
Department of Biomechanical Engineering
Erasmus Medical Centre, Sophia Children's Hospital
Department of Neonatology

Prof. Dr. J. Dankelman

Technical University Delft
Department of Biomechanical Engineering

Acknowledgement

I would like to express my gratitude towards my supervisor Ir. Tom Goos who gave me the opportunity to conduct my research at the Sophia Children's Hospital, gave me feedback and great insight in how to proceed. Besides I would like to thank Prof. Dr. J. Dankelman who put me in touch with Ir. Tom Goos and guided me through the process of graduation. For all the help with my questions about the statistics I would like to thank Katya Mauff.

I'd also like to thank my colleagues from the 'Z-flat', the technical department at the Sophia Children's Hospital and my dear friends & family.

Anna Mank
Rotterdam, March 2017

Preface

This thesis is written as the concluding part of my master Biomedical Engineering at the Technical University Delft. The aim of this study was to provide a predictive model to detect the incidence of sepsis in premature infants during hospitalization. The hypothesis was that a combination of non-invasive clinical measurements and heart rate variability could predict sepsis. If this prediction could precede the standard method to indicate sepsis, this could ensure the earlier start of treatment of the premature infants. Based on this research, a new continuous indicator can be developed to possibly prevent deaths due to sepsis in the future. Data of premature infants hospitalized in the first half year of 2016 is analysed on the relation between sepsis, HRV and other clinical predictors.

This thesis consist of a scientific paper about the research, a general introduction into heart rate variability, a section about the method of data collection and analysis, the more extensive results and an overall conclusion with recommendation.

In the appendix the build Matlab and R code can be found.

Content

Acknowledgement II
Preface II
List of figures IV
List of tables IV
List of abbreviations and definitions V
1. Scientific paper 1
2. Heart Rate Variability 12
    1. Time domain 13
    2. Frequency domain 15
    3. Non-linear 17
    4. Literature study 18
3. Data analysis 20
    Content 20
    1. Cohort selection 20
    2. Data collection 21
    3. ECG 23
    4. Covariates 25
    5. Statistical analysis 27
4. Results 31
    1. Study population 31
    2. Heart Rate Variability 32
    3. Selection of database 32
    4. Generalized Logistic Mixed Effect Regression 34
5. Recommendations 35
    General conclusion 36
References 37
Appendix 39
    Matlab 39
    R 54

List of figures

Figure 1 Study inclusion flow chart....5
Figure 2 Alterations in the raw RR interval over time preceding sepsis. Raw data of one patient is presented....6
Figure 3 RR interval (A) and HRV (B) prior to sepsis, with HRV for non-septic patients (C) as a reference. Data is averaged over three hours and presented as IQR (Q1:25%, Median, Q3:75%)....6
Figure 4 AIC and BIC values....7
Figure 5 The odds ratio of sepsis against measurement time, data of the septic group (n=18) is presented....8
Figure 6 The relation between HRV and the odds of sepsis with the histogram of the HRV in the upper right corner....8
Figure 7 The odds ratio over time for the control group....9
Figure 8 Typical ECG characteristics with the QRS complex and RR interval visualised [20]....12
Figure 9 Explanation of formula (8) [8]....14
Figure 10 Poincaré plot [28]....17
Figure 11 Derivation of RR interval from ECG signal [25]....23
Figure 12 Birth weight growth curve [40]....26
Figure 13 AIC and BIC for different HRV measures....32
Figure 14 RR and HRV compared between patients with and without sepsis, data is averaged over six hours and presented as IQR (Q1: 25% and Q3:75%)....33
Figure 15 Boxplot of HRV for the sepsis and healthy group....34

List of tables

Table 1 Baseline characteristics of the study population; exclusion based on NEC, EONS, surgical procedure or missing data....5
Table 2 Classification table of the three tested models....7
Table 3 Time domain measures of HRV [8, 9]; milliseconds (ms)....13
Table 4 Frequency domain measures of HRV [8, 9]....15
Table 5 Nonlinear measures of HRV [8, 9]....17
Table 6 Literature study....19
Table 7 Waveform output....22
Table 8 Baseline covariates....25
Table 9 Time varying covariates....25
Table 10 Response variable....27
Table 11 Example of the database for patient 3....27
Table 12 The different bacteria that cause sepsis....31
Table 13 Sepsis characteristics....32
Table 14 Demographics of the patients included in the first database....32
Table 15 Classification table with a cut off value of 0.3....33

List of abbreviations and definitions

AIC	Akaike Information Criterion (section 3.5)
b	coefficient estimate
BIC	Bayesian Information Criterion (section 3.5)
CRP	C-Reactive Protein is a biomarker used to diagnose infection since it's triggered by increased cytokines levels within 4 to 6 hours of an inflammatory stimulus.
EONS	Early ONset Sepsis develops within 72 hours after birth (section 3.1).
f_{peak}	Peak frequency
f_{resp}	Respiratory frequency
GA	Gestational Age is the duration of pregnancy measured from the first day of the last normal menstrual period
GEE	Generalized Estimating Equations (section 3.5)
HF	High Frequencies
HR	Heart Rate
HRV	Heart Rate Variability
IQR	Inter Quartile Range is a measure which divides the data into quartile. The first quartile (Q1) represents the first 25% of the data and the third quartile (Q3) the upper 25%.
LBW	Low Birth Weight
LF	Low Frequencies
LONS	Late Onset Sepsis refers to sepsis representing after the first 72 hours of life.
ms	milliseconds
NEC	Necrotizing Enterocolitis is disease that affects the intestine through the invasion of bacteria. The underdeveloped and weakened intestine becomes inflamed and eventually necrosis of the tissue occurs creating holes in the wall.
NICU	Neonatal Intensive Care Unit is the specialized unit for the intensive care of premature
PID	born infants.
Ptot	Patient Information Dossier
(p)RR50	Total power
ROC	Receiver Operating Curve (section 2.4)
SampEn	Sample Entropy (section 2.3)
SD	Standard Deviation
SDARR	Standard Deviation of all RR intervals (section 2.1)
SDRR	Standard Deviation of the average RR interval over 5 minutes (section 2.1)
SDSD	Standard Deviation of Successive RR interval Differences (section 2.1)
SE	Standard Error
SpO2	Peripheral blood Oxygen Saturation
RMSSD	Root Mean Square of the Successive Difference (section 2.1)
TIRR	Triangular interpolation of RR (section 2.1)
VLF	Very Low Frequencies

1. Scientific paper

In the following section the scientific paper written about the research conducted for this master thesis is presented.

Altered heart rate variability and clinical signs preceding sepsis in premature infants

A.B.Mank

ABSTRACT. *Background and objective.* To reduce the mortality rate of sepsis, treatment needs to be started as soon as possible. One should however first be sure about the occurrence of sepsis. Currently, the suspicion of sepsis in premature infants is based on non-specific physiological changes. The detection of alterations in the inter-heartbeat interval preceding sepsis can offer a solution. The aim of this study was to design a predictive model which captures the relation between physiological changes and variability in the inter heart beat interval, Heart Rate Variability (HRV) and the development of sepsis.

Methods. A retrospective study was performed in the Neonatal Intensive Care Unit (NICU) at the Sophia Children's Hospital from January 2016 to June 2016. The study focussed on very preterm infants, of Gestational Age (GA) <32 weeks, with a birth weight below 1500 gram. Sepsis was defined as a blood culture proven Late-ONset Sepsis (LONS) with elevated levels of C-reactive Protein (CRP). Logistic mixed effect modelling was used to estimate the probability of developing sepsis based on GA, gender, birth weight, CRP, percentage weight, bradycardia, median RR and HRV.

Results. During the inclusion period, 18 of the 60 infants developed at least one event of sepsis. Infants with sepsis had lower birth weight and GA, with higher mortality rates and length of stay. Logistic mixed effect models showed a significant relation between CRP, bradycardia, median RR, HRV and the probability of developing sepsis.

Conclusions. Bradycardia, CRP, median RR and HRV are found to be useful predictors of sepsis. Since the authorisation of a blood culture to prove sepsis takes about four days and the model predicts sepsis at the moment the blood culture is taken, the model is faster in providing results.
Keywords: Premature infant, sepsis, heart rate variability

Introduction

During the year 2012 about 13.000 infants were born premature, before 37 weeks of pregnancy, in the Netherlands. This is 7.4% of the total number of babies born that year [1]. As a cause or as a consequence of the premature birth, almost all of these infants are born with a low birth weight, <1500 gram [2]. Infants born prematurely have spent less time in uterus compared to full-term infants. Consequently, these infants are prone to getting sick and even dying in the neonatal period: the first 28 days of life [1]. Due to better quality of specialized intensive care, foetal, neonatal and perinatal mortality has decreased over the years. Despite of this 0.43 million infants die due to the consequences of sepsis and other severe infections globally [3]. Only second to respiratory distress syndrome, patent ductus arteriosus, retinopathy of prematurity and anaemia, sepsis is the disease with the highest incidence among premature born infants hospitalized on the NICU of the Sophia Children's Hospital (Rotterdam, the Netherlands). Local unpublished data show that the incidence was 20% among infants hospitalized between 2008 and 2014. Sepsis is a bacterial infection in the bloodstream, which causes physiological changes such as temperature instability, altered heart rate and abnormalities of respiration [4]. Since these physiological changes are non-specific, a blood

culture is taken to prove sepsis [5]. Moreover, biomarkers are used for diagnosis, monitoring and prediction of the outcome of sepsis. During infection, CRP levels are increased, making them a suitable indicator of infection [6]. Antibiotic treatment is started immediately when sepsis is proven [5]. In order to decrease the mortality rate of patients who suffer from sepsis, treatment needs to be started as soon as possible. Since the suspicion of sepsis is based on non-specific physiological changes, a more specific detection method is desired [7].

In healthy full term infants, normal adaptions and interactions occur causing variability in blood pressure, temperature, breathing and Heart Rate (HR). However, when an infant is getting sick, the amount of variability is thought to decrease, called de-complexification [8]. Hon and Lee studied the relation between HR and foetal distress already in 1965, showing that during foetal distress alterations in inter-heartbeat interval preceded changes in HR [8]. During the past years this research has been extended to the phase after birth and the possibility of a new diagnostic technique. Some software, e.g. Kubios HRV, is already designed to easily analyse HR data in the time and frequency domain. However, this analysis is only possible in retrospective [9].

The aim of this study is to design a predictive model which captures the relationship between HRV, baseline patient characteristics and clinical measurements (such as CRP and weight loss) on the development of sepsis in premature infants.

Methods

Patients Data was collected from a cohort of prematurely born patients hospitalized on the NICU of the Sophia Children's Hospital (Rotterdam, the Netherlands) between January and June 2016 was. Inclusion criteria encompassed born very preterm and Low Birth Weight (LBW). Very preterm infants are born with a GA ≤32 weeks (including 32 0/7) and LBW infants are born weighing less than 1500 gram. Exclusion criteria encompassed born outside of the Sophia Children's Hospital, diagnosis of Necrotizing Enterocolitis (NEC), major surgical procedures, sepsis developed within 72 hours after birth (early onset sepsis; EONS) and data recordings available of less than 6 hours. NEC is defined as an episode meeting Bell stage 2 or 3, or requiring surgical intervention. Spontaneous focal intestinal perforation in a normal appearing bowel with necrosis is not defined as NEC [10]. Diagnosis of NEC was based on review of the individual patient files by the author. An example of a surgical procedures leading to exclusion is the treatment of gastroschisis where the abdominal content protrudes from the abdomen of the infant. In the case of sepsis, patients with major surgical procedures or diagnosis of NEC within 1 week before the onset of sepsis were excluded since these events can alter HRV [11]. The cohort consists of both healthy and septic patients. Since all data is collected retrospectively, anonymized and as per standard of care during clinical practice, informed consent was not mandatory according to the Dutch Medical Research Involving Human Subjects Act.

Data collection The outcome covariate sepsis is defined as a blood culture-proven LONS with an elevated CRP concentration of greater than 10 mg/L during the admission on the NICU [12]. LONS is defined as manifesting in the first 72 hours after birth. Patient information was collected from the individual Electronic Health Record and the intensive care Patient Data Management System. Date and time of the blood culture and authorisation with outcome of sepsis (binary; positive or negative) were collected. Only the first event of sepsis of each patient was noted.

Vital signs were collected using Infinity Acute Care System M540 & C700 (Dräger, Lübeck, Germany). Data of the entire NICU admission or as much as available was collected. Data collection was only stopped when the patient developed sepsis.

Variables The baseline covariates which are included in the model are gender, GA and birth weight. The GA is group mean centred, since zero age has no value for the prediction. Birth weight is defined as the percentile of the normal growth curve based on reference curves of the '*Stichting Perinatale Registratie Nederland*'. A distinction in the used reference curves is made between primipara, multipara and gender [2]. The time-varying covariates include weight, CRP, bradycardia, median RR and HRV. Weight is defined as percentage weight loss or gain with respect to the infant's birth weight. CRP was divided on a categorical scale: no measurement (0), CRP below 10 (1) and CRP equal or greater than 10 (2). Bradycardia is defined as the amount of heart beats below 100 beats per minute within one hour of data [13] .

Electrocardiogram data, sampled with a frequency 200 Hz, were analysed using Matlab (The Mathworks, Natick, United States). The Pan Tompkins algorithm was used to retrieve the inter beat intervals; RR interval. The algorithm consists of pre-processing of the data through filtering and decision rules for the detection of the RR interval (see section 3.3) [9]. The Standard Deviation of the Average of 5 minutes of RR interval data (SDARR) was used as indicator of HRV. This choice was based on research conducted on the same dataset, resulting in SDARR as the best predictor (see section 4.2). Also the median RR interval over one hour was included.

Statistical analysis The demographic characteristics of the patients were analysed, divided into the non-septic, septic and excluded group. The outcome covariate was defined binary; negative (0) or positive (1). When the outcome of the blood culture was positive, the hour in which it was taken was marked as one and all previous time points as zero. All other covariates were, when necessary, averaged over one hour. The time-axis starting point is defined at the beginning of the measurement and not immediately after birth since there was no data measured yet. In this way each patient has its own continuous time axis consisting of blocks of one hour per data point. The data was used to build a longitudinal database. When no data was available at a certain time point, the column was left empty. A logistic mixed effect model in R (Rstudio, Boston, United Stated of America, version 3.3.2) using Generalized Linear Mixed Effects Regression (package lme4, version 1.1-12) was used to analyse the relation between the covariates and sepsis [14, 15].

Since the selected time period was the maximum duration for which data was available, a larger sample size wasn't possible. The study has an observational and retrospective design. Though to ensure the sample size would be sufficient for adequate power for logistic mixed effect modelling, a golden rule was used: a minimum of 10 patients or events for each variable included in the model. Based on this rule a minimum of 40 patients was needed for the used variables (GA, SDARR, CRP and time) [16]. The output of the logistic mixed effect model consists of the significance of the estimates of the different covariates, overall fit of the model, classification table, correlation of the fixed effect and the probability of developing sepsis [14]. To assess the overall fit of the model the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are used [17]. The probability of sepsis is represented as the normalized odds ratio; prediction of sepsis is empirically defined as odds ≥ 0.4 .

Table 1 Baseline characteristics of the study population; exclusion based on NEC, EONS, surgical procedure or missing data

	Septic (n= 18)	Control (n=42)	Exclusion (n=27)
GA (weeks)	$27(26^{3/7}$-$28)$	$29^{1/7}(27^{4/7}$ -$30^{2/7})$	$30^{6/7}$ $(27^{5/7}$-$31^{5/7})$
Sex (male : female)	14 : 4	20 : 22	11 : 16
Birth weight (g)	873 (790-1100)	1183 (920-1335)	1250 (956-1725)
Mortality (n(%))	5 (28)	2 (5)	5 (19)
Length of stay (days)	36 (23-60)	18 (6-26)	6 (3-34)

*GA, birth weight and length of stay are expressed as median (Inter Quartile Range; IQR)

The model is run different times with different combinations of the covariates, including nonlinear combinations such as splines, to gain the model with the best fit for the outcome. The role of the doctor in detecting sepsis is analysed by comparing a model with and without CRP as a covariate, since CRP is measured based on doctors' indication. Statistical significance is defined as a p-level below 0.05. A database of a cohort patients with birth weight below 2500 gram, averaged over six hours, was analysed previously with no successful results (see section 4.1).

Results

Study population During a time frame of six months, 60 eligible patients were admitted to the NICU, of whom 18 infants experienced at least one episode of sepsis. Figure 1 shows the inclusion and selection of the infants in this study. Baseline characteristics of the septic, control (non-septic) and the excluded patients are presented in table 1. Infants with sepsis had lower birth weight and GA, with higher mortality rates and length of stay. The excluded group stayed in the NICU for a shorter time than the control group.

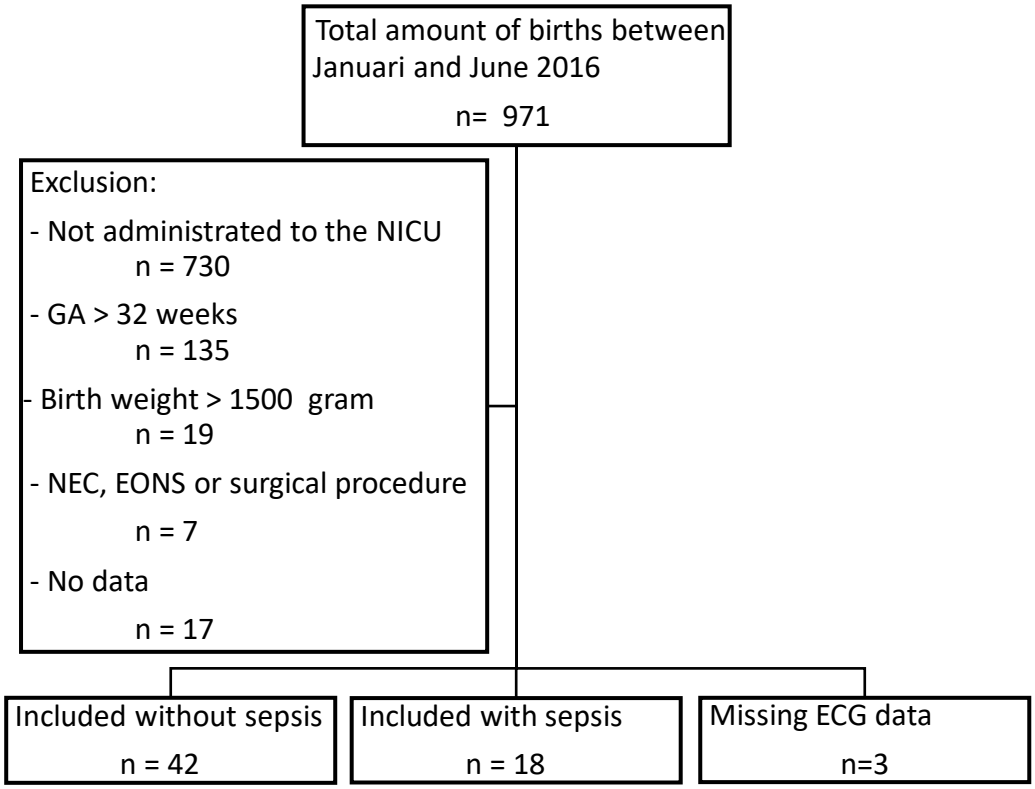

Figure 1 Study inclusion flow chart

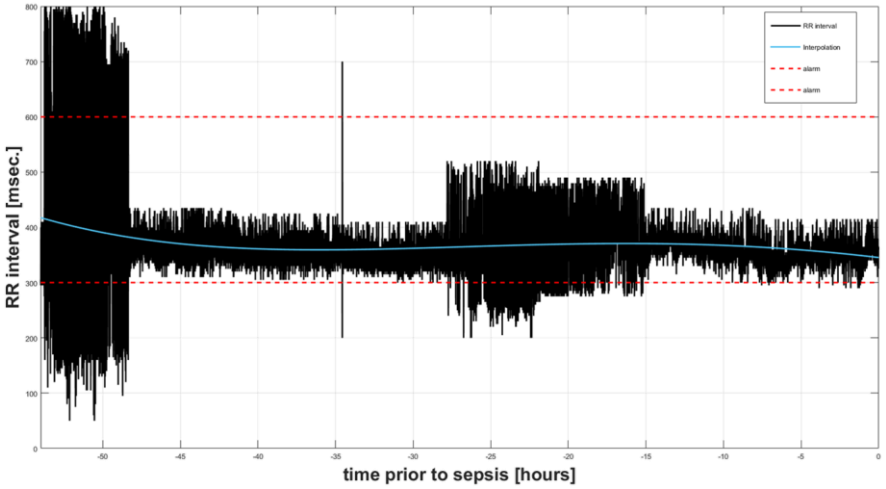

Figure 2 Alterations in the raw RR interval over time preceding sepsis. Raw data of one patient is presented

Heart Rate Variability Figure 2 shows the raw RR interval for one patient developing sepsis with alterations in variability over time. Further analysis of the averaged RR interval and HRV preceding sepsis revealed the presence of multiple dips in HRV around 50, 36 and 15 hours prior to sepsis, figure 3. In the first two cases these dips are matched with a decrease in RR interval, but in the last case an increase in RR is found. This increase of RR can be related to a rise in the amount of bradycardias expected prior to sepsis. In the last hours for sepsis both RR and HRV drop again. These drops are to a lesser extent visible in the graph of non-septic patients.

