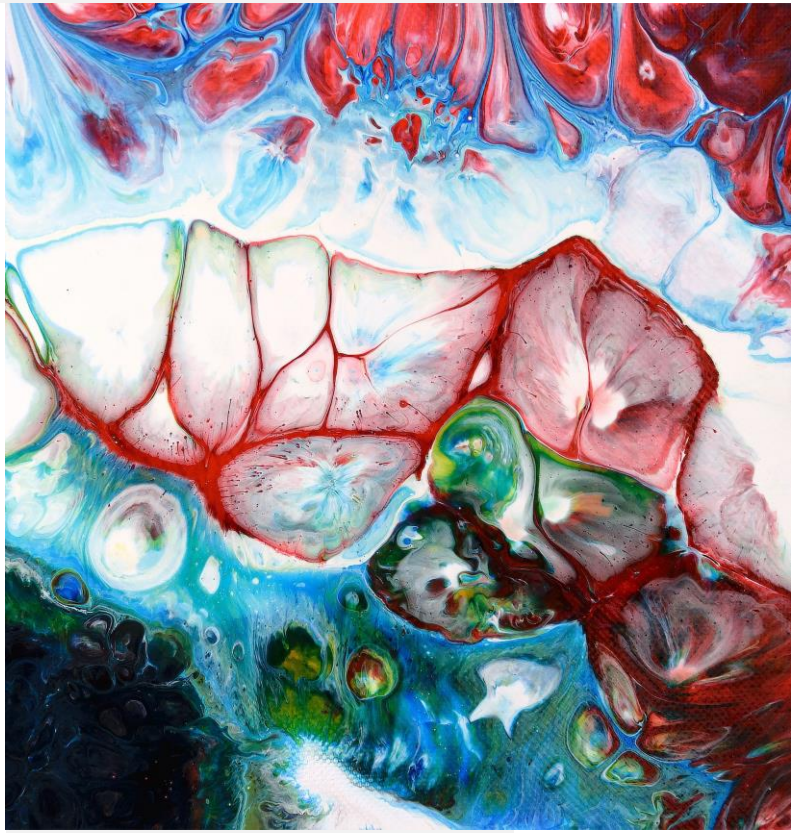
Replicate Intl
Michael Coggins, PhD

BIOMETRIC DATA SCIENCE HRV PROJECT

Current Progress

CONTENTS



PURPOSE

Why? But….why…(or why not)?

I want to understand us and data can help

- For years, I collected biometric data – mostly performance, but some relevant to basic medical health and wellness – as part of my own curiosity regarding client success (or not) as a coach and trainer
  - In 2016-2019, I created a small pilot study to look at wearable technology as a potentially unbiased data collection/analysis engine to predict improvements in a person's strength, lean muscle mass, bone density and fat loss based on different exercise and diet interventions
- During the pandemic, as gyms closed, I took online courses in Data Science to broaden my analytical techniques as most of what I used in previous research 'lives' were basic parametric and non-parametric methods
- In 2022, I thought…why not put these together?
  - But, a professor friend, suggested I use some publicly available data first, as way to have a comparison to other researchers' data analytics and conclusions.

ABOUT ME

Entrepreneur (Replicate International)
Business Analyst (Credit Suisse, Campbell Alliance)
Researcher (Yale, Cornell)
Military Engineer (USAF)

Data Scientist (in-progress)

- Started computational side projects after career change from equity research to athletic coach / trainer
- Started Coursera courses during Covid lockdown (IBM Data Science cert amongst other classes)

