Data Analytics

Course: 18-787

Patrick McSharry
patrick@mcsharry.net
www.mcsharry.net
Twitter: @patrickmcsharry

Spring 2023

ICT Center of Excellence
Carnegie Mellon University

Data Analytics

WEEK 5A

Course outline

Week	Lecture A	Lecture B
1	Data Analytics	Weather forecasting
2	Renewable energy	Wind energy
3	Solar energy	Demand forecasting
4	Risk	Extreme events
5	Health	Biomedicine
6	Early warning systems	Economic forecasting

Today's Lecture

No.	Activity	Description	Time
1	Challenge	Data-driven medicine	10
2	Discussion	History of healthcare	10
3	Case study	Growth charts	10
4	Analysis	Electronic records	20
5	Demo	Rwandan growth chart	20
6	Q&A	Questions and feedback	10

Visiting a doctor

- Enter into the doctor's office
- In the UK, GP's allocate about 5min per person
- You offer up your complaints
- Doctor measures your heart rate, blood pressure, height, weight etc
- Discussion of the symptoms
- Diagnosis is heavily reliant on human judgment
- If there is uncertainty, further assessments may be required

Why is medicine not data-driven?

- Typical diagnosis by doctor is based on experience and gut-feeling rather than empirical evidence
- Qualitative information (discussion, physical appearance of the patient, previous conditions) is viewed as being most important
- Doctors are trained to perform without complete data

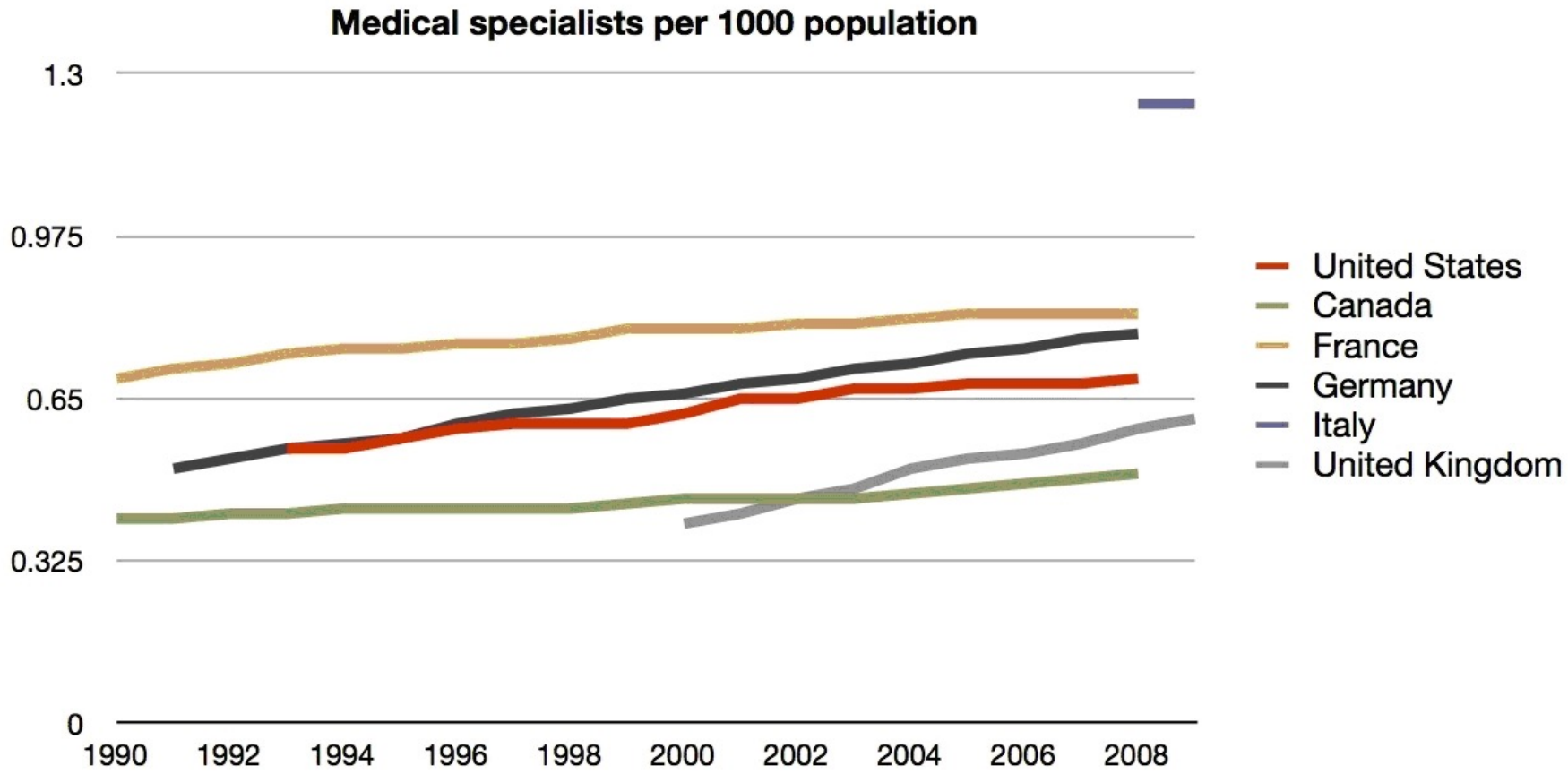
History of Healthcare

- First medical school was Schola Medica Salernitana in Italy around 1220.
- Students spent eight years training for medical degree.
- Emphasis on theory rather than clinical work.
- The 19th century was the first time that the scientific method was applied to medical research.