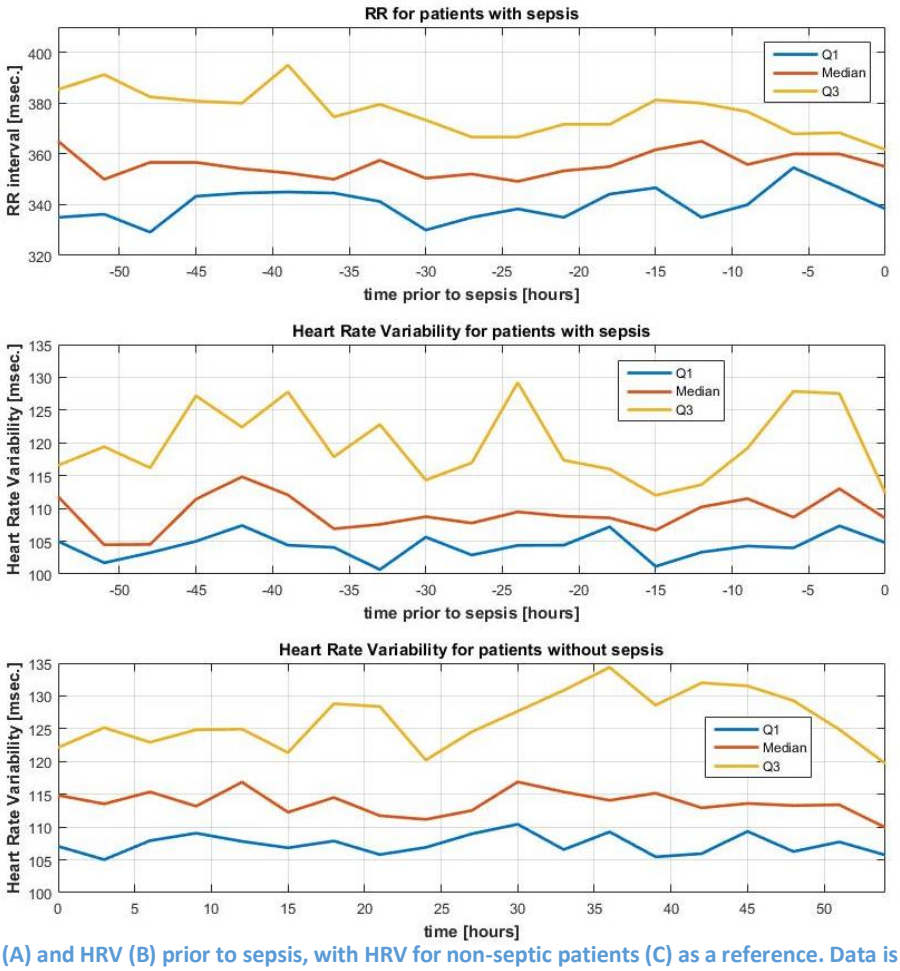

Figure 3 RR interval (A) and HRV (B) prior to sepsis, with HRV for non-septic patients (C) as a reference. Data is averaged over three hours and presented as IQR (Q1:25%, Median, Q3:75%).

Over time the median HRV is in general higher for the non-septic patients in comparison to the patients that develop sepsis.

Selection of covariates After backward elimination of both baseline and time-varying covariates, based on the significance of the estimate ($p<0.05$ as a criterion) and the lowest AIC & BIC value, the following variables remained associated with sepsis: GA, CRP, bradycardia, median RR and HRV. Birth weight, gender and percentage weight loss were discarded. The remaining variables are standardized, using the z-score, to improve the convergence of the optimization. The z-score indicates the amount of standard deviations an observation differs from the mean of variable for the entire cohort [14].

Generalized Logistic Mixed Effect Regression The remaining covariates were entered in a Generalized Logistic Mixed Effect Regression (GLMER) analysis, represented as coefficient estimate (b) and Standard Error (SE). In a GLMER model, model 1, with odds of sepsis as the dependent variable, CRP (b = 6.38, SE = 0.89), bradycardia (b = 0.27, SE = 0.10) and Median RR (b = -3.95, SE = 1.33) were significant (all $p<0.01$). Figure 3 and table 2 show assessments of the quality of the statistical model. The estimates for HRV and GA were not significant, so GA was eliminated from the model. Since HRV is expected to have a significant contribution to the detection of sepsis, the relation between HRV and the odds of sepsis is studied, see figure 4. The histogram of the HRV shows a negative skew, relating to a non-normal distribution. The relation between HRV and the odds of sepsis can never be fitted linearly. For this reason HRV was estimated through a natural cubic spline in the second model. The splines are fitted around the percentiles of the data. To define the spline, a maximum of six breakpoints were added around the following percentiles (corresponding value) of HRV: • 17 (-0.48) • 33 (-0.36) • 50 (-0.26) • 67 (-0.12) • 83 (0.20). In the second GLMER model with odds of sepsis as the dependent variable the following covariates were significant (all $p<0.05$): CRP (b = 6.34, SE = 0.68), bradycardia (b = 0.70, SE = 0.28), Median RR (b = -4.04, SE = 1.5) and Spline4 (b = 191, SE = 45.0), Spline5 (b = -5625, SE = 38.8) and Spline6 (b = -11120, SE = 66,1). The AIC and BIC values for the model are lower and the amount of true positives is higher. All correlations of fixed effect remained below 0.5 except the correlation between Spline5 and Spline6, which was equal to 0.80. A third model is analysed to evaluate the value of CRP. The removal of CRP from model 2 and the addition of GA results in model 3. Both assessments of the quality of the model are worse when CRP is removed. Based on the assessments of the overall fit of the model and the highest sensitivity, model two was analysed further. The predicted odds ratios, per patient who developed sepsis, are presented over time in figure 5. As a reference, figure 7 presents the odds ratios of the control group. The false positive measurement is visible in this graph.

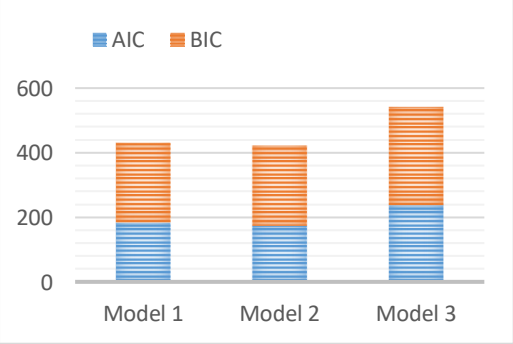

Figure 4 AIC and BIC values

Table 2 Classification table of the three tested models

	Model 1	Model 2	Model 3
True positive	4	7	2
False positive	0	1	0
True negative	42	41	42
False negative	14	11	16
Sensitivity [%]	0.22	0.39	0.11
Specificity [%]	1	0.97	1

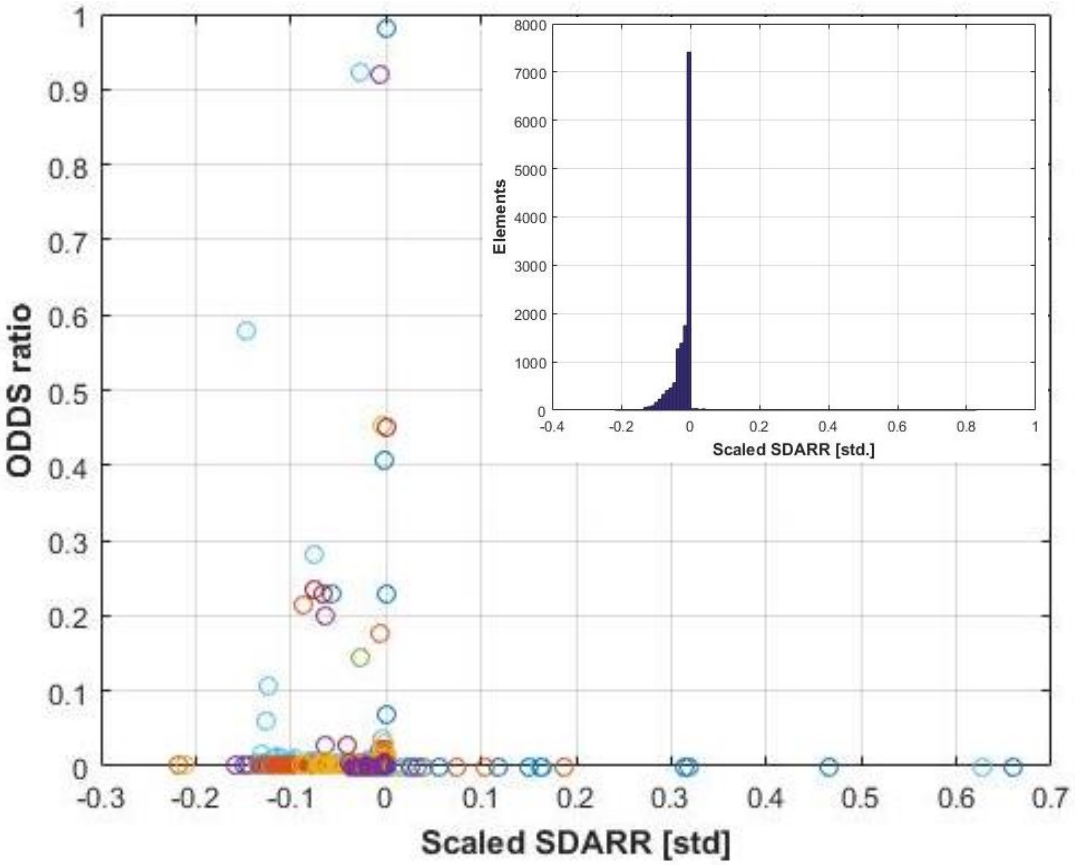

Figure 6 The relation between the scaled HRV parameter and the odds of developing sepsis, with the histogram of the HRV in the upper right corner.

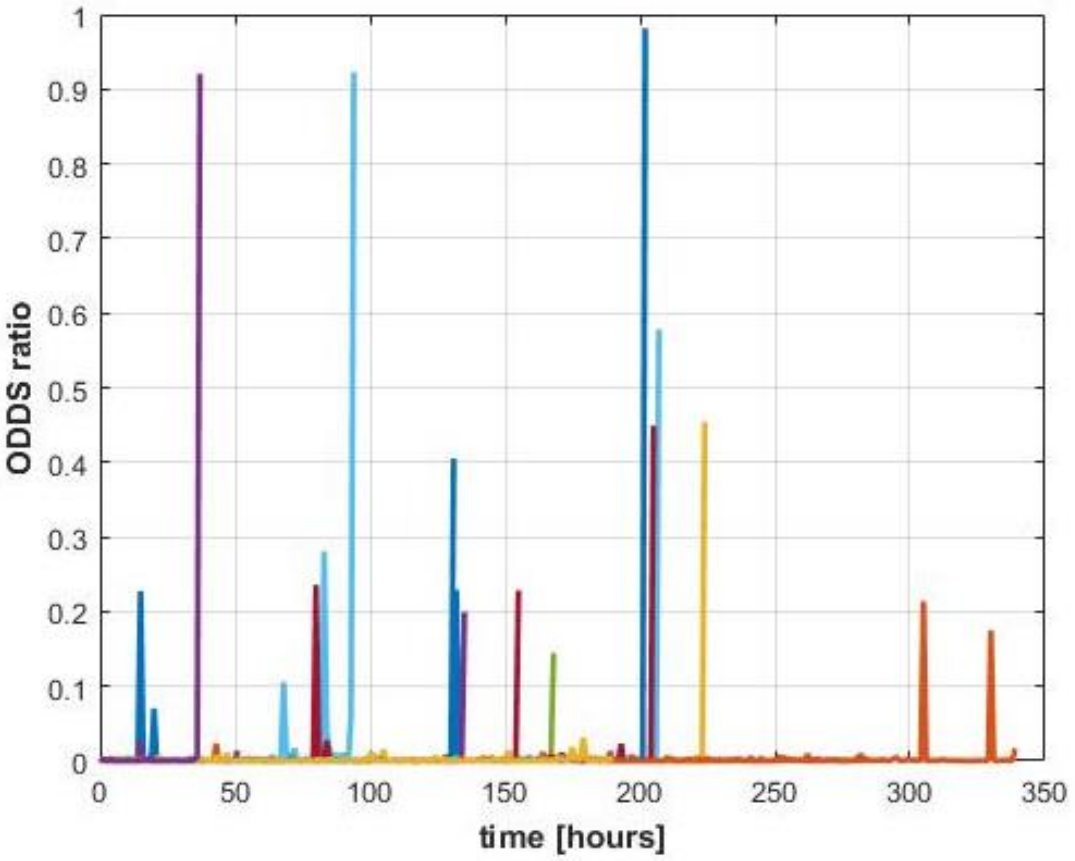

Figure 5 The odds ratio of developing sepsis over measurement time. Data of the entire septic group (n=18) is displayed, each line represents the data of different patient.

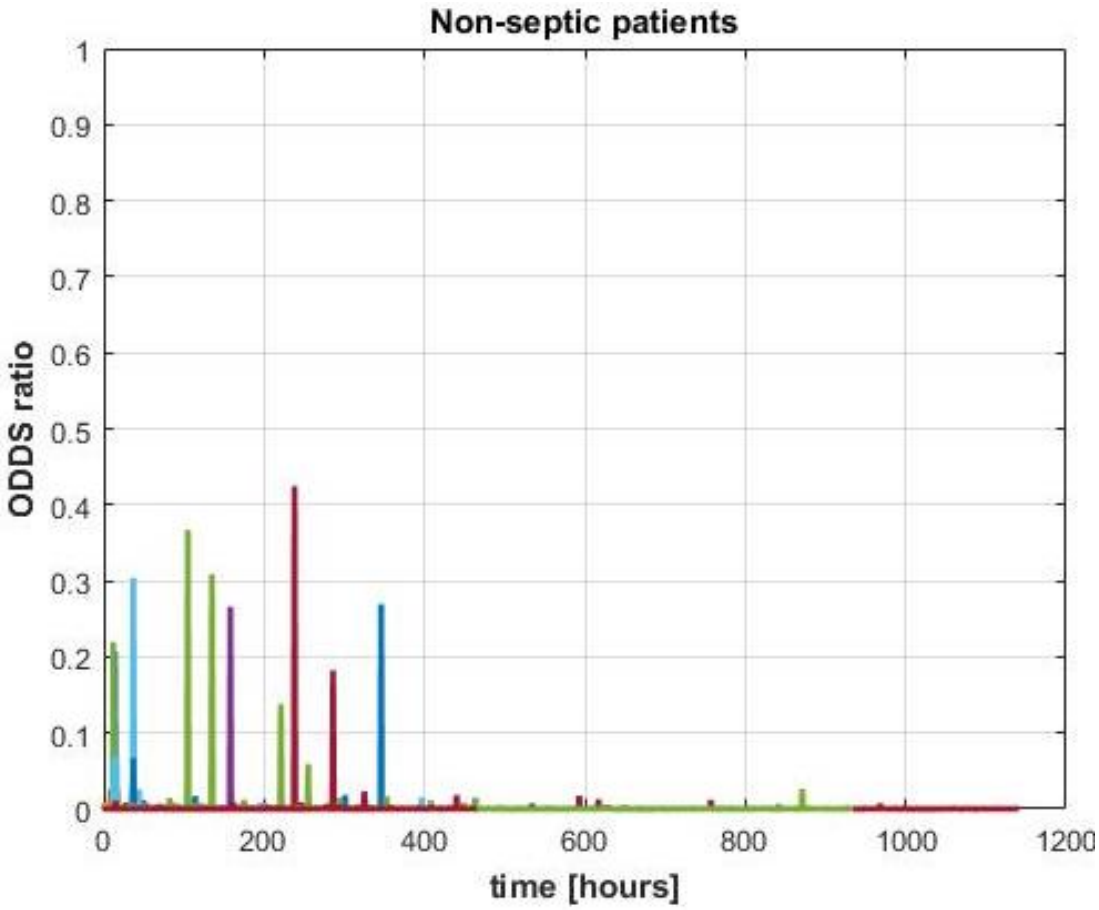

Figure 7 The odds ratio of developing sepsis over measurement time for the control group (n=42).

In 7 out of 18 cases the odds ratio becomes greater than 0.4; the model predicts an increased chance on developing sepsis. When the point in time that the odds became greater than 0.4 is compared to the time the blood culture is taken, no difference in time is found for all 7 cases. So with respect to the blood culture, no predictive value over the time is gained. However analysis of the median (IQR) of the authorisation of the blood samples showed that this process takes another 90,22 (74,63 – 108,60) hours to confirm the outcome of sepsis.

Discussion

The aim of this study is to design a predictive model which captures the relationship between HRV, baseline patient characteristics and clinical measurements on the development of sepsis in premature infants. The results of this study show that the odds of developing sepsis increase when CRP increases, the amount of bradycardias increases, the median RR decreases and (depending on the range of HRV) the spline of HRV decreases. The estimates for the baseline patient characteristics gender, GA and birth weight did not significantly differ from zero. A difference was discerned between the septic and control patients, based on the average values of these characteristics. However, this difference was of lower statistical value to the predictive model with respect to the time-varying covariates. The results of the model without CRP only showed two true positive predictions, proving that the role of the doctor remains important in predicting sepsis.

The influence of percentage weight loss or gain was negligible. This fact can be explained since when an infant gets ill, the number of measurements of weight drops because they're too fragile to be measured. The weight of a patient is furthermore highly influenced by treatment.

When the infant loses weight, the number of feedings or the amount of food is raised to increase the body's weight. On the other hand when the infant gains weight as a consequence of fluid retention, treatment is started to get rid of the additional liquid. The same rule applies to temperature measurements: if the temperature of the infant drops, the incubator will be adjusted accordingly to increase the body's temperature and vice versa. For this reason, body temperature is not included in this model [4].

A limitation of this study is the rate of missing data points. If one of the covariates is missing, the statistical model will ignore the entire time point. In this way important predictive information can be lost. Another limitation is the sampling frequency of the ECG. Only a sampling frequency ranging from 250 to 500 Hz can ensure accurate R peak detection [8]. Another important point is the threshold of the odds ratio of 0.4. In theory, when the chances are greater than 50%, the probability becomes great enough to predict the outcome with enough certainty. However, for the threshold of 0.4 only one outcome was false positive.