LinkedIn

LINEAR REGRESSION CURRENT CONCLUSION

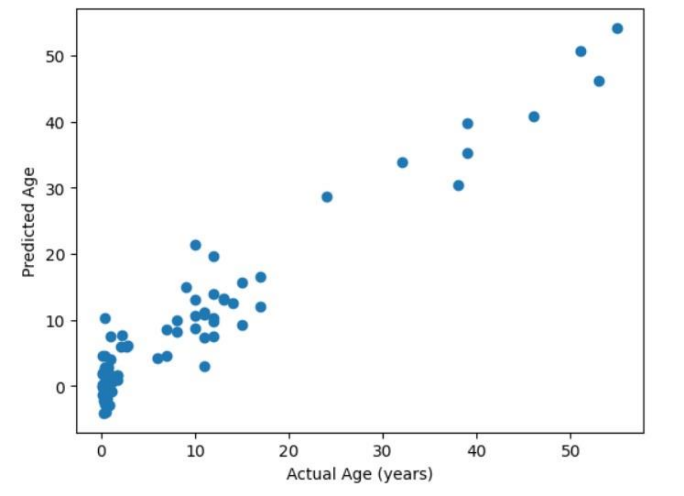
Naive linear regression using 32 features of 91 Subjects with full metadata (age) showed IBI/HR not the most predictive features

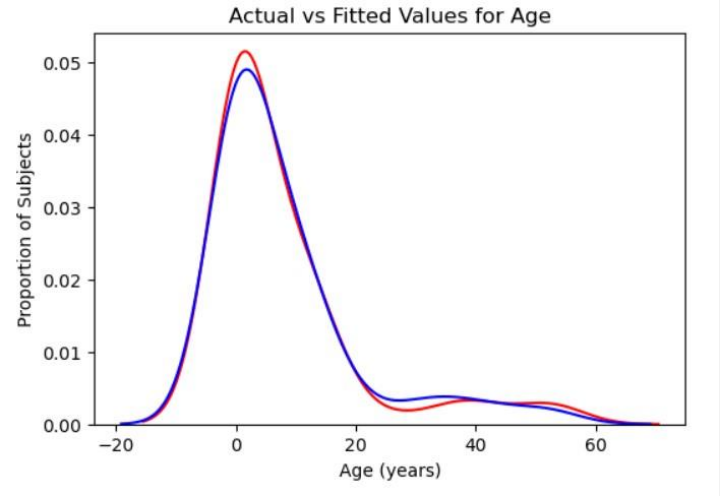
Initially I chose a supervised multiple linear regression model based on age as the dependent variable:

- Initial model will use all 91 subjects and is easy to understand and has significant literature on model benefits and limitations
- Full code uploaded to GitHub here: https://github.com/mrc7mrc49/HRV-Public-DB-Irirzun

Relative ranking of most important features (Top 6):

- No Change Count Z[SDNN, 512]", "CoV Avg IBI", "High Count Z[SDNN]256 >4 Z","CoV C[rMSSD]512", "High Count C[rMSSD]512 >4 Z"
- Avg IBI and Avg HR not among Top 6 is somewhat of a surprise given the strong, consistent relationship of infants possessing Avg HRs or ~2x adults





jupyter Initial MLR Calc Dataset3 Last Checkpoint: 3 minutes ago

File Edit View Run Kernel Settings Help

Code

Initial EDA

[3]:
```
#Initial Multinomial Regression (on Age) for entire dataset with complete metadata (Age)
#Importing of initial libraries for data manipulation and visualization
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
```

