Aidan O’Keeffe
January 8, 2023

Exploratory data analysis of heart rate variability

Neonatal sepsis is one of the leading causes of death for preterm neonates. In recent years, efforts have been made to predict the onset of sepsis using vital sign measurements. The state of the art in this area of research is the HeRO score (source), but there remains room for improvement, with other researchers making attempts to solve this problem (source in email, source in my files).

This essay communicates an effort to predict the onset of sepsis in preterm neonates using only statistics of heart rate variability. The dataset is introduced, the statistics are defined, the development of the pipeline is explained in depth, and the results are shared. The results were inconclusive, so possible modifications to the project are discussed that may yield better results.

Ask

The driving question of this project was: Can the onset of sepsis in preterm neonates be predicted using heart rate variability?

[Define HRV]

[Go through research justifying this hypothesis]. (including the one saying that frequency domain measures are not good for infants)

Prepare

As this was an exploratory analysis, a small dataset of seven infants was used. Three infants (1,5,7 [CHECK]) developed sepsis, while the others did not. Time series of ECG, respiration, and SPO2 were pulled. If an infant developed sepsis, than the time in the medical record at which a positive blood culture was reported occuped the center of the time series, and then approximately one week of data before and after was included. The non-septic infants were infants close in age to the septic infants that had “relatively uncomplicated stays”. Two weeks of data were pulled, with the center corresponding to the sepsis-time of the septic infant in terms of post-gestational age. Deidentified too.

summary data of infant ages and stuff

The files were very large, so Dr. Chang broke them into 5 smaller files that could be handled by Python one at a time. Later on, it was found that two infants saw a major gap between the files, and this was resolved by shifting the first part to match the second.

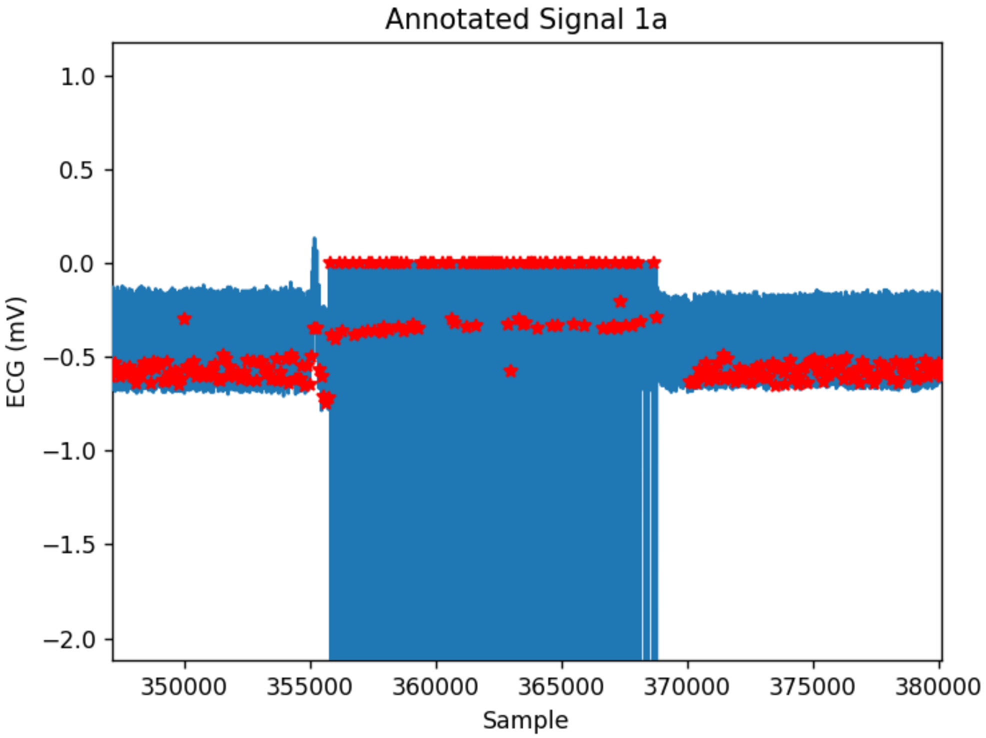
[visuals of the file pieces, particularly infants 2 and 3]

ECG is a signal of voltage vs. time, whereas the RR interval data relies on the location of the QRS complex. As such, a QRS complex detector was needed. Based on work by (QRS paper) as well as practical considerations of time and availability, we went with XQRS.

XQRS takes an argument of the sampling frequency, which we estimated with a histogram of frequencies. As you can see, they are almost universally 250 Hz.
[visual of frequency distribution histogram]

There were some missing values in the data, and sometimes a different ECG channel was used, but all ECG signals were found to be in the same range of voltages and it was rare that more than one channel was recorded at once, so we decided to combine the signals together.

A strange feature of the data was the presence of very large spikes in the data, including artefacts which we termed “troughs”.
[picture of troughs and spikes, like from infant 1]



The cause of troughs was not ascertained. Both troughs and non-trough spikes in the data disrupted QRS detection, but we found that flat regions in the signal did not disrupt QRS detection, so we decided to replace troughs with flatlines, and then throw out the artificially large RR intervals. (Troughs were usually between 20 to 30 seconds in length).

(Graphics, see slide CAN XQRS RECOGNIZE LACK OF PEAKS IN FLAT SEGMENT?)

There was a lot of noise in the data (ECG signals are always noisy), but our first examinations indicated that XQRS could handle this, so we did not apply any filters to denoise the data.
[image of QRS complexes being detected in a segment with noise]

Finally, XQRS sometimes missed a number of QRS complexes in a row, which created unrealistically large RR intervals. We decided to try and recover some single intervals from these so-called multiple intervals (although as we'll discuss later, there is room for improvement in this respect).
[Picture of multiple intervals]

Process
The data cleaning pipeline can be summarized as follows:

- Load in and complete ECG signal
    - Description of the process
    - **[Visual of what it did]**
- Erase troughs and spikes
    - Description of the algorithm
    - **[Visual of different thresholds]**
- Run XQRS
    - Description of breaking signal into chunks for accelerated detection
- Calculate RR intervals
- Clean RR intervals
    - **[Visual with multiple and single intervals overlaid]**

Make detailed descriptions for each of these things here.

Analyze
The statistics were calculated. We looked at different time windows.

- Can put the formulas for the statistics here

Share
Put visualizations in here

- **[Time series of the statistics]**

Act
The visuals show no indication of anomalies in the time series approaching sepsis. As such, we are left to suggest improvements to the pipeline.

- Determine the cause of the trough artefacts and fix it.
- Denoise the signal: may help with the missed QRS complexes
- Consider other QRS detectors
- Use a more sophisticated multiple-interval breaker; the current one artificially deflates