The odds ratio surpassed the threshold value for seven patients amongst whom sepsis was detected. This resulted in a low sensitivity score since the amount of false negatives was higher than the amount of true positives. To be able to implement a predictive model in practice, the model needs to be highly sensitive. As such this model needs to be improved and tested further before it may be implemented. A high amount of false positives would be less severe because in practice this would result in the draw of an additional blood sample, while a high amount of false negatives results in the missing of sepsis with all possible consequences. To improve the sensitivity of the model the addition of more relevant parameters could be a solution. Capillary refill time greater than two seconds is found to be a strong predictor in the 24 hours preceding the onset of sepsis [18]. Capillary refill time is the time it takes for colour to return to the capillary bed after pressure is applied to the skin [4]. During the development of sepsis, the microcirculation changes and the severity of these changes based on the capillary refill time can be an indicator of sepsis [18].

During the testing of different models, backward elimination was used to find the best fit. This process was needed since adding to many parameters to the model can result in convergence problems. Regarding these problems, it might be better to take a closer look at the covariates entered to the model instead of adding additional parameters. A baseline difference was found between the septic and control group already four days preceding sepsis. Correcting for this difference can result in a better prediction of sepsis.

Finally, the development of sepsis can take a few hours or take up to two days [7]. Defining sepsis at the moment a positive blood culture is taken, does not capture the exact moment of the onset of sepsis. Moorman et al. (2006) for example used their predictive model to define sepsis 18 hours prior to the moment of the positive blood culture. Another factor is the cases in which sepsis is missed in this data cohort, referring to clinical sepsis; sepsis without a positive blood culture [7].

Nevertheless, every true positive detection of sepsis is beneficial to the health of infant through the early start of treatment.

Conclusion A great potential is found in the predictive value of both HRV and clinical measurements on the odds of developing sepsis. However, the current model is not sensitive enough to predict all cases of sepsis. Benefits lay in the early prediction of sepsis to bring forward the draw of the blood sample in time to prove sepsis, since the authorisation of the outcome takes about four days.

References

1. *World Health Organization; Born too soon: the global action report on preterm birth*. 2012.
2. Nederland, S.P.R., *Perinatale Registratie Nederland Grote Lijnen 1999 - 2012.* Utrecht, 2013.
3. Oza, S., et al., *Neonatal cause-of-death estimates for the early and late neonatal periods for 194 countries: 2000–2013.* Bulletin of the World Health Organization, 2015. **93**(1): p. 19-28.
4. Bekhof, J., et al., *Clinical signs to identify late-onset sepsis in preterm infants.* European journal of pediatrics, 2013. **172**(4): p. 501-508.
5. Kornelisse, R., *Bacteriële infecties op de NICU van het Erasmus MC.* Kwaliteits Informatie Systeem Sophia Kinderziekenhuis, 2010.
6. Standage, S.W. and H.R. Wong, *Biomarkers for pediatric sepsis and septic shock.* 2011.
7. Moorman, J.R., D.E. Lake, and M.P. Griffin, *Heart rate characteristics monitoring for neonatal sepsis.* IEEE transactions on biomedical engineering, 2006. **53**(1): p. 126-132.
8. Cardiology, T.F.o.t.E.S.o., *Heart rate variability standards of measurement, physiological interpretation, and clinical use.* Eur Heart J, 1996. **17**: p. 354-381.
9. Tarvainen, M.P., et al., *Kubios HRV–heart rate variability analysis software.* Computer methods and programs in biomedicine, 2014. **113**(1): p. 210-220.
10. Samuels, N., et al., *Necrotising enterocolitis and mortality in preterm infants after introduction of probiotics: a quasi-experimental study.* Scientific Reports, 2016. **6**.
11. Sullivan, B.A., et al., *Infection and other clinical correlates of abnormal heart rate characteristics in preterm infants.* The Journal of pediatrics, 2014. **164**(4): p. 775-780.
12. Stoll, B.J., et al., *Late-onset sepsis in very low birth weight neonates: the experience of the NICHD Neonatal Research Network.* Pediatrics, 2002. **110**(2): p. 285-291.
13. Alferink, K., *Monitor bewaking met betrekking tot de hartactie bij kinderen (Versie 3).* Kwaliteits Informatie Systeem Sophia Kinderziekenhuis, 2015.
14. Faraway, J.J., *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. Vol. 124. 2016: CRC press.
15. Carey, V. and Y.-G. Wang, *Mixed-effects models in S and S-PLUS*. 2001, Taylor & Francis.
16. Lenth, R.V., *Post hoc power: tables and commentary.* Iowa City: Department of Statistics and Actuarial Science, University of Iowa, 2007.
17. Dobson, A.J. and A. Barnett, *An introduction to generalized linear models*. 2008: CRC press.
18. Bohanon, F.J., et al., *Heart rate variability analysis is more sensitive at identifying neonatal sepsis than conventional vital signs.* The American Journal of Surgery, 2015. **210**(4): p. 661-667.

2. Heart Rate Variability

In literature many options to analyse HRV are found, the following section will discuss these possibilities.

HR decelerations, drops in HR of at least 15 bpm and 2 minutes duration, are present in data of neonatal patients. Besides decelerations, drops in HRV have been found and shown to relate to the development of sepsis in neonatal patients [18]. Besides the relationship to sepsis, alterations in HRV are also related to smoking, drugs and alcohol in adults. Even in patients suffering from diabetes, decreased HRV is proven to be related to neuropathy. These alterations are found before clinical symptoms of neuropathy were present. Another important relationship concerning HRV is the one with gender and age. Healthy new-born boys have a decreased HRV compared to girls. Also over time HRV changes, which could be present in time and frequency indexes of HRV. It goes so far that indexes measured during the day are significant different from the ones measured during the night [19]. So it's very important to take these facts into account and chose the correct related HRV index.

HRV can be quantitated in different manners. These manners can be divided into time -, frequency domain and non-linear options. The methods to determine HRV are based on the determination of the RR interval from the QRS complex in ECG data, see figure 8. The variation in the RR interval is defined as the HRV. The detection of the R peak will be discussed in section 3.3.

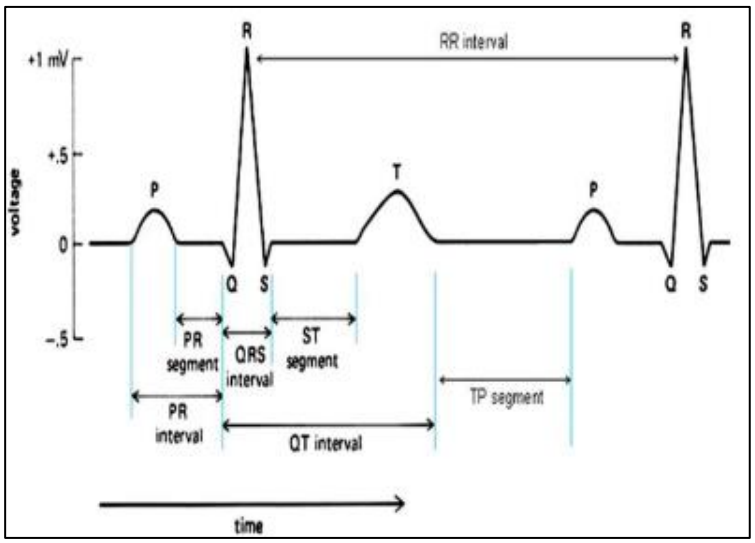

Figure 8 Typical ECG characteristics with the QRS complex and RR interval visualised [20]

1. Time domain

Table 3 gives an overview of possible measurements of HRV in time domain.

Table 3 Time domain measures of HRV [8, 9]; milliseconds (ms)

Variable	Unit	Explanation	Formula
$\overline{RR}$	ms	The mean of all RR intervals	$\overline{RR} = \frac{1}{n}\sum_{i=1}^{n}(t_n - t_{n-1})_i$ (1)
SDRR	ms	Standard deviation of all RR intervals	$SD_{RR} = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}\left(RR_j - \overline{RR}\right)^2}$ (2)
SDSD	ms	Standard deviation of successive RR interval differences	$SDSD = \sqrt{E\{\Delta RR_j^2\} - E\{\Delta RR_j\}^2}$ (3)
$\overline{HR}$	1/min	Mean of all HR data	See (1)
SDHR	1/min	Standard deviation of HR	See (3)
RMSSD	ms		$RMSSD\sqrt{\frac{1}{N-1}\sum_{j=1}^{N-1}\left(RR_{j+1} - RR_j\right)^2}$ (4)
RR_{50}	-	The number of successive intervals differing more than 50 ms	$RR_{50} = \sum_{j=1}^{N-1}\left(RR_{j+1} - RR_j\right) > 50$ (5)
pRR_{50}	%	Relative NN50 to total amount of NN intervals	$pRR_{50} = \frac{RR50}{N-1} x\ 100\ \%$ (6)
HRV triangular index	-	Integral of the RR histogram divided by the height of the histogram	$HRV_{index} = \frac{N}{Y}$ (7)
TIRR	ms	Triangular interpolation of RR width of a triangle fitted to the histogram of RR intervals	$TIRR = M - N$ (8)
SDARR	ms	Standard deviation of the average of RR intervals per 5 min of data	See (2)
p10, p20 etc.	%	Percentile	$p10 = \mu - Z * \sigma$ (9)
Skewness		Measure of symmetry of the data	$skewness = \sum_{i=1}^{N}\frac{(RR_i - \overline{RR})^3/N}{\sigma^3}$ (10)
Kurtosis		Measure of the tail of the plotted data compared to the normal distribution	$kurtosis = \sum_{i=1}^{N}\frac{(RR_i - \overline{RR})^4/N}{\sigma^4} - 3$ (11)

Time domain methods have a few benefits above others: they are relative simple to compute since they don't require any alterations the successive RR intervals before calculation. The obvious and most easy choice to calculate is the mean of the RR intervals or of the HR itself over a certain amount of time [9]. However, of course the mean of RR intervals can easily be the same for different sets of data and will not give a specific insight in the relation between the data and possibly other correlates. Also for asymmetrical distributions the mean doesn't add a lot of information and the median could be a better solution. The skewness and kurtosis, though calculated around the mean, of the RR interval could add additional information about the asymmetrical part of a distribution of RR intervals [21]. These measures also remains reliable when dealing with missing heart beats [22].

The standard deviation (SDRR) of the RR intervals is a measurement of the overall variability within RR intervals, both short- and long term. SDRR is not very sensitive when there are a lot of HR decelerations in the signal [22, 23].

Variation is equal to the total amount of power using spectral analysis, so it represents all the cyclic components responsible for variability in the period of recording. It's not useful when comparing different recording times since the HRV variance is dependent on time (day/night).

A variation on the normal standard deviation is the SDARR. This is the standard deviation of the average RR interval over usually 5 minutes. This is an estimate of the long term components of heart rate variability and requires minimum 24 hours measurements [9].The SDSD is another variation on the standard deviation, since it's the deviation of the difference between successive RR intervals, in literature this value is also called SDNNi.

The Root Mean Square of the Successive Difference (RMSSD) is a well know method and is a measurement of short term variability. The amount of successive RR intervals differing more than 50 ms from each other is called the RR50, pRR50 is the percentage in comparison to the total amount of RR intervals. The RMSSD and (p) RR50 both estimate short-term variation in high frequency variations of heart rate and so are highly correlated with each other [8].

The triangular index is calculated from the histogram and dependent on the bin width of the histogram. A recommended value for bin with is 1/128 sec [9]. The triangular interpolation of the RR interval (TIRR) is the width when y = 0 of the minimum square difference of the highest peak of the histogram of all RR intervals, see figure 9. For the calculation of these parameters data of at least 24 hours is needed to ensure the value of the parameters.

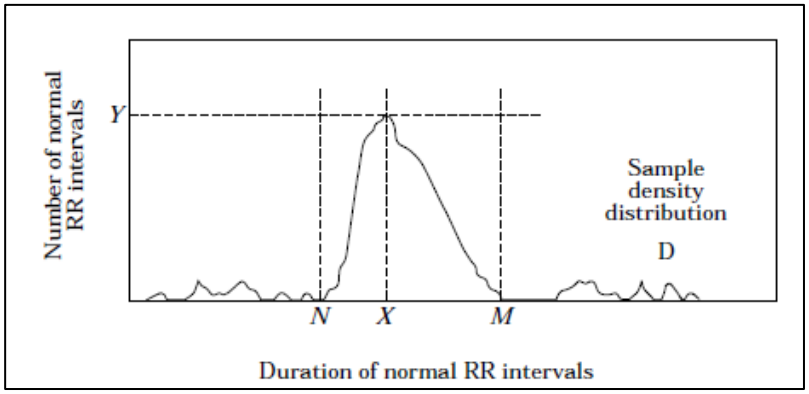

Figure 9 Explanation of formula (8) [8]

Both skewness and kurtosis are also measures of the histogram of the RR intervals. The skewness and kurtosis are calculated only around the mean, so are not useful for asymmetrical distributions [21].

To measure the percentile of the distributions data needs to be ordered, in this way it contains no information anymore about the order of RR intervals and individual decelerations [24].

Recommendations

Though the mean itself might give insufficient information, it's important to monitor the trend of HR during the measurement. Since this can have an impact on the parameters measuring the overall variability like SDRR and TIRR. Logically most time domain measures high correlate with each other, so it doesn't make sense to measures them all.

To give an estimate of all kind of variations the following is recommended [8]:

- SDRR Measure of overall variability
- SDARR Measure of long-term components of variability
- RMSSD Measure of short-term components of variability

2. Frequency domain

Table 4 shows the frequency domain measures of HRV.

Table 4 Frequency domain measures of HRV [8, 9]

Variable	Unit	Explanation	Formula
f_{peak}	Hz	Peak frequencies for all frequency bands	$f_{peak} = f(\max(P))$ (1)
P_{band}	ms^2	Absolute powers of all frequency bands	$P_{band} = \int_{f_{lower}}^{f_{upper}} S(f)df$ (2)
P_{rel}	%	Relative power of all frequency bands	$[HF, LF, VLF] = \frac{[HF,LF,VLF]}{P_{tot}} * 100\%$ (3)
LF/HF		Ratio between two frequency bands	$Ratio = \frac{LF}{HF}$ (4)
P_{norm}	n.u.	Power of frequency bands normalized for the VLF band	$P_{norm} = \frac{HF}{(P_{tot} - VLF)}$ (5.1) $P_{norm} = \frac{LF}{(P_{tot} - VLF)}$ (5.2)
P_{tot}	ms^2	Total spectral power	$P_{tot} = \int_{f_{lower}}^{f_{upper}} S(f)df$ (6)
f_{resp}	Hz	ECG derived respiratory frequency	

For the measures of HRV, the data first needs to be transformed to the frequency domain. An estimate of the power distribution over frequencies is made. The standard method is the Fourier transform. Fourier transform assumes a signal to be the sum of periodic components, also periodicity, with equal behaviour (mean and variance) over time, also stationarity. Due to the non-stationarity of RR data, normal Fourier transform can't be used [7]. Cubic spline interpolation is used to transform the signal into an equidistantly sampled series by fitting a polynomial function [25]. Note, this also works as a low pass filter. To estimate the spectrum the following options can be used:

- Welch periodogram

First the data is divided into segments, which have some overlap. The segments are then windowed. The final spectrum is the average of the Fast Fourier Transform spectra of the windowed segments.

- Autoregressive modelling

Spectrum estimation by estimating parameters of a autoregressive model using spectral factorization. There is no guarantee that the model order is correct.

- Lomb periodogram

Least-squares fitting of the data to a sinusoid; minimize the error between sinusoid and data points [26].

The sampling frequency should be twice the maximum signal frequency, if not aliasing occurs (power of components greater than the Nyquist is added to frequencies below the Nyquist). Nyquist is half the sampling rate, around this frequency folding occurs.

Peak frequencies (f_{peak}) in the spectral estimation of the RR intervals occur within set frequency band ranges:

- Very Low Frequencies	(VLF)	0	–	0.04 Hz
- Low Frequencies	(LF)	0.04	–	0.15 Hz
- High Frequencies	(HF)	0.15	–	0.4 Hz

For short measurements, t<5 min, the VLF component isn't useful.
The HF result from parasympathetic activity and the LF from a combination of parasympathetic and sympathetic activity. That is why the ratio LF/HF is of interest, since this gives an indication of the sympathetic activity. The parasympathetic activity results in a resting state of the body, slows the HR. While the sympathetic activity becomes active when the body is stimulated and increases the HR. The power at these bands can also normalized or taken relative to the total amount of power or the signal. The total power and the power in the specific frequency bands are calculated by integrating the spectrum over the entire spectrum or over the specific bands. Since the total power (Ptot) is equal to the variance of the signal, reduced power at certain frequency bands is also related to reduced variance of the signal [26].

Peaks in the data, for example due to motion artefacts in the ECG, increase power easily at higher frequencies [9]. Another factor having influence in the HF band are the respiratory rate and the neonatal HR. So for frequency analysis the respiratory frequency (f_{resp}) needs to be taken into account, this varies between subject and is dependent on physiological status of the subject.
This phenomena is called Respiratory Sinus Arrhythmia; the natural variation in HR due to respiration and is the HF component. To give an indication, the neonatal HR should be between 120 and 160 beats/min. This equals 2 – 2.67 Hz. The respiratory rate should be between 40 and 80 breaths/min, this equals 0.67 – 1.33 Hz [26].

Recommendations
The HF are mostly within the range of interest and linked to the respiratory variability. However in literature no consensus is found on a single optimal method. High correlations are found between SDRR,pnn50, RMSSD and spectral analysis. This raises the question if frequency domain analysis adds useful information [27].

3. Non-linear

Table 5 shows the nonlinear measures of HRV.

Table 5 Nonlinear measures of HRV [8, 9]

Variable	Unit	Explanation	Formula
SD_1 ,SD_2	ms	Standard deviations of the Poincaré plot	-
SampEn	-	Sample entropy	$SampEn = -\log(\frac{A}{B})$ (1)
ApEn	-	Approximate entropy	$ApEn = \phi^m(r) - \phi^{m+1}(r)$ *(2)
REC	%	Percentage of recurrence points	$REC = \frac{1}{N^2}\sum_{i,j=1}^{N} R_{i,j}$ (3)
DET	%	Determinism	$DET = \frac{\sum_{l=l_{min}}^{N} lP(l)}{\sum_{i,j}^{N} R_{i,j}}$ **(4)
L mean	beats	Average line length of diagonal lines in recurrence plot	$L\ mean \frac{\sum_{l=l_{min}}^{N} lP(l)}{\sum_{l=l_{min}}^{N} P(l)}$ (5)
L max	beats	Maximum line length of diagonal lines in recurrence plot	$L\max = \max(L\ mean)$ (6)

* $\phi^m(r) = (N-m+1)^{-1}\sum_{i=1}^{N-m+1}\log(C_i^m(r))$

** P(l) = histogram of the lengths (l) of the diagonal lines

A Poincaré plot is shown in figure 10, it represents the correlation between successive RR intervals. The RR interval on y-axis is equal to RR_{j+1} and it is plotted as a function of R_j. The data plotted in the graph clearly has the shape of an ellipse and is characterized by its width and length. The first is also called SD1 and is the standard deviation of the data points perpendicular to the line $RR_j = RR_{j+1}$ (line of identity). The latter is called the SD2 and is the standard deviation along the line of identity. Short term variability is captured by SD1 and long term variability by SD2 [9].