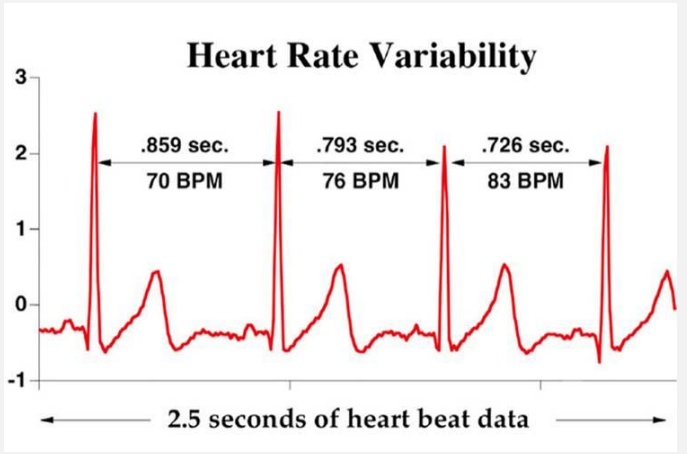
[6]:
```
#Download into python dataset (from Excel worksheets) and conversion into dataframes
path = 'C:\\Users\\mikec\\OneDrive\\Composite Exp Data for GraphPad 9.15.24a.xlsx'

#df will be dataframe of all 92 subjects with Age and calculated metrics
df = pd.read_excel(path, sheet_name='Preworked before code')
df = df.drop(['Subject', 'Gender'], axis=1)

#df1 will be dataframe of 4 subjects without Age metadata from author's zipfile
df1 = pd.read_excel(path, sheet_name='Data for Age Guess')
df1 = df1.drop(['Subject', 'Age (years)', 'Gender'], axis=1)

#dfy will be single column dataframe of Age for linear regression statistics
dfy = df['Age (years)']

#Importing useful libraries for model fitting and model accuracy/relevance statistics
import seaborn as sns
from sklearn import preprocessing, svm
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error

#Data wrangling and removing of dumm
df1_new = pd.get_dummies(df1, dummy_na=True, dtype=int)
```

INITIAL IDEATION / PROBLEM DEFINE

Initial idea and first look at a public database

IDEATION, PROBLEM DEFINITION OR PURPOSE

Can we learn something new from old data



- There exist various publicly-accessible databases of human biometric data with potential for preventative health decisions, or performance-based tailoring of performance/recovery programs
- There appears to be untapped potential for data to be analyzed by alternative methods including ML programs
- Why not run some of the public data through alternative methods to look for possible uncovered predictors/relevant metrics?
- EKG data represent a simple test case with significant years of study (and possibly a number of databases), yet little consensus about predictive value

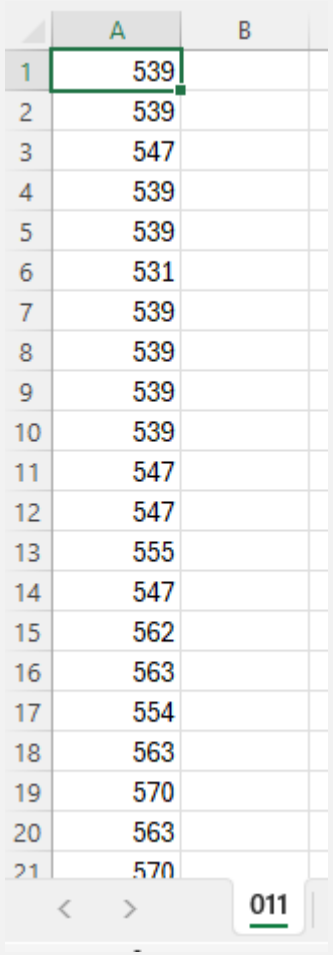
Can we identify new predictive metrics from EKG data?

INITIAL DATA VISUALIZATION

First, we sought to look at the data to get a feel for the potential for different classes based on age (sex)

Notes:

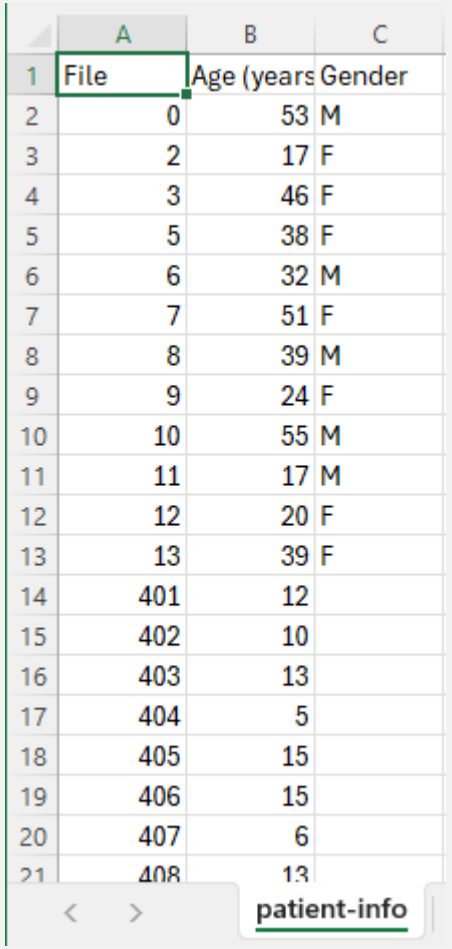
- Inter-Beat Intervals (IBIs, recorded in milliseconds per heart beat) are the reported data. We also converted IBIs to Instantaneous Heart Rates (iHRs, beats/min)
  - Due to relationship between IBIs and iHRs: iHR = 60 (seconds per min) / IBI (in seconds)
- Database included 24-hour Holter monitor recordings of different aged subects:
  - 56 Infants aged (1 month – 3 years 11 months)
  - 11 Adults aged (17 – 53)
  - 24 Adolescents (4 – 15)
  - 4 Unknown



	A
1	539
2	539
3	547
4	539
5	539
6	531
7	539
8	539
9	539
10	539
11	547
12	547
13	555
14	547
15	562
16	563
17	554
18	563
19	570
20	563
21	570