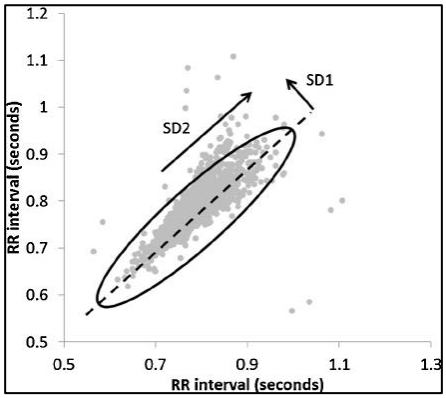

Figure 10 Poincaré plot [28]

The principle of sample entropy is based on approximate entropy, both are measures of complexity. Sample entropy describes regularity and detects time series with spikes (heart rate decelerations in the case of sepsis) [22]. The sample entropy drops in series with a lot of spikes, spikes increase variance. The calculation of sample entropy is based on the embedding dimension m, default is 2, and the tolerance r, default is 0.2 SDRR. In formula 1, A represents multiple vector pairs having length m+1 and B having length m. When the mean and variance of the data are the same, sample entropy describes the unpredictability of fluctuations over the data [29]. Sample entropy has a computational advantage over approximate entropy and is mostly used.

The calculation of the recurrence plot is also based on m and r. The default values are no m=10 and $r = \sqrt{m}SDRR$. When two embedding vectors are close to each other, the value in the binary square matrix gets the value 1, if not the value will be 0. In this way short lines are formed in the matrix, which are parallel to the diagonal. This results in the parameters recurrence point and average or maximum line length. Finally determinism is the percentage of recurrence points which form diagonal lines in the recurrence plot.

Recommendations
The Poincaré plot seems a solid method to both analyse the short and long term variability without being dependent choosing the right values for the embedding dimension or tolerance, which are needed when calculating entropy and the recurrence plot.

4. Literature study

Finally when all the different measures of HRV are discussed it's important to look into literature and to give an overview of the current insight into HRV and sepsis in premature infants. That's why a small literature study is performed using the following search terms: 'heart rate variability & premature infants & sepsis', >2000. Table 6 gives an overview of the results for 8 analysed studies, note not all found references are presented. Many studies are found the HeRO score. The HeRO score is a HRV index developed to predict the chance on developing sepsis within the next 24 hours. The values range between 1 and 5, with 5 having the greatest chance on developing sepsis [30].

Conclusion
The cohort sizes of the presented studies in table 6 were almost all quite big. All patients were hospitalized on a NICU. All different sort of results are found, the most important points will be discussed. Multiple studies confirmed the benefits of monitoring HeRO at the bedside to predict sepsis and even reduce mortality. Although the results of one study did not confirm these believe since they stated HeRO had limited ability in detecting a bloodstream infection [31].
Concerning the decision between the use of time or frequency domain measurements, the first was found to be more sensitive than the latter due to the non-stationarity of HR. Only LF proves to indicate between septic and non-septic patients [18]. Although not shown in table 6 also positive results about non-linear methods are found: the Poincare plot is different for ill in comparison to healthy patients due to acceleration and decelerations of HR [32]. Also sample entropy drops in patients whom develop sepsis, but this drop is highly dependent on the presence of spikes. Since spikes are present in the RR data, this isn't a useful method [29]. Again due to the influence of spikes, the mean and standard deviation of HR and RR are predicted not to differ between septic and non-septic patients [22].

In literature a great potential is found for the relation between, and the predictive value of HRV and sepsis. On the other hand, no conclusions can be drawn since the results are so different. Except for the one that vital signs alone are not specific enough to predict sepsis and can benefit from HRV calculations [4].

Table 6 Literature study

Researcher	Year	Study population	HRV measure, $p < 0.05$	HRV measure $p > 0.05$	Statistics	Conclusion
Bekhof et al. [4]	2012	142 GA <34 weeks	Respiratory support, capillary refill & grey skin	Hypo-/ hyperthermia& apnoea	Logistic regression	Not all vital signs are specific enough to identify clinical LONS
Bohanon et al. [18]	2015	10 LBW	SDRR, pNN50, SD2, SD1/SD2 and HR & SpO2*	VLF, LF, HF, LF/HF and cardiac output & mean arterial pressure	Unpaired student t-test	HRV more sensitive than vital signs to indicate sepsis, increased HR & decreased SpO2 for septic patients, spectral analysis not significant
Cao et al. [24]	2004	89 infants NICU	Empirical Cumulative Distribution Function	SampEn, birthweight	Kolmogorov-Smirnov test and ROC curve**	HR is not stationary and becomes more not stationary before developing sepsis
Chang et al. [26]	2001	84 infants NICU	LF, p(10,25,75,90), skewness	HF	Fuller statistics	Spectral power decreased at all frequencies prior to the diagnosis sepsis. LF adds info to BW, GA and days in hospital data
Coggins et al. [31]	2015	2384 infants	HeRO score, 37% of septic infants HeRo≥2	-	-	HeRO limited ability to detect bloodstream infection
Fairchild et al. [33]	2013	2989 LBW	HeRO score septic group	HeRO score non-septic group	-	HRC monitoring resulted in lower septic mortality
Griffin & Moorman [23]	2001	69 infants NICU	RMSSD, Skewness, p50	Mean, SDRR	Mann-Whitney or ANOVA, ROC curve*, Multivariate Regression Analysis	Skewness and p(10,25,50,75,90) of normalized RR distinguish non- and septic patients
Lake et al. [29]	2002	89 infants NICU	SampEn		Wald chi square test, ROC curve*	SampEn drops in the presence of spikes instead of increased regularity, but deals well with missing data

* Peripheral Oxygen saturation (SpO2)

** Receiver Operating Curve = sensitivity (true positive) against 1-specificity (false positive = 1- specificity), area under the curve is accuracy

3. Data analysis

The following section will focus on the choices and steps made during the data collection and analysis required for this study. First, the selection of the cohort and the choice of data of these subjects is explained. Consequently, the solid method for detecting RR interval from raw ECG is elaborated upon. The possible options for HRV have already been discussed in the previous section and in section 4.2 the final selection of the HRV parameter will be explained. Finally, the selection of the covariates and the steps made during the building of the predictive model are elaborated upon.

Content

1. Cohort selection
2. Data collection
3. ECG
4. Covariates
5. Statistical analysis

1. Cohort selection

The selection of inclusion and exclusion criteria is based on literature about comparable studies [4, 11, 12]. The list of criteria was then evaluated with experts whom had relevant clinical experience. This resulted in the following list of inclusion criteria for the cohort selection:
- patients hospitalized on the NICU of the Sophia Children's Hospital (Rotterdam, the Netherlands) between January and June 2016. The collection of data was started in January 2016, so a longer time period wasn't possible.
- very preterm

Very preterm is defined as being born with a GA ≤ 32 weeks, including 32 0/7. This range is defined based on the incidence of sepsis. The incidence is namely inversely related to GA, with rates of 20% at 28 weeks GA and 58% at 22 weeks GA [12].
- LBW

LBW infants are born weighing less than 2500 gram.
Very preterm and LBW infants are prone to developing sepsis because of their compromised immunity, long hospital stay, use of indwelling catheters, tubes and if needed surgeries [34]. Also a sub cohort of infants weighing less than 1500 gram at birth was made; defined as very low birth weight. The sub cohort is made because the incidence of sepsis is even greater for this cohort. Studies from the National Institute of Child Health and Human Development report an incidence of 21% among very low birth weight infants [12, 35].
Besides the inclusion criteria another list of exclusion criteria were defined:
- born outside of the Sophia Children's Hospital.

The reason for the former criteria is purely practical; not all of the necessary data was retrieved at the Sophia Children's Hospital. Since not all data is communicated between hospitals, including patients from other hospital will result in missing data points.
- data recordings available of less than 6 hours.

Despite of being born in the Sophia Children's Hospital, data collection is no guarantee and not always necessary.
If data recordings were too short, defined at a minimum of 6 hours, not enough useful information could be gained.
- surgical procedures
Surgical procedures have an enormous impact on the body. For the measurement of ECG especially anaesthetic agents are causers of alteration since anaesthetic agent effect the heart rate. Depending on the administrated agent, the heart rate increases or decreases or can even alter the peaks seen in the ECG. Patients recovering from general anaesthesia also show altered patterns of HRV [36].
- EONS
Sepsis is defined as a blood culture-proven late-onset sepsis (LONS) with an elevated CRP concentration of greater than 10 mg/L during the admission on the NICU [12]. A distinction is made between early and late onset sepsis. LONS is defined as sepsis occurring after 72 hours of life (on the NICU) or for term infants after 7 days of life until 120 days of life. Patients suffering from early onset sepsis are excluded since this infection type is mostly caused by the mother and given onto the child during labour.
- diagnosis of NEC
NEC is defined as an episode meeting Bell stage 2 or 3, or requiring surgical intervention. Spontaneous focal intestinal perforation in a normal appearing bowel with necrosis is not defined as NEC [10]. Diagnosis of NEC was based on reviewing of the individual patient files by the author. Statements about the diagnosis are made in these files by neonatologist. In case of sepsis, only major surgical procedures or the diagnosis of NEC within 1 week before the onset of sepsis were excluded. Since septic patients are in most cases also the ones developing other diseases, excluding all cases would result in barely any patients left for inclusion [35].
The final cohort of patients consists of non – and septic patients.

2. Data collection

For this retrospective cohort study data from patients hospitalized on the NICU of the Sophia Children's Hospital (Rotterdam, the Netherlands) between January and June 2016 was collected. All measurements were standard of care during clinical practise so informed consent was not mandatory according to the Dutch Medical Research Involving Human Subjects Act. For each patient who met the inclusion criteria data of the entire NICU admission or as much as available was collected. When a patient was administrated at the end of the study period, data collection was continued until the moment of discharge. Data collection was only stopped when the patient developed sepsis, since this study is interested in the prediction of sepsis and not in the alterations during sepsis. Baseline patient information was collected from the individual Electronic Health Record and the intensive care Patient Data Management System. The following information was retrieved:
- Patient Information Dossier (PID) number
- Gender
- Birth date and if applicable mortality date
- Gestational Age; duration of the pregnancy measured from the first day of the last normal menstrual period.
- Start and end date of hospitalization
- Patient weight; also including birth weight.
- CRP

CRP is a biomarker measured through a blood test. During infection CRP levels are increased and so can be used as an indicator of infection. However CRP alone is not specific of sepsis, since levels can also rise due to bacterial and non-infectious inflammatory conditions [6]. In the Sophia Children's Hospital antibiotic treatment is started among other things based on an elevated level of CRP, greater than 10 mg/L [5].

All patients dossiers were scanned for the incidence of sepsis, NEC and other confounders such as major surgical procedures. ECG data were collected using Infinity Acute Care System M540 & C700 (Dräger, Lübeck, Germany). Within the study period all patient data from the Dräger monitor is collected on the NICU server. This data isn't anonymized, so the first step was to assign a study number to each patient name or PID number and delete all names and PID numbers in the data logging.
Data from the patient monitor has the .cpz file extension, so after the collection a transformation into a useful file format was needed. The files have a LoseThos C+ Source extension and can't be opened by normal data analysis software. A software engineer of the Erasmus Mc, BSc M.J.B. van Ettinger, has designed a specific export application for this problem; the cap processor. This tool allows the conversion of .cpz into .csv and other like .mat. Special attention needs to be payed to cases when a new patients is administrated to the same bed side. This results in multiple administrations to one data recording and can cause errors in the data analysis.
The output folder consists of multiple files:
- Waveforms

Table 7 Waveform output

Parameter	Unit	Sample frequency [Hz]
ECG	mV	200
Respiration	Ohm	50
Plethysmograph	1/min	100
Time	dd/mm/yyyy hh:mm:ss	1

Respiration is measured as impedance of the chest, the respiratory rate is the amount of times the curve goes through zero per minute. The plethysmograph detects changes in the volume of arterial blood with each pulse beat. From the graph the peaks can be detected and the heart rate can be derived.
- Trends
A text file containing trend data with a sample frequency equal to 0.1 Hz: Time [sec.], Date [mm/dd/yyyy], Heart Rate[beats/min], Respiratory Rate [breaths/min], Non-invasive blood pressure [mmHg], peripheral Oxygen Saturation [%] and Body Temperature [°C].
- Alarm
A registration of the alarm log and settings.
- Events
Log of possible other events.
- Patient Index
All recordings belonging to one patient name and or bed place.

For this study only the ECG and Heart Rate data is used.

3. ECG

To retrieve HRV, a translation from ECG to RR interval is made, see figure 11.

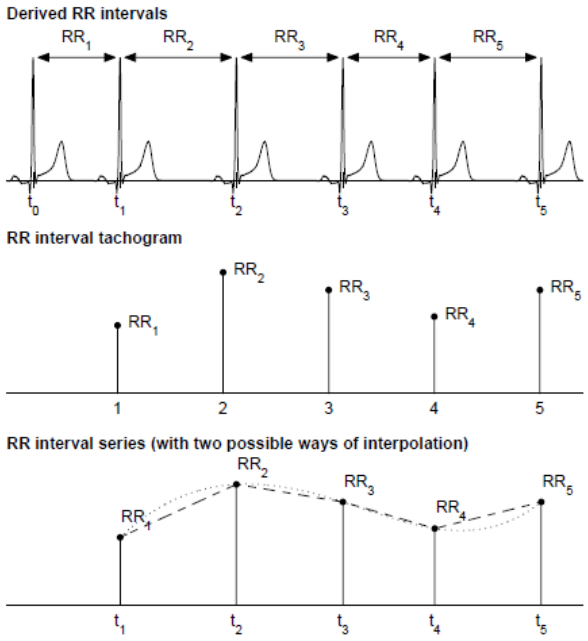

Figure 11 Derivation of RR interval from ECG signal [25]

Different options exist for deriving the RR interval:

1. Heart rate

A simple calculation of RR interval can be made using the following formula:

$$RR\ [sec.] = \frac{60}{HR}$$

However,, since the HR is measured in beats per min and with a sampling frequency of 0.1 Hz, timing becomes an issue. Especially since the HR itself is also based on a moving average over several heart beats. An option is for example to take the average of multiple HR measurements over 1 minute and calculate the average RR interval for that minute. Another option is to calculate the RR for every sample. Neither way this option isn't very accurate.

2. Pulse oximeter

A pulse oximeter measures the blood flow in capillaries near the sensor, when the heart contracts a pulse of arterial blood passes the sensor which is seen in the data. In this shape peaks can be detected to calculate the RR interval. However the shape is quite different from the ECG and due to inertia of the blood flow often stretched out. The calculation of the RR interval can become unreliable.

3. ECG feature extractor

In Labview (National Instruments, Austin, United States of America, version 2015 SP1) a special biomedical toolkit is designed to extract features from the ECG. One of the features is the time of the R wave. However Labview is most suited to analyse a continuous real time input of data and log this, while showing a nice visualisation of the data in the meantime. Since the ECG data is retrieved retrospectively Matlab (the MathWorks, Massachusetts, United States of America, version R2015b) was chosen as a better suited program for the analysis.

4. Peak detection

Peak detection can easily be performed in Matlab. A peak can be defined as a data point larger than the 2 neighbouring samples or the point where the derivative becomes zero. When you know the index

of the peak, time between peaks can be calculated. However, to be able to detect the R peak in the ECG, filters are recommended to improve the signal to noise ratio [9]. Common noise sources in ECG signal are:

- muscle noise
- artefacts due to electrode motion
- power-line interference
- baseline wander
- T waves with high frequency characteristics similar to QRS complexes

5. Pan-Tompkins algorithm

The Pan-Tompkins algorithm is developed to detect the R peaks and derive the amplitude [37]. The algorithm consist of a pre-processing part and some decision rules. This is a very solid method and used in other HRV software tools like Kubios. The Pan-Tompkins algorithm is chosen to derive the RR interval, since this method is proven to be specific in detecting R peaks [9].

The pre-processing consists of filtering, squaring and a moving average.
The first filter is a bandpass filter between 5-15 Hz to remove baseline wander and high frequency muscle noise. The second filter is a derivative filter to high light the QRS complex, but also amplifies higher frequency noise components which could be still in the signal. Another amplification is done by normalizing the signal through squaring. Then the signal is averaged with a moving window to get rid of random white noise; a moving average over 30 points (15 to the left and 15 to the right). The signal now shows some clear R peaks, but it's now important to decide if these correspond to a QRS complex with a decision rule algorithm. For each peak the following options are possible:

- QRS complex
- High sloped T wave
- Noise artefact

The QRS complexes are localized using peak detection. Two threshold are set based on a short sample of the signal:

1. Signal threshold; 25% of the maximum amplitude .
2. Noise threshold; 50% of the mean of the signal.

The QRS complexed are checked based on the threshold. The signal and noise threshold are continuously updated. If the peak is smaller than the current signal threshold, the algorithm assumes the peak isn't a QRS complex. However when this happens often on a row, a search back is done to detect the missing peaks until 1.66 times the current RR interval. Physically the ECG can't change faster than this time. On the other hand a minimum distance of 40 samples between peaks is set as a minimum. It's physiological not possible that this distance is smaller, since 40 samples correspond to 200 msec. and a HR of 300 bpm. Ventricular depolarization can't occur during the refractory period of the heart, found R peaks within in this period are false positives and removed.
The last option to check is a T wave. The slope of the T wave is less steep than the R wave, so the slope of the previous R peak is checked. If the slope is less than half of previous value it's assigned to a T wave. Based on the index of the R peak and the sampling frequency, the time between to peaks is calculated. In the appendix the complete implementation in Matlab is found.

A possible limitation is the sampling rate of 200 Hz. The optimal sampling range for detecting R peaks is 250 to 500 Hz, lower sampling rates can produce errors in the estimation of the R wave [8].

4. Covariates

Covariates are the predictors of the model, they are selected based on literature and expert experience [4, 18]. Covariates are divided into baseline, time-varying and outcome parameters. The baseline covariates are covariates that remain the same during the entire measurement period, see table 8. This is a selection of the collected baseline patient information which could be useful when predicting sepsis.

Table 8 Baseline covariates

Parameter	Unit	Formula	Source
Gender	-	Male = 1, female = 2	PDMS
Gestational Age	Days	$GA_{adj} = GA_i - \frac{\sum_{i=1}^{n} GA}{N}$	PDMS
Birthweight	Gram	Percentile	PDMS

In literature evidence may be found concerning the elevated susceptibility of males to bacterial infections, since compared to females, males have a lower immune response [38, 39]. However no literature has been found about this difference in premature infants, this study could have a relevant contribution about this aspect. Gender is added as a binary covariate to the model.

Gestational Age is rescaled around the mean of the entire cohort. Rescaling for parameters like age is recommended since zero age has no true meaning. After rescaling the intercept b_0 can be interpreted. The intercept equals the dependent variable when the independent variable is equal to zero. Gestational Age is physically never equal to zero, therefore the value of intercept becomes useless without rescaling. When analysing the data the rescaling needs to be taken into account, since the covariate now represents a deviation from the group mean. $y = b_o + b_1 x$

Another form of rescaling is used for birth weight. The relation between birthweight and the chance of developing sepsis is already explained. In health care growth curves are used as an auxiliary to monitor deviations in grow of the child compared to peer. Special birth weight curves are developed for premature infants, see figure 12 [40]. The figure shows that infants born after 32 weeks of pregnancy, normally weight 1500 gram at birth, explaining this combination of GA and birthweight as inclusion criterion. To model birth weight, the percentile of the normal birth weight growth curve adjusted for GA, gender and the amount of pregnancies the mother endured is used.

The time-varying covariates vary over time and are presented in table 9.

Table 9 Time varying covariates

Parameter	Unit	Sampling frequency	Formula		Source
Weight	Gram	No standard	$\%Weight(i) = \frac{Weight(i) - Weight(1)}{Weight(1)}$ For i = 1 : t		PDMS
RR	-	-	-		Dräger
CRP	mg/L	No standard	CRP = 1	No measurement	PDMS
			CRP = 2	CRP<10	
			CRP = 3	CRP≥10	
Bradycardia	-	0.1 Hz	Count (HR >10 & <100)		Dräger
Time	Hours	Continuous	Linear		Dräger

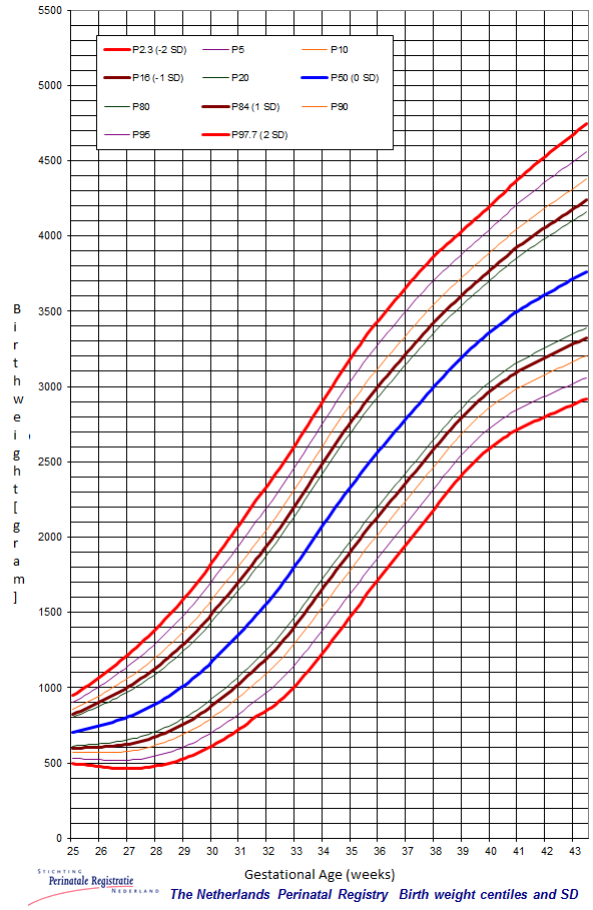

Figure 12 Birth weight growth curve [40]

Birth weight growth curves cannot be used to monitor grow over time because they tend to underestimate the growth restriction. The growth curves used for term infants aren't suited either, since premature infants tend to lose weight in the first period after birth. This weight loss is mostly a result of the intolerance to food, since the gestational system isn't fully developed, and the loss of liquid [41]. No suited growth curves adjusted for GA are found, so percentage weight loss or gain over time is used as a predictor.

To indicate HRV multiple variables are possible, information in the relevant literature is retrieved about the relation between these parameters and sepsis. In the previous section the problems with using the frequency domain are already mentioned; aliasing of power beyond Nyquist frequency, uneven sampling and the disturbance due to respiration [26]. Besides these problems, Bohanon et al. (2015) already proved in a small cohort of patients that time domain and nonlinear methods are more sensitive than frequency domain methods and significant different between non- and septic patients [7, 18]. The choice of which parameter is used in the final model is based on tested performed on the cohort of patients. The results are discussed in section 4.

CRP is only measured based on indication of the neonatologist since it requires a blood sample. Therefore to model CRP a factorial scale is used with three options. A distinction is made between no measurement, a low outcome and a high outcome [12]. The option of no measurement needs to be included since CRP isn't measured often and this doesn't mean that the value is low at those times.

Besides changes in HRV, dips in HR called bradycardia are often present preceding sepsis [35]. Bradycardia is defined as a baseline HR below 110 bpm, the baseline HR should be the average of at least 10 min. of data [42]. In this study the amount of drops during one hour is taken, since this matches the time axis.

Every patient has an individual linear time axis. This can be used to interpret the output of the predictive model. The axis starts at zero, this is the start of the first measurement of ECG and not identical to the birth time since measurement of ECG almost never starts immediately.

The outcome covariate is presented in table 10.

Table 10 Response variable

Parameter	Unit	Formula
Date & time	d-m-year 00:00:00	-
Authorisation date & time	d-m-year 00:00:00	-
Outcome	-	Positive = 1, negative = 0
Bacteria (if outcome is positive)	-	-
Type of sepsis	hours	EOS: Date sepsis – date of birth ≤ 24 hours LONS: Date sepsis – date of birth > 24 hours

Only the first event of sepsis of each patient was noted. When no ECG data was presented around the moment of sepsis, data 24 hours before or after the moment of sepsis were assigned to the moment of sepsis. When the time of the blood culture was unknow it was set to 00:00:00. The moment of sepsis marks the end of data collection and one block of an hour was finished after the time of sepsis. Type of sepsis was needed to exclude EOS.

5. Statistical analysis

In Matlab (the MathWorks, Massachusetts, United States of America, version R2015b) a longitudinal database was created. For longitudinal analysis the measurements are taken for the same response variable at multiple occasions for each subject, resulting in a 'response profile' per subject, see table 11. Two databases are build, one with an averaging window over six hours and one over one hour.

Table 11 Example of the database for patient 3

Patient number	Time	GA	Gender	Birth weight	CRP	Weight	Response	HRV
3	0	11	2	20	0	0	0	119,6
3	6	11	2	20	0	-0.01	0	128
3	12	11	2	20	0	-0.02	0	156

If the value of a certain parameter wasn't known at a certain time point the field was left empty. The resulting longitudinal dataset contains, per subject a different amount of data points.

For the statistical analysis the programme R (Rstudio, Boston, United Stated of America, version 3.3.2) is used. Multiple models are suited for longitudinal analysis, however only some for binary responses. The data is not independent and identically distributed due to repeated measures. Two statistical approaches to this problem are discussed.

1. Generalized Estimating Equations (GEE)

This approach is based on general linear models: $y \sim N(x_i^T \beta, \sigma^2)$

The response variable is a combination of known covariates x_i and estimates β. The model is fit through least squares. Generalized linear models are and extension of these model, allowing the covariates to exist of a nonlinear combination.
GEE add another structure to the formula; working correlation matrix. The matrix contains the within-subject association. The estimates are now derived based on the response variable, covariates and a variance-covariance matrix, V, with subject specific characteristics on the diagonal.
V can be exchangeable (all correlations are the same), autoregressive (correlation depends on distance between observations) or unstructured (all correlation terms are different). In practise GEE is not robust to missing data points in the longitudinal database. Unfortunately the rate of missing data points is high in this study [17, 43].

2. Mixed Effect Models
To account for the repeated measures a subject specific random effect is added in mixed effect models. The within subject correlation is induced by the same random effect over time for a certain subject. The term mixed refers to the addition of fixed and random effects.
The levels of fixed effect will stay the same if experiment would be repeated. The variation in the dependent variable is explained by independent variables.
The levels of random effect are randomly selected from a population, the variation is not explained by independent variables. For example the cohort of patients will be different when the study is repeated, since the cohort is random sample of the population. The within-subject covariance matrix now consists of the variance of the subject effect and of a random error. In the formula this results in a patient specific intercept.

$$y_{ij} = (\beta_1 + a_i) + \beta_2 x_{2ij} + e_{ij}$$

Since the response variable is binary (yes or no), a link function is needed to make the response variable continuous and normally distributed.

$$\eta = logit(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$$

The result is the logarithm of the odds of probability of success. To analyse the outcome of the model a back transformation is needed to regain the odds [15].

$$\pi = \frac{\exp(\eta)}{1+\exp(\eta)}$$

In R the choice is made to use a logistic mixed effect model; Generalized Linear Mixed Effects Regression (package lme4, version 1.1-12). The estimations are performed through maximum likelihood; find the combination of estimates that maximize the likelihood of the response given the covariates. Maximum likelihood tends to underestimate the variance components of the random effect since it assumes that the fixed parameters are completely know. A restricted version of the maximum likelihood exists to solve for this problem, unfortunately then the fixed effects are biased. To compare models with different fixed effects maximum likelihood is recommended [44].

The output of the logistic mixed effect model using maximum likelihood consists of the following component to assess the fit of the model:
- Akaike Information Criterion

$$AIC = -2\ln(\hat{L}) + 2k$$

$\hat{L}$ = maximized value of likelihood, k = model degrees of freedom, N = number of observations

- Bayesian Information Criterion

$$BIC = -2\ln(\hat{L}) + \ln(N)\,k$$

The best model is the one with the lowest AIC and BIC score. AIC tends to overfit the model since it doesn't take into account the number of observations, while BIC tends to underfit due to the different way it takes into account the number of free parameters. In practise a combination of both values is used to score the model [45].

- Variance of random effect

The variance explains the variability of the random effect between the groups within the chosen random effect. For example if patient number is added as random effect, the variance is the variability of the intercept across patients.

- Estimates of the covariates

The estimates (including standard error, z-value and significance) of the covariates are presented. A significant estimator has a value significantly different from zero.

- Odds

Through back transformation the odds of the outcome can be derived based on the model output.

- Correlation of fixed effects

If the correlation between fixed effects is too high, greater than 0.8, this covariate doesn't add any new information to the model

Besides assessing the fit of the model, the model also needs to be validated:

- Classification table

Based on the odds of the outcome a classification table can be made to validate the outcome.

- Histogram of residuals

An assumption of the logistic mixed effect model is the normal distribution of the residuals. Residuals are the deviations between the observed data points and the regression model.

- Plot of residuals versus the odds of the outcome

The residuals should remain within the same range independent of the odds; assumption of constant variance of the error.

- Plot of residuals versus covariates
- Histogram of random effect

The mean of the random effects should be around zero and the histogram should look normally distributed.

In R the model is run different times, with different covariates and random effects:

1. Compare different HRV options to find the HRV with the best predictive score
2. Compare with and without CRP
3. Build best model with the best fit based on step 1 and 2

The find the model with the best fit a stepwise process is needed. First the database is analysed to check all variables for strange outliers and scale. When difference between scales of the covariates is too big, this can cause problems with the converging of the model in R. The solution is to standardize the database using the z-score for each covariate:

$$X = \frac{x - \bar{x}}{\sigma_x}$$

The resulting variable is dimensionless and represents the amount of standard deviations from the mean, negative is below the mean and positive is above the mean). Second, all possible random and fixed effects are added to the model. Step for step random and fixed effect out are left out based on significance, AIC & BIC value and correlation of fixed effect to find the model with the best fit. Additionally non linearity can be added to the covariates if a nonlinear relation is expected. A plot of

the odds of the outcome versus the covariate can reveal this nonlinear relation. A polynomial or splines can be fitted to the covariate, for example using generalized additive models.

Sample size

Since the time period was chosen based on the availability of data, all available data was used and the study has an observational design. To assure the samples is large enough to ensure adequate power for logistic mixed effect modelling, the golden rule was used: a minimum of 10 patients or events for each variable included in the model. A minimum of 40 patients was needed. Typically, for logistic mixed effect modelling simulation of data is used to derive the minimum detectable difference for the between- and within-subject factor. There does not appear to be a consensus in the relevant literature about the use of power analysis for mixed effect models with simulated data.

4. Results

Besides the results elaborated in section 1, some additional information about the study population, the choice of the HRV parameter, selection of the database and additional graphs are discussed in this section.

1. Study population

Exclusion The total amount of births between January and June 2016 was 971. A number of 884 patients were excluded because they didn't met the inclusion criteria of GA, birthweight and being born in the Sophia Children's Hospital. Four patients were excluded since they underwent one or more major surgical operations:

1. Laparotomy for atresia of the terminal ileum and stenosis of the small intestine.
2. Repair of the abdominal wall due to gastroschisis; the abdominal content sticks outside of the body through a hole in the belly of the infant.
3. Rickham OK; placement of the ommaya reservoir for aspiration of the cerebral fluid.
4. Ileocecal resection; removal of parts of the small bowel.

One case of NEC stage 3 with pneumatosis was detected and two cases of EONS. For another 20 patients no data was available.

Sepsis During the study period there was an outbreak of serratia marcescens and staphylococcus aureus. Table 12 shows the different bacteria causing sepsis in this cohort. In five case two kind of bacteria were causing sepsis.

Table 12 The different bacteria that cause sepsis

Bacteria	N (%)
Enterococcus faecalis	1 (4,77)
Escheria coli	4 (19,04)
Klebsiella pneumoniae	1 (4,77)
Staphylococcus aureus	4 (19,04)
Staphylococcus capitis	8 (38,10)
Staphylococcus epidermidis	7 (33,33)
Staphylococcus haemolyticus	1 (4,77)

Serratia marcescens only found when patient already developed sepsis earlier with different bacteria, on the other hand four cases of sepsis due to staphylococcus aureus are detected. The kind of bacteria causing sepsis isn't expected to have influence on the way sepsis can be detected. Table 13 describes the characteristics of sepsis of in- and excluded patients. Septic patients were excluded based on EONS (2 times), no ECG data available (3 times) or presence of a major surgical operation (3 times). The excluded patients developed sepsis on average four days earlier after birth than the included patients. The authorisation of the blood culture takes about 3.5 days.

Table 13 Sepsis characteristics

Sepsis characteristic *Med(IQR)*	Included (n=21)	Excluded (n=8)
Authorisation – sample [hours]	90,22 (74,63 – 108,60)	80,28 (72,34 – 84,93)
Times sepsis per administration [-]	1,0 (1,0 – 2,0)	1,0 (1,0 – 3,5)
Blood culture – date of birth [days]	9,00 (6,75 – 14,00)	5,00 (2,25 – 15,5)

2. Heart Rate Variability

The choice of the HRV parameter is based on a simple predictive model. The general fit of the model for the different HRV parameters is compared. Besides HRV, only the baseline covariates, time and the response of sepsis are included in the model. Figure 13 presents the results.

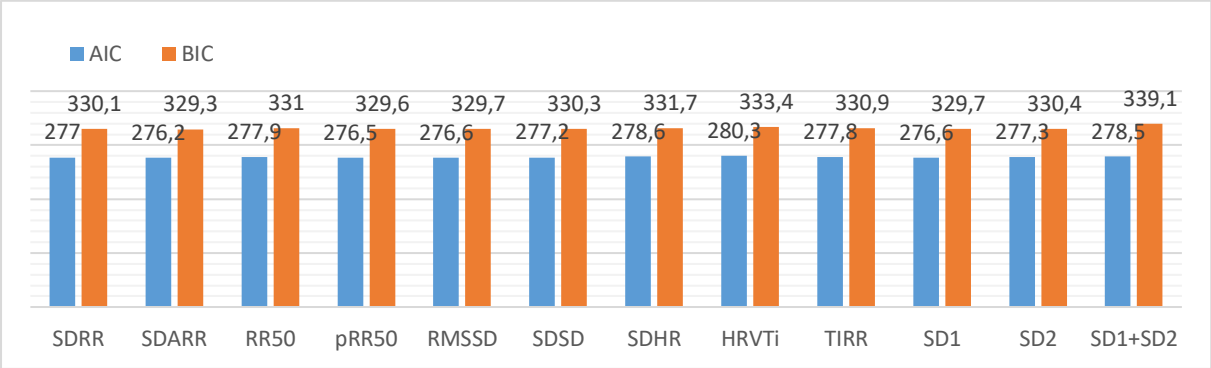

Figure 13 AIC and BIC for different HRV measures

Figure 13 shows that SDARR has both the lowest AIC as BIC value. However, the differences are minimal. Therefore also the significance of the estimate is analysed. The estimate of SDARR was the only one whom was significantly different from zero. That's why SDARR is chosen for further analysis.

3. Selection of database

DATABASE 1 For the first database all patients with a weight below 2500 gram at birth are included. The values in the database are averaged over six hours. Table 14 describes the patient characteristics.

Table 14 Demographics of the patients included in the first database

	Septic (n= 21)	Control (n=58)	Excluded (n=27)
GA (weeks)	$27^{1/7}(26^{3/7}\text{-}28^{3/7})$	$29^{4/7}(28^{2/7}$ -31)	$30^{6/7}$ ($27^{5/7}$-$31^{5/7}$)
Sex (male : female)	16 : 5	31 : 27	11 : 16
Birth weight (g)	880 (1303-805)	1330 (980-1550)	1250 (956-1725)
Mortality (n(%))	5 (24)	2 (3)	5 (19)
Length of stay (days)	33 (21-62)	10 (5-20)	6 (3-34)

More males suffered from sepsis in this cohort than females and overall more males were included than females. In the control group the ratio was almost equal. Patients in the septic group endured shorter gestation, were born with a lower birthweight, had an increased risk of death and stayed longer in the hospital.

In a generalized linear mixed effect model with odds of developing sepsis as outcome as the dependent variable, only CRP (b=2.94 , SE=0.52) and GA (b=-0.05 , SE=0.03) were significant (all p <.05). The variance of the intercept was 0.96. The estimation of HRV was not significantly different from zero. Table 16 shows the resulting classification table with a cut off value of 0.3.

Table 15 Classification table with a cut off value of 0.3

		Predicted		
Response		0	1	Sum
	0	2783	2	2785
	1	17	3	20
	Sum	2800	5	2805

99,3 % was predicted correct. However, a cut off value of 0.3 or lower gave odds ratio's high enough to predict any case of sepsis as outcome. The moment the odds became above the threshold value is compared to the moment the blood culture was taken, resulting in the following delta:

Delta [hours]	365	76	202

This outcome means that the model is able to predict sepsis up to 10 days beforehand, which doesn't makes sense. The result can be explained by looking at figure 14 and 15. Both the RR interval for septic as the control group decreases over time. The HRV curve remains flat almost the entire time, it even increase in the hours before sepsis. Figure 15 also shows that the median of both groups lay within the boxplot of the other group, making it hard to distinguish between groups. By taking an average over six hours, any effect is probably completely averaged out.

Due to these unsatisfying results database 1 is considered to be not usable.

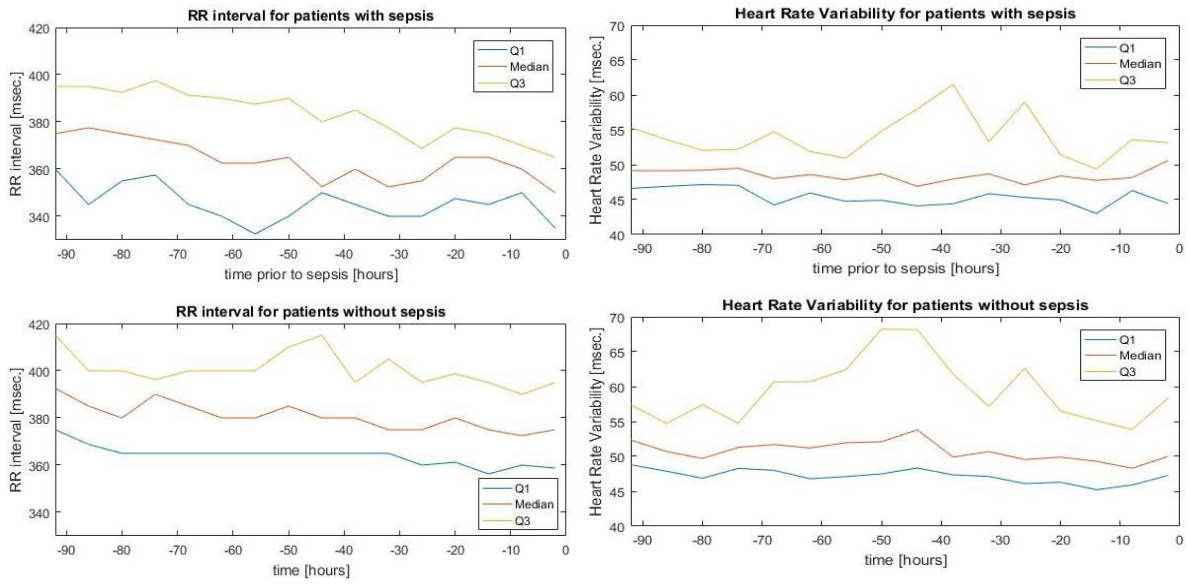

Figure 14 RR and HRV compared between patients with and without sepsis, data is averaged over six hours and presented as IQR (Q1: 25% and Q3:75%).