011

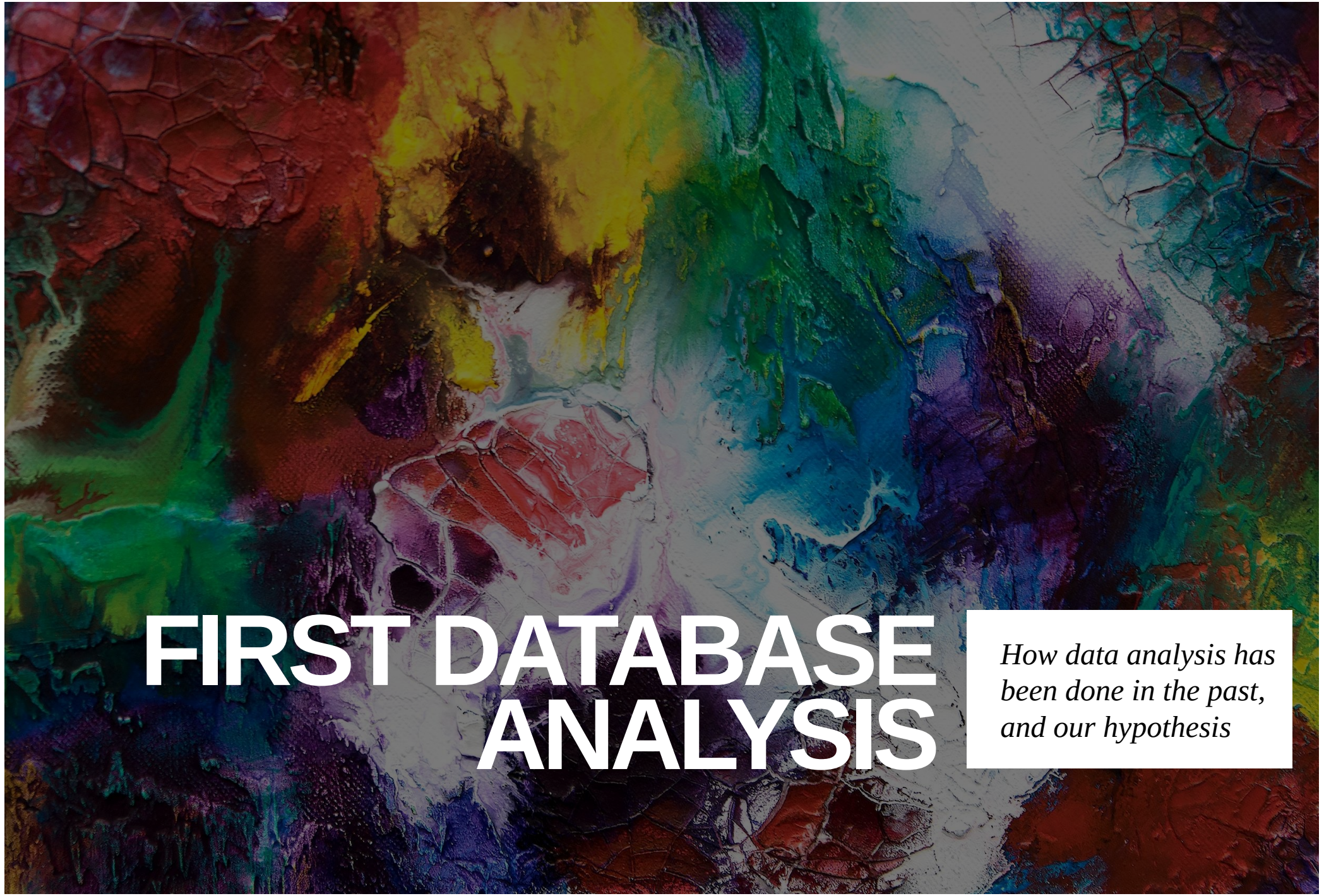
Single Subject IBIs



	A	B	C
1	File	Age (years	Gender
2	0	53	M
3	2	17	F
4	3	46	F
5	5	38	F
6	6	32	M
7	7	51	F
8	8	39	M
9	9	24	F
10	10	55	M
11	11	17	M
12	12	20	F
13	13	39	F
14	401	12	
15	402	10	
16	403	13	
17	404	5	
18	405	15	
19	406	15	
20	407	6	
21	408	13	

patient-info

Metadata



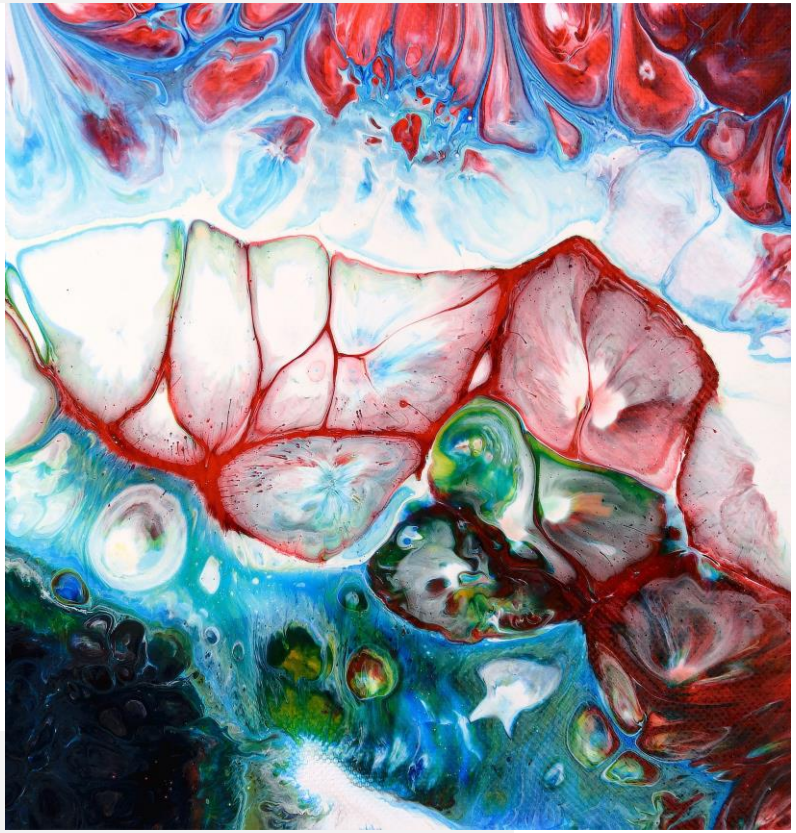
FIRST DATABASE ANALYSIS

How data analysis has been done in the past, and our hypothesis

TIME DOMAIN ANALYSIS – IN THE PAST

Two main measurements:

- SDNN – Standard Deviation (from mean over chosen time period)
  - Evaluates deviations from each beat to the mean
  - Corresponds to slower time course, lower frequency changes (*possibly* parasympathetic)
- rMSSD – Root Mean Square Standard Deviation (from immediately previous IBI)
  - Evaluates deviation from immediately previous recorded heart beat
  - Corresponds to faster time course, higher frequency changes (sympathetic nervous system and respiration, *possibly*)



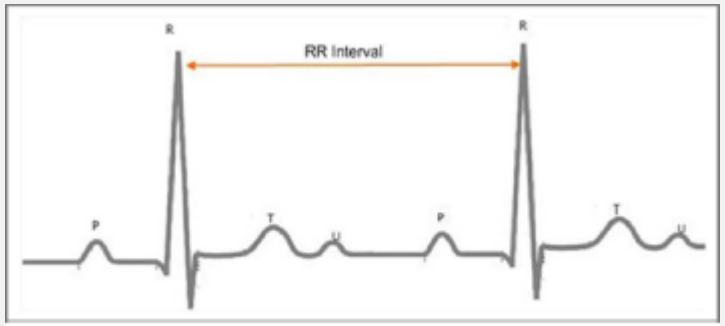


Fig. 1 RR interval used for HRV measurement

Murukesan et al., 2013

HYPOTHESIS

Versions of normalizing TD data to account for variable expected values given different average HRs

My idea for analysis creates values similar to Z Scores

- Z = (x - μ) / σ as a new version of a 'normalized' comparator given a pre-existing IBI (or HR)
  - I chose to use a simple version of an 'expected noise' or 'expected jitter' appropriate for SDNN (around the time period mean IBI) and rMSSD (immediately previous IBI)
- Shown is a single subject with created Z scores (for SDNN comparison metric)

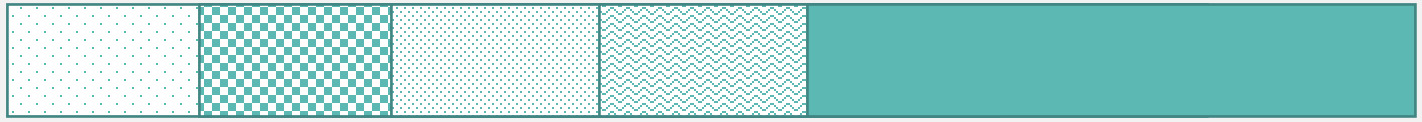
Trace	aPR	IBI(ms)	((n) - (n-1)	E[SDNN] 512	Z[SDNN]
461	130.1518	0	0	1.735612767	0
453	132.4503	8	64	1.705362594	13.62410465
438	136.9863	15	225	1.657180472	64.82695304
437	137.2998	1	1	1.631944205	0.149950134
461	130.1518	24	576	1.651921188	183.0213445
461	130.1518	0	0	1.665443658	1
469	127.9318	8	64	1.683254414	14.08274278
469	127.9318	0	0	1.69680065	1
469	127.9318	0	0	1.707449986	1
484	123.9669	15	225	1.726485454	59.10796569
469	127.9318	15	225	1.732668954	58.63216679
484	123.9669	15	225	1.746710209	57.57125161
469	127.9318	15	225	1.750456973	57.2926512
476	126.0504	7	49	1.757323967	8.900253548
477	125.7862	1	1	1.763789356	0.187522666
461	130.1518	16	256	1.762008285	65.29524855
453	132.4503	8	64	1.756761066	12.6297431
453	132.4503	0	0	1.752116523	1
453	132.4503	0	0	1.747976474	1
446	134.5291	7	49	1.741473361	9.117892873
453	132.4503	7	49	1.738264039	9.162766311
445	134.8315	8	64	1.732471725	13.08760844
453	132.4503	8	64	1.729952747	13.13630267

CREATED COMPOSITE DATA

Given different size datasets (row length) and disparities in known attributes, what might better represent individual data

Created a sliding windows of time periods (now many HRV ‘epochs’)

- I chose two time periods of 512 IBIs (513 heart beats) and 256 IBIs (257 IBIs)
  - Given infants have ~2x as many heart beats per time period, this allows for a comparison of similar #IBIs or similar periods of time
  - 512 IBIs is ~6-10 minutes for adults, ~3.5-5 minutes for infants
  - Now there are thousands of short time overlapping epochs (short windows) to potentially better capture periods of stabilized physiological state



T=0, IBI 1 | T=x1, IBI 2 | T=x2, IBI 3 | T=x3, IBI 4 | T=x4, IBI 5

Subect	Age (years)	Gender	Avg IBI (ms, 512)	CoV Avg IBI	Avg iPR (bpm, 512)	CoV Avg iPR	No Change Count Z[SDNN, 512]	Low Count Z<2 [SDNN, 512]	High Count Z>4 [SDNN, 512]	Avg C[SDNN, 512]	CoV C[SDNN, 512]	Low Count <2 C[SDNN]	High Count >4 C[SDNN]
0	53	M	943.3	0.17	66.0	0.21	0.04	0.16	0.63	7.39	0.44	0.00	0.94
2	17	F	841.5	0.19	74.3	0.21	0.03	0.09	0.76	10.88	0.27	0.00	1.00
3	46	F	711.4	0.11	85.4	0.12	0.08	0.16	0.57	6.76	0.74	0.00	0.83
5	38	F	769.4	0.17	82.3	0.18	0.06	0.14	0.62	7.82	0.45	0.00	0.96
6	32	M	858.4	0.17	70.2	0.17	0.07	0.20	0.48	5.04	0.33	0.00	0.70

CREATED COMPOSITE DATA

Given different size datasets (row length) and disparities in known attributes, what might better represent individual data

Created a group of simple statistical data as comparators

- Sample descriptors such as mean, coefficient of variation, count, etc. help minimize ectopic and missing beats
  - Some metrics not shown below

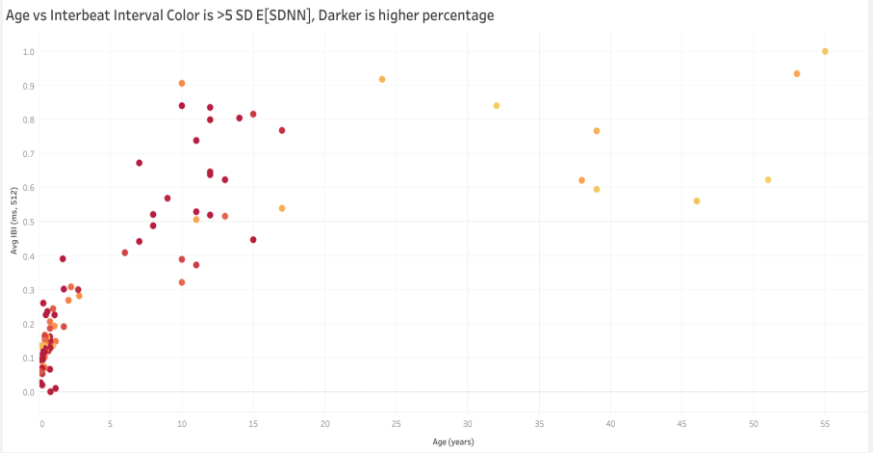
Subect	Age (years)	Gender	Avg IBI (ms, 512)	CoV Avg IBI	Avg iPR (bpm, 512)	CoV Avg iPR	No Change Count Z[SDNN, 512]	Low Count Z<2 [SDNN, 512]	High Count Z>4 [SDNN, 512]	Avg C[SDNN, 512]	CoV C[SDNN, 512]	Low Count <2 C[SDNN]	High Count >4 C[SDNN]
0	53	M	943.3	0.17	66.0	0.21	0.04	0.16	0.63	7.39	0.44	0.00	0.94
2	17	F	841.5	0.19	74.3	0.21	0.03	0.09	0.76	10.88	0.27	0.00	1.00
3	46	F	711.4	0.11	85.4	0.12	0.08	0.16	0.57	6.76	0.74	0.00	0.83
5	38	F	769.4	0.17	82.3	0.18	0.06	0.14	0.62	7.82	0.45	0.00	0.96
6	32	M	858.4	0.17	70.2	0.17	0.07	0.20	0.48	5.04	0.33	0.00	0.70