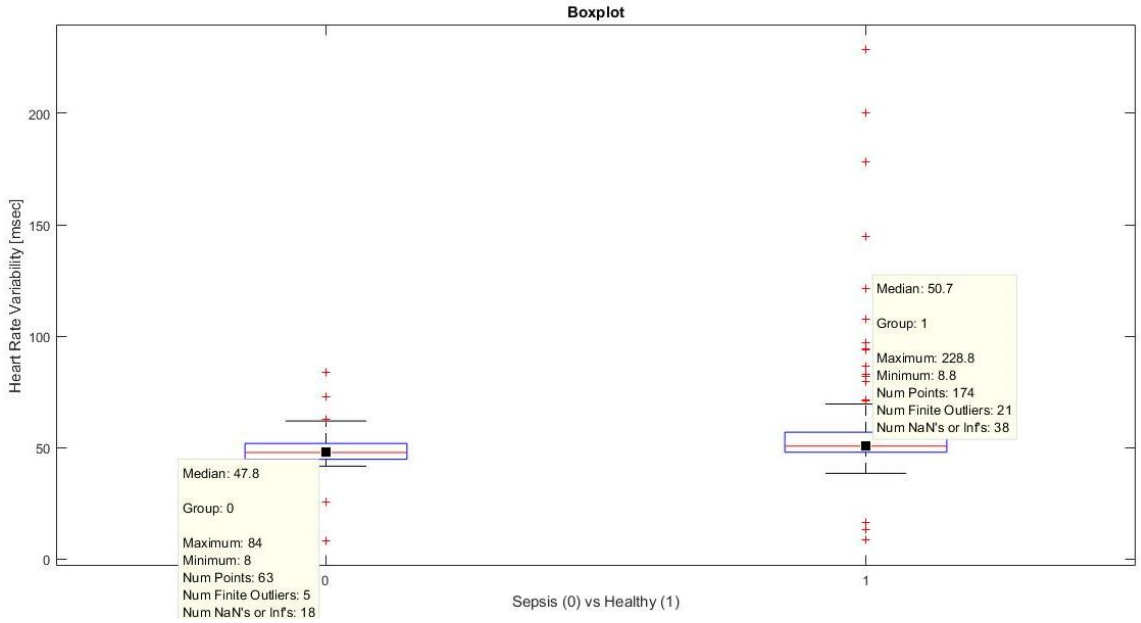

Figure 15 Boxplot of HRV for the sepsis and healthy group

DATABASE 2 For the second database only patients weighing below 1500 gram at birth are included. The values are averaged over one hour. From now on this database is used. The patient characteristics are described in section 1.

4. Generalized Logistic Mixed Effect Regression

The characteristics of model 2, discussed in section 1 and chosen as the best model, are elaborated upon. Model building was done through backward elimination of both baseline and time-varying covariates. The variance of the random effect for patient number was 0 and for bradycardia 0.5075 . Estimates of the fixed effects for gender, weight and GA were not significant. In the second GLMER model with odds of sepsis as the dependent variable the following covariates were significant (all $p<0.05$): CRP (b = 6.34, SE = 0.68), bradycardia (b = 0.70, SE = 0.28), Median RR (b = -4.04, SE = 1.5) and Spline 4 (b = 191, SE = 45.0), Spline 5 (b = -5625, SE = 38.8) and Spline 6 (b = -11120, SE = 66,1). The residuals were centred around zero and the shape of the histogram was normally distributed. The correlations between the fixed effect were all weak, expect for the correlation between Spline 4 and Spline 5, marked as moderate, and Spline 5 and Spline 6, as strong. Empirically, the threshold for the odds ratio is determined. A low threshold resulted in false positive outcomes and a high threshold resulted in many false negatives. The threshold of 0.4 scored gave the best results on both conditions. Finally, the time difference between the sepsis prediction and outcome is analysed. Resulting in zero hours as outcome for all patients in which sepsis was predicted.

5. Recommendations

Worldwide, one in ten babies is born before the expected date of birth. Complications related to premature birth are a leading cause of death among children under the age of five [1]. Mortality rates are twice as high for patients who develop sepsis with prolonged hospital stay [7, 12]. To prevent mortality related to sepsis, the methods to detect sepsis need to be improved so treatment can be started on time. The detection of sepsis can be improved through the development of a predictive model based on non-invasive clinical measurements. One of the potential measurements is HRV based on RR interval [23]. The aim of this study is to design a predictive model which captures the relationship between HRV, baseline patient characteristics and clinical measurements on the development of sepsis in premature infants. During a time frame of six months, 60 eligible patients were admitted to the NICU, of whom 18 infants experienced at least one episode of sepsis. The mortality rate of the group that developed sepsis was indeed higher and hospital stays were longer. As may be expected on the basis of the hypothesis of de-complexification amongst infants getting ill, initial inspection of the raw RR intervals showed alterations in the variability preceding sepsis. HRV is derived to capture the alterations. The odds of developing sepsis are analysed by means of a logistic mixed effect model. This study showed that the odds of developing sepsis increase when CRP increases, the amount of bradycardias increases, the median RR decreases and (depending on the range of HRV) the spline of HRV decreases. However, the model could only predict sepsis in 7 out of 18 cases, resulting in a low sensitivity of 0.39. Nevertheless in practice, every detection of sepsis is beneficial to the treatment and health of the infants. Especially, since the false positive rate is very low, no infants will get the wrong medical treatment.

A number of recommendations may be made which can possibly improve this model.
- ECG measurements are prone to artefacts due to misplaced electrodes, muscle noise or a wandering baseline. Artefacts cause missing data points in both the ECG and RR data since RR interval cannot be derived from bad ECG data. The missing data points are filled in by NaN (not a number). In this database for the HRV measures approximately 10% of the data was missing and has been replaced by nan; namely a total of 1600 out of 18000 points for one HRV measure. When these missing data points occur around the same time as the moment when sepsis develops, important predictive information is lost. Interpolation of the missing points provides a solution to this complication. Previous research suggests reconstruction based on linear prediction of the missing ECG signal results in more accurate estimates of HRV than by removal of abnormal beats in the ECG data [46].
- A baseline difference was found between the septic and control group. Consequently, during the modelling the difference in offset could be emphasised instead of the smaller relative differences over time that are related to sepsis. Baseline correction for each patient poses a better solution than the used standardization method, since the latter is derived around the entire group mean.
- The demographics of the patients show differences between the two groups. These differences are desired, since baseline covariates are entered into the model to predict sepsis. However, they did not turn out to have a significant predictive value after all. A paired samples test ensures both groups match in baseline and could be more sensitive on time-varying predictive covariates.

- The current dataset can be used to calculate the sample size that ensures adequate power. The expectation is that an increased sample size is necessary, since, as a result of longitudinal database characteristic, only 18 of the 18000 points per covariate match sepsis.

General conclusion

A great potential is found in the predictive value of both HRV and clinical measurements on the odds of developing sepsis. However, the current model is not sensitive enough to predict all cases of sepsis. The authorisation of a blood culture to prove sepsis takes about four days. Since this model predicts sepsis at the moment the blood culture is taken, this gives a four day advantage in the treatment of sepsis.

References

1. *World Health Organization; Born too soon: the global action report on preterm birth*. 2012.
2. Nederland, S.P.R., *Perinatale Registratie Nederland Grote Lijnen 1999 - 2012.* Utrecht, 2013.
3. Oza, S., et al., *Neonatal cause-of-death estimates for the early and late neonatal periods for 194 countries: 2000–2013.* Bulletin of the World Health Organization, 2015. **93**(1): p. 19-28.
4. Bekhof, J., et al., *Clinical signs to identify late-onset sepsis in preterm infants.* European journal of pediatrics, 2013. **172**(4): p. 501-508.
5. Kornelisse, R., *Bacteriële infecties op de NICU van het Erasmus MC.* Kwaliteits Informatie Systeem Sophia Kinderziekenhuis, 2010.
6. Standage, S.W. and H.R. Wong, *Biomarkers for pediatric sepsis and septic shock.* 2011.
7. Moorman, J.R., D.E. Lake, and M.P. Griffin, *Heart rate characteristics monitoring for neonatal sepsis.* IEEE transactions on biomedical engineering, 2006. **53**(1): p. 126-132.
8. Cardiology, T.F.o.t.E.S.o., *Heart rate variability standards of measurement, physiological interpretation, and clinical use.* Eur Heart J, 1996. **17**: p. 354-381.
9. Tarvainen, M.P., et al., *Kubios HRV–heart rate variability analysis software.* Computer methods and programs in biomedicine, 2014. **113**(1): p. 210-220.
10. Samuels, N., et al., *Necrotising enterocolitis and mortality in preterm infants after introduction of probiotics: a quasi-experimental study.* Scientific Reports, 2016. **6**.
11. Sullivan, B.A., et al., *Infection and other clinical correlates of abnormal heart rate characteristics in preterm infants.* The Journal of pediatrics, 2014. **164**(4): p. 775-780.
12. Stoll, B.J., et al., *Late-onset sepsis in very low birth weight neonates: the experience of the NICHD Neonatal Research Network.* Pediatrics, 2002. **110**(2): p. 285-291.
13. Alferink, K., *Monitor bewaking met betrekking tot de hartactie bij kinderen (Versie 3).* Kwaliteits Informatie Systeem Sophia Kinderziekenhuis, 2015.
14. Faraway, J.J., *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. Vol. 124. 2016: CRC press.
15. Carey, V. and Y.-G. Wang, *Mixed-effects models in S and S-PLUS*. 2001, Taylor & Francis.
16. Lenth, R.V., *Post hoc power: tables and commentary.* Iowa City: Department of Statistics and Actuarial Science, University of Iowa, 2007.
17. Dobson, A.J. and A. Barnett, *An introduction to generalized linear models*. 2008: CRC press.
18. Bohanon, F.J., et al., *Heart rate variability analysis is more sensitive at identifying neonatal sepsis than conventional vital signs.* The American Journal of Surgery, 2015. **210**(4): p. 661-667.
19. Acharya, U.R., et al., *Heart rate variability: a review.* Medical and biological engineering and computing, 2006. **44**(12): p. 1031-1051.
20. Mathworks. 15-9-2016; Available from: https://nl.mathworks.com/help/dsp/examples/real-time-ecg-qrs-detection.html.
21. Kovatchev, B.P., et al., *Sample asymmetry analysis of heart rate characteristics with application to neonatal sepsis and systemic inflammatory response syndrome.* Pediatric research, 2003. **54**(6): p. 892-898.
22. Griffin, M.P., et al., *Abnormal heart rate characteristics preceding neonatal sepsis and sepsis-like illness.* Pediatric research, 2003. **53**(6): p. 920-926.
23. Griffin, M.P. and J.R. Moorman, *Toward the early diagnosis of neonatal sepsis and sepsis-like illness using novel heart rate analysis.* Pediatrics, 2001. **107**(1): p. 97-104.
24. Cao, H., et al., *Increased nonstationarity of neonatal heart rate before the clinical diagnosis of sepsis.* Annals of biomedical engineering, 2004. **32**(2): p. 233-244.
25. Tarvainen, M.P., *Kubios HRV User's Guide version 2.2.* Biosignal Analysis and Medical Imaging Group, 2014.

26. Chang, K.L., et al., *Comparison and clinical application of frequency domain methods in analysis of neonatal heart rate time series.* Annals of biomedical engineering, 2001. **29**(9): p. 764-774.
27. Allen, J.J., A.S. Chambers, and D.N. Towers, *The many metrics of cardiac chronotropy: A pragmatic primer and a brief comparison of metrics.* Biological psychology, 2007. **74**(2): p. 243-262.
28. Jelinek, H.F., et al., *Using meta-regression data mining to improve predictions of performance based on heart rate dynamics for Australian football.* Applied Soft Computing, 2014. **14**: p. 81-87.
29. Lake, D.E., et al., *Sample entropy analysis of neonatal heart rate variability.* American Journal of Physiology-Regulatory, Integrative and Comparative Physiology, 2002. **283**(3): p. R789-R797.
30. Fairchild, K.D., *Predictive monitoring for early detection of sepsis in neonatal ICU patients.* Current opinion in pediatrics, 2013. **25**(2): p. 172-179.
31. Coggins, S.A., et al., *Heart rate characteristic index monitoring for bloodstream infection in an NICU: a 3-year experience.* Archives of Disease in Childhood-Fetal and Neonatal Edition, 2015: p. fetalneonatal-2015-309210.
32. Guzik, P., et al., *Heart rate asymmetry by Poincaré plots of RR intervals.* Biomedizinische Technik, 2006. **51**(4): p. 272-275.
33. Fairchild, K.D., et al., *Septicemia mortality reduction in neonates in a heart rate characteristics monitoring trial.* Pediatric research, 2013. **74**(5): p. 570-575.
34. *World Health Organization; Fact sheet 363 Preterm birth (online available)*. 2014 [cited view date: 30-7-2015.
35. Simonsen, K.A., et al., *Early-onset neonatal sepsis.* Clinical microbiology reviews, 2014. **27**(1): p. 21-47.
36. Ireland, N., et al., *Heart rate variability in patients recovering from general anaesthesia.* British journal of anaesthesia, 1996. **76**(5): p. 657-662.
37. Pan, J. and W.J. Tompkins, *A real-time QRS detection algorithm.* Biomedical Engineering, IEEE Transactions on, 1985(3): p. 230-236.
38. García-Gómez, E., B. González-Pedrajo, and I. Camacho-Arroyo, *Role of sex steroid hormones in bacterial-host interactions.* BioMed research international, 2012. **2013**.
39. Schröder, J., et al., *Gender differences in human sepsis.* Archives of Surgery, 1998. **133**(11): p. 1200-1205.
40. Schaap, H.W.H.P., *Wat is het normale foetale gewicht?* Nederlands Tijdschrift Geneeskunde, 2009. **153**.
41. Butler, A.S. and R.E. Behrman, *Preterm Birth:: Causes, Consequences, and Prevention*. 2007: National Academies Press.
42. Development, N.I.o.C.H.a.H., *Standard Terminology of fetal heart rate monitoring.* 2008.
43. Hothorn, T. and B.S. Everitt, *A handbook of statistical analyses using R*. 2014: CRC press.
44. Oehlert, G.W., *A few words about REML*. 2014.
45. McCulloch, C.E. and J.M. Neuhaus, *Generalized linear mixed models*. 2001: Wiley Online Library.
46. Theera-Umpon, N., P. Phiphatkhunarnon, and S. Auephanwiriyakul. *Data reconstruction for missing electrocardiogram using linear predictive coding*. in *Mechatronics and Automation, 2008. ICMA 2008. IEEE International Conference on*. 2008. IEEE.

Appendix

The appendix consists of the made Matlab (The Mathworks, Natick, United States) and R (Rstudio, Boston, United Stated of America, version 3.3.2) code to generate the results presented in this thesis.