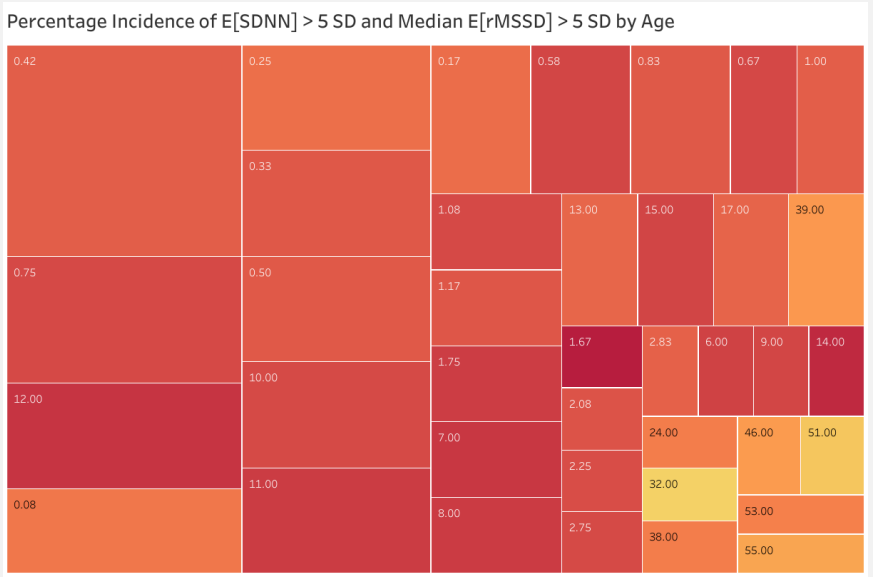
TIME DOMAIN COMPOSITE DATA VISUALIZATION

Do any of the composite metrics show variation by age or possible 2-3 groupings (infant, adolescent, adult)

Created metric data presented (used for AI/ML processing)

- Non-normalized data showed a number of expected trends (and are also visible after normalization)
  - Infants have significantly smaller IBIs / higher iHRs: top row
- Normalized data showed potential for some metrics to group ages into 2 or 3 groups by our created 'expectation' Z metric
  - Low Count represents the percentage of time bin windows that had IBIs < X (initially we chose X to be 2 representing a change of < 3 standard deviations)
  - High Count represents the percentage of time bin windows that had IBIs > X (initially we chose X to be 4 representing a change of > 5 standard deviations)





LINEAR REGRESSION CURRENT CONCLUSION

Naive linear regression using 32 features of 91 Subjects with full metadata (age) showed IBI/HR not the most predictive features

Relative ranking of most important features (Top 6):

- "No Change Count Z[SDNN, 512]", "CoV Avg IBI", "High Count Z[SDNN]256 >4 Z", "CoV C[rMSSD]512", "High Count C[rMSSD]512 >4 Z"
- Avg IBI and Avg HR not among Top 6 is somewhat of a surprise given the strong, consistent relationship of infants possessing Avg HRs or ~2x adults

Next steps:

- Gen AI for increasing Adult/Adolescent data
- Removal of infants <4 years old to account for possible upright bias