Matlab

```
clc
clear
close all

%BW<1500 gram
Sep = [9 10 11 18 19 71 74 34 35 36 41 60 63 80 84 90 95 104];
pnmbr = [3 4 6 9 10 11 13 15 16 18 19 20 21 22 27 29 31 33 34 35 36 38 41
43 44 46 49 52 53 54 55 56 58 59 60 61 62 63 65 71 73 74 76 77 78 79 80 84
85 86 90 91 92 93 95 96 98 103 104 105];
ismem = ismember(pnmbr,Sep);
control = pnmbr(~ismem);
gen_data = xlsread('D:\Prediction model\Final\Cohort'); % general patient
information
CRP_dat = xlsread('D:\Prediction model\Final\CRPdata');
Weight = xlsread('D:\Prediction model\Final\Gewichten_tot_adj');
input = [200,1];  % fs and window

for i = 1:length(control)
 try
 tic
 [Datab2] = Datab(input,control(i),gen_data,CRP_dat,Weight);
 Name=num2str(control(i));
 Col_header = {'Pnmbr', 'time_adj', 'GA_adj'
,'Gender','BW_datab','Weight_datab','CRP_datab','CRP_adj','Response_datab',
'Max','Min','Mean','Median','SDRR','SDARR','RR50','pRR50','RMSSD','SDSD','m
eanHR','SDHR','HRVTi','TIRR','SD1','SD2','Brady'};
 xlswrite(['D:\R\Database\' Name],Col_header,'Blad1','A1:Z1') %Write column
header
 xlswrite (['D:\R\Database\' Name],Datab2,'Blad1','A2:Z1500');
 clear Datab2 Name
 toc
 catch ME
  fprintf('Error:%s\n', ME.message);
  continue; %try next pnmbr
 end
 fprintf('Succes\n')
end

% bw>1500 gram en bw<2500 gram
sep2 = [9,10,11,18,19,30,34,35,36,41,60,63,70,71,74,80,83,84,90,95,104];
pnmbr_ex =
[1,2,5,7,12,14,23,24,25,28,39,40,42,45,47,50,51,64,67,68,69,88,99,100,101,1
02,106];
Pnmbr = [1:106];

ismem = ismember(Pnmbr,pnmbr_ex);
pnmbr1 = Pnmbr(~ismem);   %remove all excluded
ismem2=ismember(pnmbr1,pnmbr);
pnmbr2=pnmbr1(~ismem2);   %remove all BW<1500 gram

for i = 1:length(pnmbr2)
 try
```

```
 tic
 [Datab2] = Datab(input,pnmbr2(i),gen_data,CRP_dat,Weight);

 Name=num2str(pnmbr2(i));
 Col_header = {'Pnmbr', 'time_adj', 'GA_adj'
,'Gender','BW_datab','Weight_datab','CRP_datab','CRP_adj','Response_datab',
'Max','Min','Mean','Median','SDRR','SDARR','RR50','pRR50','RMSSD','SDSD','m
eanHR','SDHR','HRVTi','TIRR','SD1','SD2','Brady'};
 xlswrite(['D:\R\Database\Excl\' Name],Col_header,'Blad1','A1:Z1') %Write
column header
 xlswrite (['D:\R\Database\Excl\' Name],Datab2,'Blad1','A2:Z1500');
 clear Datab2 Name
 toc
 catch ME
  fprintf('Error:%S\n',ME.message)
 end
 fprintf('Succes\n')
end

function [Datab2] = Datab(input,pnmbr,gen_data,CRP_dat,Weight)

%% Choices input parameters
fs = input(1);
conv = 693960;            % correctie excel tijd
Window = input(2);           % window for averaging in hours

%% Open general patient info xls file
j = find(gen_data(:,1)==pnmbr);       % index patient
gender = gen_data(j,2);          % 1 = male, 2 = female
GA = gen_data(j,3);           % [days]
GA_mean = round(mean(gen_data(:,3)));
GA_adj = GA - GA_mean;          % centered GA
multiple = gen_data(j,4);         % 1 = primi, 2 = multiparea
response = gen_data(j,5);         % 1 = neg, 2 = sepsis
if response == 2;
 date_sep = gen_data(j,6);        % date + time
end

clear j GA_mean gen_data
%% CRP
CRP_nmbr = CRP_dat(:,1);
i = find (CRP_nmbr == pnmbr );
if isempty(i) ==1;           % no data
 CRP_adj = nan;
 CRP_time = nan;
else CRP = CRP_dat(i,3);
 CRP(CRP<10) = 1;          % makes CRP categorical
 CRP(CRP>=10) = 2;
 CRP_adj = CRP;
 CRP_time = CRP_dat(i,2);
end
clear CRP CRP_dat CRP_nmbr i

%% BW

z = find(Weight(:,1)==pnmbr);
zz = Weight(z,3) == Weight(z,4);       % indicate birthweight
Weights = Weight(z,2);          % weight in kg
bw = Weights(zz)*1000;          % Growthcurve in grams
[perc_bw] = Birthweight(gender,multiple,GA,bw);
```

```
%% Percentage Weight
Dates = Weight(z,4);
[Weight_time,I]=sort(Dates);        % chronological order
Weight_adj = Weights(I);
perc_tot = zeros(nnz(Weight_adj),1);
for i = 1:size(Weight_adj);
 perc = (Weight_adj(i)-Weight_adj(1))/Weight_adj(1); % first value always
zero
 perc_tot(i) = perc(1);
end
clear Weight z zz Weights bw Dates Weight_adj i GA multiple perc

%% open ECG file & start loop

pathname = ['D:\Final data\p',num2str(pnmbr)];
filename = dir([pathname,'\*.csv']);

[m,~]=size(filename);
number_def=zeros(m,1);

for i = 1:m          %sort filenames according to number
 t = strfind(filename(i).name,'_');
 t2 = strfind(filename(i).name,'.');
 z = filename(i).name;

 number= str2double(z(t+1:t2-1));
 number_def(i)=number;
end

[~,I]=sort(number_def);


for i = 1:m          % run loop for amount of ecg files

 path_def = [pathname,'\',filename(I(i)).name];
 ECG = importdata(path_def);

 % Define Start and End
 Start_meas = datenum(ECG.textdata(9,1),'yyyy/mm/dd') - conv;
 End_meas = datenum(ECG.textdata(end,1),'yyyy/mm/dd') - conv;

 Start_time =
etime(datevec(ECG.textdata(9,2),'HH:MM:SS'),datevec('00:00:00',
'HH:MM:SS'))/3600/24; %time as fraction
 End_time =
etime(datevec(ECG.textdata(end,2),'HH:MM:SS'),datevec('00:00:00',
'HH:MM:SS'))/3600/24;

 Start(i) = Start_meas + Start_time;
 End(i) = End_meas + End_time;

 clear Start_meas End_meas Start_time End_time text

 data = ECG.data;
 l = length(data);
 ecg(1:l,i) = data(:,1);
 clear ECG data l
```

```
 % Cut data based on incidence of sepsis
 if response == 1;          % No sepsis
 n(i) = floor((End(i)-Start(i)) * 24);     % duration of time block in
hours
 Response(1:n(i),i) = ones(n(i),1);      % column for each file

 else response = 2;          % 2 = sepsis

 n = 0;
  if date_sep > Start(i) && date_sep < End(i);
  delta = round((date_sep - Start(i)) * 24);

  End(i) = Start(i)+ delta/24;
  datapunten = delta*60*60*fs;    %datapoints
  ecg2 = ecg(:,i);
  ecg(:,i) = [];          %make column empty
  ecg(1:datapunten,i) = ecg2(1:datapunten);

  %change resp.
  Response(1:delta,i) = [ones(delta-1,1);2];%stop opening data

  clear ecg2 delta datapunten
  break
  elseif date_sep > End(i);       % open next file in for loop till moment
of sepsis
  n(i) = floor((End(i)-Start(i)) * 24);
  Response(1:n(i),i) = ones(n(i),1);
   if i == m(end) && date_sep < End(i)+1;   %if this is the last data file
and sepsis within next 24 hours, mark this point as sepsis
   Response(n(i),i) = 2;
   end
  else date_sep < Start(i);       % sepsis before measurement due to
missing data
  datapunten = 6*60*60*fs;
  ecg2 = ecg(:,i);
  ecg(:,i) = [];
  ecg(1:datapunten,i) = ecg2(1:datapunten);
  End(i) = Start(i) + 6/24;       % use next 6 hours of known data
  Response(1:6,i) = [ones(5,1);2];
  clear ecg2 datapunten

  break

  end

 end


end

clear gen_data i n response lengt filename filenames path_def path conv
date_sep

%% Time axis
time = (0:1:((End(end)-Start(1))*24))';     % hours
N = length(time);          % amount of datapoints

%% Open trend data and calculate bradycards
```

```
pathname2 = ['D:\Final data\p',num2str(pnmbr)];

filename2 = dir([pathname2,'\*.txt']);

[m2,~]=size(filename2);

fs_tr = 0.1;
punten2 = round(fs_tr*Window*60*60);      % datapoints per 1 hours
Bradycard_def = [];
conv2 = 1452077680;

% Startpunt tijdsas

for l = 1:m2
 path_def = [pathname2,'\',filename2(l).name];
 trends = tdfread(path_def);
 Time =trends.JULIAN;

 Tim_adj(l) = (Time(1)-conv2)/3600/24 + 42375.45;

end

[Time,I]=sort(Tim_adj); %chronological order



for l = 1:m2
 path_def = [pathname2,'\',filename2(I(l)).name]; %openen op chronologische
volgorde
 trends = tdfread(path_def);

 if ischar(trends.HR) == 1;
 HR =str2double(cellstr(trends.HR));
 else
 HR=trends.HR;
 end


 if (Time(l)-Start(1)) < 0
  delta= (Start(1)-Time(l))*24; %pas starten als ecg is begonnen
  HR_def = HR(round(delta*fs_tr*3600):end);
  tt=0;
 else (Time(l)-Start(1))>= 0;
  tt= round((Time(l)-Start(1))*24); %[per blokken van 6 (24/6 =*4) offset,
nu per uur]
  HR_def = HR;
 end

 sz = floor(length(HR_def)/punten2);
 Bradycard = [];

 for ll = 1:sz
  HR_t = HR_def((ll-1)*punten2+1:ll*punten2);
  i = find(HR_t < 100 & HR_t > 10);     % bradycards HR<100, not when HR is
zero
  Bradycard(ll) = length(i);
 end
```

```
 Bradycard_def(tt+1:tt+ll,:) = Bradycard';
 clear Bradycard HR_def
end
clear fs_tr punten2 lengt2 filename2 filenames2 path_def path2 trends HR sz
ll i Bradycard c HR_t l conv2 Time I Tim_adj tt

%% Filter ECG
ecg_cor = [];
[~,C] = size(ecg);
for a = 1:C
 ecg2 = ecg(:,a);
 ecg2(ecg2>400) = nan;          % Filter large movement artifacts
 ecg2(ecg2<-400) = nan;

 ecg_cor(:,a) = ecg2(:,1);         %#ok<*SAGROW>
end
clear ecg ecg2 a C

%% Output

[Output] = HRV2(ecg_cor,Window,fs,N,Start);
%[RR_Filt] = RR_calc(ecg_cor,Window,fs,N,Start)

%% Database

%CRP

if isnan(CRP_adj(1)) == 1          % no CRP data
 CRP_datab = zeros(N,1);
else
del = (CRP_time - Start(1))*24;         % index
CRP_datab = zeros(N,1);
Del = del(del>=0 & del<=N);          % within measurement time
 if isempty(Del) ==1           % no data, so database is zeros
 else
  for zz = 1: length(Del)
  Del2(zz) = find(del==Del(zz));
  CRP_datab(round(Del(zz))) = CRP_adj(Del2(zz));
  end
 end
end

clear del Del Del2 zz CRP_adj Window

%Weight

delta2 = ((Weight_time +0.5) - Start(1))*24;    %every measurement at
12:00:00
Weight_datab = nan(N,1);
Delta2 = delta2(delta2>=0 & delta2<=N);      %index time
if isempty(Delta2) == 1

else
  for z = 1: length(Delta2)
   D = find (delta2 == Delta2(z));
   Delta3(z) = D(1); % index output
   if Delta2(z) ==0
    Delta2(z)=1;
   else
```

```
   end
   Weight_datab(round(Delta2(z))) = perc_tot(Delta3(z));
  end
end


clear delta2 Delta2 z Delta3

% Response

delta = round((Start-Start(1))*24);       % startingpoint
Response_datab = ones(N,1);

[~,c] = size(Start);
for cc = 1:c
 start = delta(cc);          % startingpoint
 Response2 = Response(:,cc);        % selection colomn
 Response2(Response2 == 0) = [];
 [r2,~] = size(Response2);
%  if start == 0
%   Response_datab(start+1:start+r2)=Response2;
%  else
 Response_datab(start+1:start+r2)=Response2;
 %end
end

clear delta r c cc start Response2 r2 Response

% Final database
Pnmbr = pnmbr* ones(N,1);
GA_adj = GA_adj* ones(N,1);
Gender = gender* ones(N,1);
BW_datab = perc_bw * ones(N,1);

Brady_datab = zeros(N,1);
if length(Bradycard_def) < N
Brady_datab(1:length(Bradycard_def)) = Bradycard_def;
else
Brady_datab = Bradycard_def(1:N);
end

clear pnmbr gender perc_bw Bradycard_def

% linear interpolation
Weight_databn = (inpaint_nans(Weight_datab));

CRP_adj = CRP_datab;
CRP_adj(CRP_adj>=2) = 1;

%% Complete database
Datab2 = [Pnmbr, time, GA_adj
,Gender,BW_datab,Weight_databn,CRP_datab,CRP_adj,Response_datab,Output,
Brady_datab];

%xlswrite (['D:\Prediction model\Pilot\Databases\'
Name],RR_Filt,'Blad2','A2:U1000');
end

function [ Output] = HRV2(ecg,window,fs,N,Start)
```

```
%calculate heart rate variability

punten = window*fs*60*60;       %point for one block of data
[~,c] = size(ecg);

Output = nan(N,16);
for i = 1:c           %per file
 ecg2 = ecg(:,i);
 ecg2= ecg2(find(ecg2,1,'first'):find(ecg2,1,'last'));%correctie
automatische nullen

 sz = size(ecg2);
 sz = floor(sz(1)/punten);      %hele blokken van 1 uur

 tt = round((Start(i)-Start(1))*24);

  for j = 1: sz
  ecg_t = ecg2((j-1)*punten+1:j*punten);  %opsplitsen data
  ecg_t(ecg_t==0) = [];
  ecg_t = ecg_t(find(~isnan(ecg_t) > 0, 1 ,'first'):find(~isnan(ecg_t) > 0,
1 , 'last'));
   if length(ecg_t) < 2*fs

   clear ecg_t
   else
   [~,qrsind,~]=pan_tompkin(ecg_t,fs,0);  %RR interval
   t= (1/fs)*qrsind;
   k=1:(size(t')-1);
   RR_t=((t(k+1))-t(k))';      %ibi voor HRV analyse
    if isempty(RR_t) ==1
    clear RR_t

    else

    [RR_filt] = deleteoutliers(RR_t,0.05,0);
    t_filt = cumsum(RR_filt) ;     %new time axis

    win = 300;  %sdnn [s]
    xx = 50;  %nn50 [ms]
    output = timeDomainHRV([t_filt,RR_filt],win,xx);
    output2 = poincareHRV([t_filt,RR_filt]);
    %output =
freqDomainHRV(ibi,VLF,LF,HF,AR_order,window,noverlap,nfft,fs,methods,flagPl
ot)

    OUTput(1,1) = output.max;
    OUTput(1,2) = output.min;
    OUTput(1,3) = output.mean;
    OUTput(1,4) = output.median;
    OUTput(1,5) = output.SDNN;
    OUTput(1,6) = output.SDANN;
    OUTput(1,7) = output.NNx;
    OUTput(1,8) = output.pNNx;
    OUTput(1,9) = output.RMSSD;
    OUTput(1,10)= output.SDNNi ;
    OUTput(1,11)= output.meanHR;
    OUTput(1,12)= output.sdHR;
    OUTput(1,13)= output.HRVTi;
    OUTput(1,14) = output.TINN;
```

```
    OUTput(1,15)=output2.SD1;
    OUTput(1,16)=output2.SD2;

    t2 = tt+j;
    Output(t2,:) = OUTput;

    clear ecg_t RR_t RR_filt OUTput output
    end

   end
  end


 clear ecg2
end
end
```

Pan-Tompkin algorithm [1]

```
function [qrs_amp_raw,qrs_i_raw,delay]=pan_tompkin(ecg,fs,gr)
if ~isvector(ecg)
 error('ecg must be a row or column vector');
end

if nargin < 3
 gr = 1; % on default the function always plots
end
ecg = ecg(:); % vectorize

%% Initialize
qrs_c =[]; %amplitude of R
qrs_i =[]; %index
SIG_LEV = 0;
nois_c =[];
nois_i =[];
delay = 0;
skip = 0; % becomes one when a T wave is detected
not_nois = 0; % it is not noise when not_nois = 1
selected_RR =[]; % Selected RR intervals
m_selected_RR = 0;
mean_RR = 0;
qrs_i_raw =[];
qrs_amp_raw=[];
ser_back = 0;
test_m = 0;
SIGL_buf = [];
NOISL_buf = [];
THRS_buf = [];
SIGL_buf1 = [];
NOISL_buf1 = [];
THRS_buf1 = [];

%% Plot differently based on filtering settings
if gr
 if fs == 200
 figure, ax(1)=subplot(321);plot(ecg);axis tight;title('Raw ECG Signal');
 else
 figure, ax(1)=subplot(3,2,[1 2]);plot(ecg);axis tight;title('Raw ECG Signal');
 end
end
%% Noise cancelation(Filtering) % Filters (Filter in between 5-15 Hz)
if fs == 200
%% Low Pass Filter H(z) = ((1 - z^(-6))^2)/(1 - z^(-1))^2
b = [1 0 0 0 0 0 -2 0 0 0 0 0 1];
a = [1 -2 1];
h_l = filter(b,a,[1 zeros(1,12)]);
ecg_l = conv (ecg ,h_l);
ecg_l = ecg_l/ max( abs(ecg_l));
delay = 6; %based on the paper
if gr
ax(2)=subplot(322);plot(ecg_l);axis tight;title('Low pass filtered');
end
```

```
%% High Pass filter H(z) = (-1+32z^(-16)+z^(-32))/(1+z^(-1))
b = [-1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 32 -32 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1];
a = [1 -1];
h_h = filter(b,a,[1 zeros(1,32)]);
ecg_h = conv (ecg_l ,h_h);
ecg_h = ecg_h/ max( abs(ecg_h));
delay = delay + 16; % 16 samples for highpass filtering
if gr
ax(3)=subplot(323);plot(ecg_h);axis tight;title('High Pass Filtered');
end
else
%% bandpass filter for Noise cancelation of other sampling frequencies(Filtering)
f1=5; %cuttoff low frequency to get rid of baseline wander
f2=15; %cuttoff frequency to discard high frequency noise
Wn=[f1 f2]*2/fs; % cutt off based on fs
N = 3; % order of 3 less processing
[a,b] = butter(N,Wn); %bandpass filtering
ecg_h = filtfilt(a,b,ecg);
ecg_h = ecg_h/ max( abs(ecg_h));
if gr
ax(3)=subplot(323);plot(ecg_h);axis tight;title('Band Pass Filtered');
end
end
%% derivative filter H(z) = (1/8T)(-z^(-2) - 2z^(-1) + 2z + z^(2))
h_d = [-1 -2 0 2 1]*(1/8);%1/8*fs
ecg_d = conv (ecg_h ,h_d);
ecg_d = ecg_d/max(ecg_d);
delay = delay + 2; % delay of derivative filter 2 samples
if gr
ax(4)=subplot(324);plot(ecg_d);axis tight;title('Filtered with the derivative filter');
end
%% Squaring nonlinearly enhance the dominant peaks
ecg_s = ecg_d.^2;
if gr
ax(5)=subplot(325);plot(ecg_s);axis tight;title('Squared');
end

%% Moving average Y(nt) = (1/N)[x(nT-(N - 1)T)+ x(nT - (N - 2)T)+...+x(nT)]
ecg_m = conv(ecg_s ,ones(1 ,round(0.150*fs))/round(0.150*fs));
delay = delay + 15;

if gr
ax(6)=subplot(326);plot(ecg_m);axis tight;title('Averaged with 30 samples length,Black
noise,Green Adaptive Threshold,RED Sig Level,Red circles QRS adaptive threshold');
axis tight;
end

%% Fiducial Mark
% Note : a minimum distance of 40 samples is considered between each R wave
% since in physiological point of view no RR wave can occur in less than
% 200 msec distance
[pks,locs] = findpeaks(ecg_m,'MINPEAKDISTANCE',round(0.2*fs));

%% initialize the training phase (2 seconds of the signal) to determine the THR_SIG and
THR_NOISE
THR_SIG = max(ecg_m(1:2*fs))*1/3; % 0.25 of the max amplitude

THR_NOISE = mean(ecg_m(1:2*fs))*1/2; % 0.5 of the mean signal is considered to be noise
SIG_LEV= THR_SIG;
NOISE_LEV = THR_NOISE;

%% Initialize bandpath filter threshold(2 seconds of the bandpass signal)
THR_SIG1 = max(ecg_h(1:2*fs))*1/3; % 0.25 of the max amplitude
THR_NOISE1 = mean(ecg_h(1:2*fs))*1/2; %
SIG_LEV1 = THR_SIG1; % Signal level in Bandpassed filter
NOISE_LEV1 = THR_NOISE1; % Noise level in Bandpassed filter
%% Thresholding and online desicion rule

for i = 1 : length(pks)

 %% locate the corresponding peak in the filtered signal
 if locs(i)-round(0.150*fs)>= 1 && locs(i)<= length(ecg_h)
   [y_i x_i] = max(ecg_h(locs(i)-round(0.150*fs):locs(i)));
  else
   if i == 1
   [y_i x_i] = max(ecg_h(1:locs(i)));
   ser_back = 1;
```

```
   elseif locs(i)>= length(ecg_h)
   [y_i x_i] = max(ecg_h(locs(i)-round(0.150*fs):end));
   end

 end

%% update the heart_rate (Two heart rate means one the moste recent and the other selected)
if length(qrs_c) >= 9

 diffRR = diff(qrs_i(end-8:end)); %calculate RR interval
 mean_RR = mean(diffRR); % calculate the mean of 8 previous R waves interval
 comp =qrs_i(end)-qrs_i(end-1); %latest RR
 if comp <= 0.92*mean_RR || comp >= 1.16*mean_RR
  % lower down thresholds to detect better in MVI
   THR_SIG = 0.5*(THR_SIG);
   %THR_NOISE = 0.5*(THR_SIG);
   % lower down thresholds to detect better in Bandpass filtered
   THR_SIG1 = 0.5*(THR_SIG1);
   %THR_NOISE1 = 0.5*(THR_SIG1);

 else
  m_selected_RR = mean_RR; %the latest regular beats mean
 end

end

 %% calculate the mean of the last 8 R waves to make sure that QRS is not
 % missing(If no R detected , trigger a search back) 1.66*mean

 if m_selected_RR
  test_m = m_selected_RR; %if the regular RR availabe use it
 elseif mean_RR && m_selected_RR == 0
  test_m = mean_RR;
 else
  test_m = 0;
 end

if test_m
  if (locs(i) - qrs_i(end)) >= round(1.66*test_m)% it shows a QRS is missed
   [pks_temp,locs_temp] = max(ecg_m(qrs_i(end)+ round(0.200*fs):locs(i)-round(0.200*fs))); %
search back and locate the max in this interval
   locs_temp = qrs_i(end)+ round(0.200*fs) + locs_temp -1; %location

   if pks_temp > THR_NOISE
   qrs_c = [qrs_c pks_temp];
   qrs_i = [qrs_i locs_temp];

   % find the location in filtered sig
   if locs_temp <= length(ecg_h)
   [y_i_t x_i_t] = max(ecg_h(locs_temp-round(0.150*fs):locs_temp));
   else
   [y_i_t x_i_t] = max(ecg_h(locs_temp-round(0.150*fs):end));
   end
   % take care of bandpass signal threshold
   if y_i_t > THR_NOISE1

     qrs_i_raw = [qrs_i_raw locs_temp-round(0.150*fs)+ (x_i_t - 1)];% save index of bandpass
     qrs_amp_raw =[qrs_amp_raw y_i_t]; %save amplitude of bandpass
     SIG_LEV1 = 0.25*y_i_t + 0.75*SIG_LEV1; %when found with the second thres
   end

   not_nois = 1;
   SIG_LEV = 0.25*pks_temp + 0.75*SIG_LEV ; %when found with the second threshold
   end

  else
   not_nois = 0;

  end
end

%% find noise and QRS peaks
if pks(i) >= THR_SIG

    % if a QRS candidate occurs within 360ms of the previous QRS
    % ,the algorithm determines if its T wave or QRS
    if length(qrs_c) >= 3
```

```
      if (locs(i)-qrs_i(end)) <= round(0.3600*fs)
      Slope1 = mean(diff(ecg_m(locs(i)-round(0.075*fs):locs(i)))); %mean slope of the waveform
at that position
      Slope2 = mean(diff(ecg_m(qrs_i(end)-round(0.075*fs):qrs_i(end)))); %mean slope of
previous R wave
        if abs(Slope1) <= abs(0.5*(Slope2)) % slope less then 0.5 of previous R
         nois_c = [nois_c pks(i)];
         nois_i = [nois_i locs(i)];
         skip = 1; % T wave identification
         % adjust noise level in both filtered and
         % MVI
         NOISE_LEV1 = 0.125*y_i + 0.875*NOISE_LEV1;
         NOISE_LEV = 0.125*pks(i) + 0.875*NOISE_LEV;
        else
         skip = 0;
        end

      end
     end

  if skip == 0 % skip is 1 when a T wave is detected
  qrs_c = [qrs_c pks(i)];
  qrs_i = [qrs_i locs(i)];

  % bandpass filter check threshold
   if y_i >= THR_SIG1
      if ser_back
       qrs_i_raw = [qrs_i_raw x_i]; % save index of bandpass
      else
       qrs_i_raw = [qrs_i_raw locs(i)-round(0.150*fs)+ (x_i - 1)];% save index of bandpass
      end
       qrs_amp_raw =[qrs_amp_raw y_i];% save amplitude of bandpass
   SIG_LEV1 = 0.125*y_i + 0.875*SIG_LEV1;% adjust threshold for bandpass filtered sig
   end

  % adjust Signal level
  SIG_LEV = 0.125*pks(i) + 0.875*SIG_LEV ;
  end


 elseif THR_NOISE <= pks(i) && pks(i)<THR_SIG

   %adjust Noise level in filtered sig
   NOISE_LEV1 = 0.125*y_i + 0.875*NOISE_LEV1;
   %adjust Noise level in MVI
   NOISE_LEV = 0.125*pks(i) + 0.875*NOISE_LEV;



 elseif pks(i) < THR_NOISE
  nois_c = [nois_c pks(i)];
  nois_i = [nois_i locs(i)];

  % noise level in filtered signal
  NOISE_LEV1 = 0.125*y_i + 0.875*NOISE_LEV1;
  %end

   %adjust Noise level in MVI
  NOISE_LEV = 0.125*pks(i) + 0.875*NOISE_LEV;


 end

 %% adjust the threshold with SNR
 if NOISE_LEV ~= 0 || SIG_LEV ~= 0
  THR_SIG = NOISE_LEV + 0.25*(abs(SIG_LEV - NOISE_LEV));
  THR_NOISE = 0.5*(THR_SIG);
 end

 % adjust the threshold with SNR for bandpassed signal
 if NOISE_LEV1 ~= 0 || SIG_LEV1 ~= 0
  THR_SIG1 = NOISE_LEV1 + 0.25*(abs(SIG_LEV1 - NOISE_LEV1));
  THR_NOISE1 = 0.5*(THR_SIG1);
 end

% take a track of thresholds of smoothed signal
SIGL_buf = [SIGL_buf SIG_LEV];
```

```
NOISL_buf = [NOISL_buf NOISE_LEV];
THRS_buf = [THRS_buf THR_SIG];

% take a track of thresholds of filtered signal
SIGL_buf1 = [SIGL_buf1 SIG_LEV1];
NOISL_buf1 = [NOISL_buf1 NOISE_LEV1];
THRS_buf1 = [THRS_buf1 THR_SIG1];
 skip = 0; %reset parameters
 not_nois = 0; %reset parameters
 ser_back = 0; %reset bandpass param
end

if gr
hold on,scatter(qrs_i,qrs_c,'m');
hold on,plot(locs,NOISL_buf,'--k','LineWidth',2);
hold on,plot(locs,SIGL_buf,'--r','LineWidth',2);
hold on,plot(locs,THRS_buf,'--g','LineWidth',2);
end

%% overlay on the signals
if gr
figure,az(1)=subplot(311);plot(ecg_h);title('QRS on Filtered Signal');axis tight;
hold on,scatter(qrs_i_raw,qrs_amp_raw,'m');
hold on,plot(locs,NOISL_buf1,'LineWidth',2,'Linestyle','--','color','k');
hold on,plot(locs,SIGL_buf1,'LineWidth',2,'Linestyle','-.','color','r');
hold on,plot(locs,THRS_buf1,'LineWidth',2,'Linestyle','-.','color','g');
az(2)=subplot(312);plot(ecg_m);title('QRS on MVI signal and Noise level(black),Signal Level
(red) and Adaptive Threshold(green)');axis tight;
hold on,scatter(qrs_i,qrs_c,'m');
hold on,plot(locs,NOISL_buf,'LineWidth',2,'Linestyle','--','color','k');
hold on,plot(locs,SIGL_buf,'LineWidth',2,'Linestyle','-.','color','r');
hold on,plot(locs,THRS_buf,'LineWidth',2,'Linestyle','-.','color','g');
az(3)=subplot(313);plot(ecg-mean(ecg));title('Pulse train of the found QRS on ECG
signal');axis tight;
line(repmat(qrs_i_raw,[2 1]),repmat([min(ecg-mean(ecg))/2; max(ecg-
mean(ecg))/2],size(qrs_i_raw)),'LineWidth',2.5,'LineStyle','-.','Color','r');
linkaxes(az,'x');
zoom on;
end
end
```

1. Adjusted version of H. Sedghamiz (2014). Complete implementation of Pan-Tompkins algorithm, Linkoping university

```
function output = timeDomainHRV(ibi,win,xx)
%timeDomainHRV: calculates time-domain hrv of ibi interval series
% ibi = 2dim ibi array
% win = window size to use for sdnni (s)
% xx = value to use for NNx and pNNx (ms)

 t=ibi(:,1)-ibi(1,1);
 ibi=ibi(:,2);
 %check inputs
 ibi=ibi.*1000; %convert ibi to ms
 %assumes ibi units are seconds

%  if abs(range(ibi))<50 %assume ibi units are seconds
%    ibi=ibi.*1000; %convert ibi to ms
%  end
%  if abs(range(diff(t)))>50 %assume time unites are ms
%   t=t./1000; %convert time to s
%  end

%  if t<1000 %assume win units are (s)
%   t=t*1000; %convert to (ms)
%  end
```

```
 %calculate and round to nearest 1 decimal point
 output.max=round(max(ibi)*10)/10;
 output.min=round(min(ibi)*10)/10;
 output.mean=round(mean(ibi)*10)/10;
 output.median=round(median(ibi)*10)/10;
 output.SDNN=round(std(ibi)*10)/10;
 output.SDANN=round(SDANN(ibi,win*1000)*10)/10;
%output.SDANN=round(SDNNi(ibi,win*1000)*10)/10;
 [p n]=pNNx(ibi,xx);
 output.NNx=round(n*10)/10;
 output.pNNx=round(p*10)/10;
 output.RMSSD=round(RMSSD(ibi)*10)/10;
 output.SDNNi=round(SDNNi(ibi,win*1000)*10)/10;
%output.SDNNi=round(SDANN(ibi,win*1000)*10)/10;
 %heart rate
 hr=60./(ibi./1000);
 output.meanHR=round(mean(hr)*10)/10;
 output.sdHR=round(std(hr)*10)/10;

 %GEOMETRIC HRV

 %calculate number of bins to use in histogram
 dt=max(ibi)-min(ibi);
 binWidth=1/128*1000; %1/128 seconds. Reference: (1996) Heart rate
variability: standards of measurement, physiological interpretation and
clinical use.
 nBins=round(dt/binWidth);

 %temp
 nBins=32;

 output.HRVTi=round(hrvti(ibi,nBins)*10)/10;
 output.TINN=round(tinn(ibi,nBins)*10)/10;

end

function output = SDANN(ibi,t)
%SDANN: SDANN index is the std of all the mean NN intervals from each
%segment of lenght t.
 a=0;i1=1;
 tmp=zeros(ceil(sum(ibi)/t),1);
 for i2=1:length(ibi)
  if sum(ibi(i1:i2)) >= t
   a=a+1;
   tmp(a)=mean(ibi(i1:i2));
   i1=i2;
  end
 end
 output=std(tmp);
end

function output = SDNNi(ibi,t)
%SDNNi: SDNN index is the mean of all the standard deviations of
%NN (normal RR) intervals for all windows of lenght t.
 a=0;i1=1;
 tmp=zeros(ceil(sum(ibi)/t),1);
 for i2=1:length(ibi)
  if sum(ibi(i1:i2)) >= t
   a=a+1;
   tmp(a)=std(ibi(i1:i2));
```

```
    i1=i2;
   end
  end
  output=mean(tmp);
end

function [p n] = pNNx(ibi,x)
%pNNx: percentage of successive/adjacent NN intervals differing by x (ms)
or more
  differences=abs(diff(ibi)); %successive ibi diffs (ms)
  n=sum(differences>x);
  p=(n/length(differences))*100;
end

function output = RMSSD(ibi)
%RMSSD: root mean square of successive RR differences
  differences=abs(diff(ibi)); %successive ibi diffs
  output=sqrt(sum(differences.^2)/length(differences));
end

function output=hrvti(ibi,nbin)
%hrvti: HRV triangular index

  %calculate samples in bin (n) and x location of bins (xout)
  [n,xout]=hist(ibi,nbin);
  output=length(ibi)/max(n); %hrv ti

end

function output=tinn(ibi,nbin)
%tinn: triangular interpolation of NN interval histogram
%Reference: Standards of Measurement, Physiological Interpretation, and
Clinical Use
%    Circulation. 1996; 93(5):1043-1065.

  %calculate histogram of ibi using nbin bins
  [nout,xout]=hist(ibi,nbin);

  D=nout;
  peaki=find(D==max(D));
  if length(peaki)>1
   peaki=round(mean(peaki));
  end

  i=1;
  d=zeros((peaki-1)*(nbin-peaki),3);

  for m=(peaki-1):-1:1
   for n=(peaki+1):nbin
    %define triangle that fits the histogram
    q=zeros(1,length(D));
    q(1:m)=0;
    q(n:end)=0;
    q(m:peaki)=linspace(0,D(peaki),peaki-m+1);
    q(peaki:n)=linspace(D(peaki),0,n-peaki+1);

    %integrate squared difference
    d(i,1)=trapz((D-q).^2);
    d(i,2:3)=[m,n];%d(i,2:3)=[m,n];
```

```
   %plot(D); hold on; plot(q,'r'); hold off;
   %title(['d^2 = ' num2str(d(i,1))])
   i=i+1;
  end
 end
 %find where minimum square diff occured
 i=find(d(:,1)==min(d(:,1)));
 r = isempty(i);
 if r==1;
  output = nan;
 else r=0;
 i=i(1); %make sure there is only one choise
 m=d(i,2); n=d(i,3);
 %calculate TINN in (ms)
 output=abs(xout(n)-xout(m));
 end


end
```

R

```
# Mixed effects model

#install and load packages
install.packages('Matrix')
install.packages('lme4')
install.packages('lattice')
install.packages('foreign')
install.packages('rJava')
install.packages('xlsxjars')
install.packages('xlsx')
install.packages('JMbayes')
install.packages('MASS')
install.packages('nlme')
install.packages('splines')
install.packages('survival')

library(Matrix)
library(lme4)
library(lattice)
library(foreign) #spss
library(rJava)
library(xlsxjars)
library(xlsx)

# read in your data
am.data <- read.xlsx(file.choose(), 1) # opens a pop-up window, allowing you to select your dataset

# check data
head(am.data)
sapply(am.data, data.class)

# changing your 1's and 2's in Response to 0's and 1's:
am.data$Response_datab[am.data$Response_datab == 1] <- 0
am.data$Response_datab[am.data$Response_datab == 2] <- 1
table(am.data$Response_datab)

# similarly for gender
am.data$Gender[am.data$Gender == 1] <- 0
am.data$Gender[am.data$Gender == 2] <- 1
table(am.data$Gender)

# changing pat number,gender, CRP and response(?) to factors
am.data$Pnmbr <-as.factor(am.data$Pnmbr)
am.data$Gender <-as.factor(am.data$Gender)
am.data$CRP_datab <- as.factor(am.data$CRP_datab)
am.data$CRP_adj <- as.factor(am.data$CRP_adj)

#am.data$Response_datab <- as.numeric(am.data$Response_datab)

sapply(am.data, data.class)
levels(am.data$Gender)

# time_adj + GA_adj + BW_datab + Gender
# Weight_datab + CRP_datab + CRP_adj
# SDRR + SDARR + RR50 +pRR50 + RMSSD + SDSD + SDHR + HRVTi + TIRR
# SD1 + SD2 + Brady
```

```
#scaling
df <-
data.frame(am.data$time_adj,am.data$GA_a
dj,am.data$BW_datab,am.data$Weight_data
b,am.data$Median, am.data$SDRR,
am.data$SDARR,am.data$RR50,am.data$pRR
50,am.data$RMSSD,am.data$SDSD,am.data$S
DHR,am.data$HRVTi,am.data$TIRR,am.data$S
D1,am.data$SD2,am.data$Brady)

dfs<-df
dfs<-scale(dfs, center= TRUE, scale=TRUE) #all
the centring in R
am.data2 <-
data.frame(am.data$Pnmbr,am.data$Gender,
am.data$CRP_adj, am.data$CRP_datab,
am.data$Response_datab, dfs)

install.packages("splines")
library(splines)
spline2=ns(am.data2$am.data.SDARR, df=6)


final1 <- glmer(am.data.Response_datab ~
am.data.time_adj + am.data.CRP_datab +
am.data.Brady + am.data.Median +
spline2[,4:6] + (1 |am.data.Brady), data =
am.data2, family = binomial)
summary(final1)

final12 <- glmer(am.data.Response_datab ~
am.data.time_adj + am.data.CRP_datab +
am.data.Brady + am.data.Median +
spline1[,4:5] + (1 |am.data.Pnmbr), data =
am.data2, family = binomial)
summary(final12)

final2 <-glmer(am.data.Response_datab ~
am.data.time_adj + am.data.CRP_datab +
am.data.Brady + am.data.Median +
am.data.SDARR + (1 |am.data.Pnmbr), data =
am.data2, family = binomial)
summary(final2)

final3 <- glmer(am.data.Response_datab ~
am.data.time_adj + am.data.GA_adj +
am.data.Brady + am.data.Median +
spline2[,4:6] + (1 |am.data.Brady), data =
am.data2, family = binomial)
summary(final3)

Response<-
final@frame$am.data.Response_datab
Yhat = final@resp$eta
Y= exp(Yhat)/(1+exp(Yhat))

x= data.frame(final3@frame)

write.xlsx(x,'D:/R/Outcome/final/xfinal3.xlsx')

resp= final3@resp$eta
write.xlsx(resp,'D:/R/outcome/final/final3.xlsx
')

# choose a threshold for dichotomizing
according to predicted probability
thresh <- 0.5
Predicted <- cut(Y, breaks=c(-Inf, thresh, Inf),
labels=c("0", "1"))
# contingency table and marginal sums
cTab <- table(Response, Predicted)
addmargins(cTab)
#percentage correct
sum(diag(cTab)) / sum(cTab)

#R2
r2.corr.mer <- function(m) {
 lmfit <- lm(model.response(model.frame(m))
~ fitted(m))
 summary(lmfit)$r.squared
}

r2.corr.mer(final1)

basic <- glmer(am.data.Response_datab ~
am.data.time_adj + am.data.GA_adj +
am.data.CRP_datab + am.data.Brady +
am.data.Median + (1 |am.data.Pnmbr), data =
am.data2, family = binomial)
summary(basic)

install.packages('polynom')
library(polynom)

poly <- (am.data2$am.data.SDARR)
poly2 <- poly+
I((am.data2$am.data.SDARR)^2)
poly3 <-poly+ I((am.data2$am.data.SDARR)^2)
+ I((am.data2$am.data.SDARR)^3)
Poly3s <- scale(poly3[,3])

Poly = bs(am.data$SDARR,3) #poly spine
Poly2= Poly[,1] + Poly[,2] + Poly[,3]
```

```
model <- glmer(am.data.Response_datab ~
am.data.time_adj + am.data.CRP_datab +
am.data.Brady + am.data.Median + (1
|am.data.Pnmbr) + (1 |am.data.Brady), data =
am.data2, family = binomial)
summary(model)

spline= ns(am.data2$am.data.SDARR, df=5) #,
knots = seq(-1, 1, by=2))
spline2=ns(am.data2$am.data.SDARR, df=6)
sdarr = spline[,1]*am.data2$am.data.SDARR +
spline[,2]*am.data2$am.data.SDARR +
spline[,3]*am.data2$am.data.SDARR +
spline[,4]*am.data2$am.data.SDARR

Test <- glmer(am.data.Response_datab~
spline2 + (1|am.data.Pnmbr), data=am.data2,
family=binomial)
summary(Test)

Rand2 <-glmer(am.data.Response_datab ~
am.data.time_adj + am.data.CRP_datab +
am.data.Brady + am.data.Median +
(am.data.Pnmbr |am.data.Brady), data =
am.data2, family = binomial)
summary(Rand2)

plot(sdarr)
points(am.data2$am.data.SDARR,col='red')

test <-glmer(am.data.Response_datab ~ spline
+ (1|am.data.Pnmbr), data=am.data2,
family=binomial)

model5<-glmer(am.data.Response_datab ~
am.data.time_adj + am.data.GA_adj +
am.data.CRP_datab + + am.data.Brady + (1
|am.data.Pnmbr) + (1|am.data.Brady),
data=am.data2, family=binomial)
summary(model5)

#control = glmerControl(optimizer =
c("bobyqa", "Nelder_Mead"), restart_edge =
FALSE, boundary.tol = 1e-5, calc.derivs=TRUE,
use.last.params=FALSE, sparseX = FALSE,
tolPwrss=1e-7, compDev=TRUE,
nAGQ0initStep=TRUE))
anova(model)

install.packages('ggplot2')
library(ggplot2)

# classification table option 1
Response<-
Rand@frame$am.data.Response_datab
Yhat2 <- fitted(model)
Yhat = Rand@resp$eta
Y= exp(Yhat)/(1+exp(Yhat))

plot(Y,residuals(model))

# test relationship
std =sd(am.data$SDARR,na.rm=TRUE)
avg = mean(am.data$SDARR,na.rm=TRUE)
x =(model@frame$am.data.SDARR)*std+avg
#back transform HRV
plot(x,Y)

# choose a threshold for dichotomizing
according to predicted probability
thresh <- 0.5
Predicted <- cut(Y, breaks=c(-Inf, thresh, Inf),
labels=c("0", "1"))
# contingency table and marginal sums
cTab <- table(Response, Predicted)
addmargins(cTab)
#percentage correct
sum(diag(cTab)) / sum(cTab)

# histogram of residuals
qplot(residuals(model))

#histogram of random effects
rand = ranef(model)
qplot(model@pp$delu)
plot(Predicted,residuals(model))
plot(model@frame$SDARR, residuals(model))
#save data
x= data.frame(Rand@frame)
write.xlsx(x,'D:/R/Outcome/xBest.xlsx')
resp= Rand@resp$eta
write.xlsx(resp,'D:/R/outcome/Best.xlsx')
```