Scientific Method

- **Formulation of a question:** Posing a previously unanswerable proposition.
- **Hypothesis:** A statement of the answer to the question, without worrying about whether the statement is accurate or not.
- **Prediction:** A contemplative view of what the data will prove the answer to be.
- **Testing:** Extensive trials to assess variables and factors related to the hypothesis.
- **Analysis:** Compilation and reconciliation of the data derived from testing.

Specialists



Rise of specialists means more opinion-based medicine.

Bad Science – Ben Goldacre

- **Nostradamus:** Many people believe he predicted the future, but they ignore the hundreds of other people that predicted things that did not happen. They pay attention only to the person who was right (perhaps out of luck), as opposed to all those who were wrong.
- **Medical trials:** A publication will cite a study showing that a drug cured a certain medical condition in eight out of ten people tested. However, that publication conveniently ignores ten other tests, of the same drug and same condition, where only two out of ten people were cured.

See: *Bad Science by Ben Goldacre* (Harper Perennial, 2008),

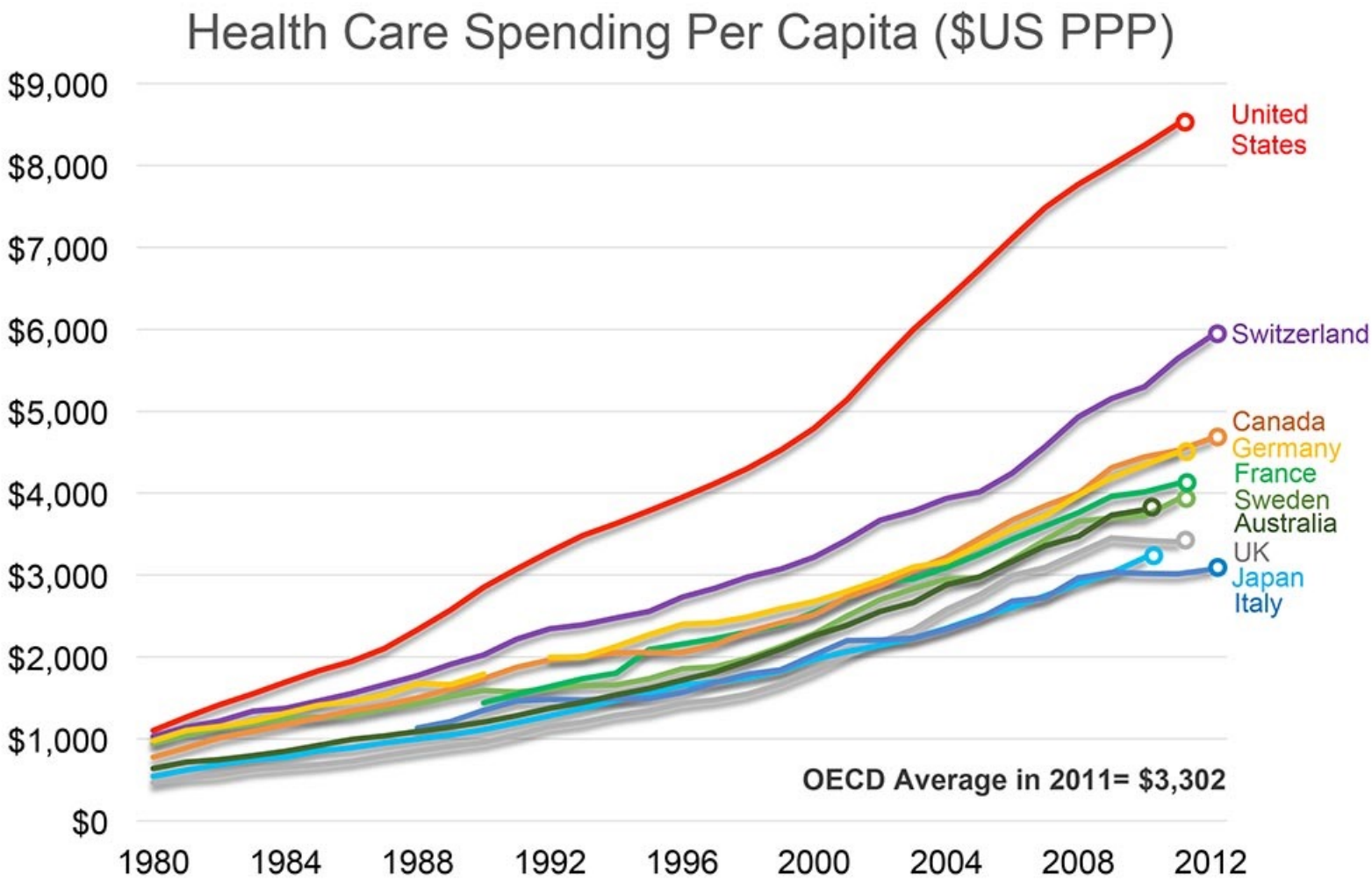
Doctors are human

- Doctors, like everyone else, have cognitive limitations.
- Some are naturally smarter than others or have deeper knowledge about a particular topic.
- The latter leads to biases in how they think, act, and prescribe.
- Doctors often decide on a patient diagnosis in the first 5 to 10 seconds of the observation.
- Essentially, they base their diagnosis on a gut reaction to the symptoms that they can see or are described to them.

Future Medical Schools

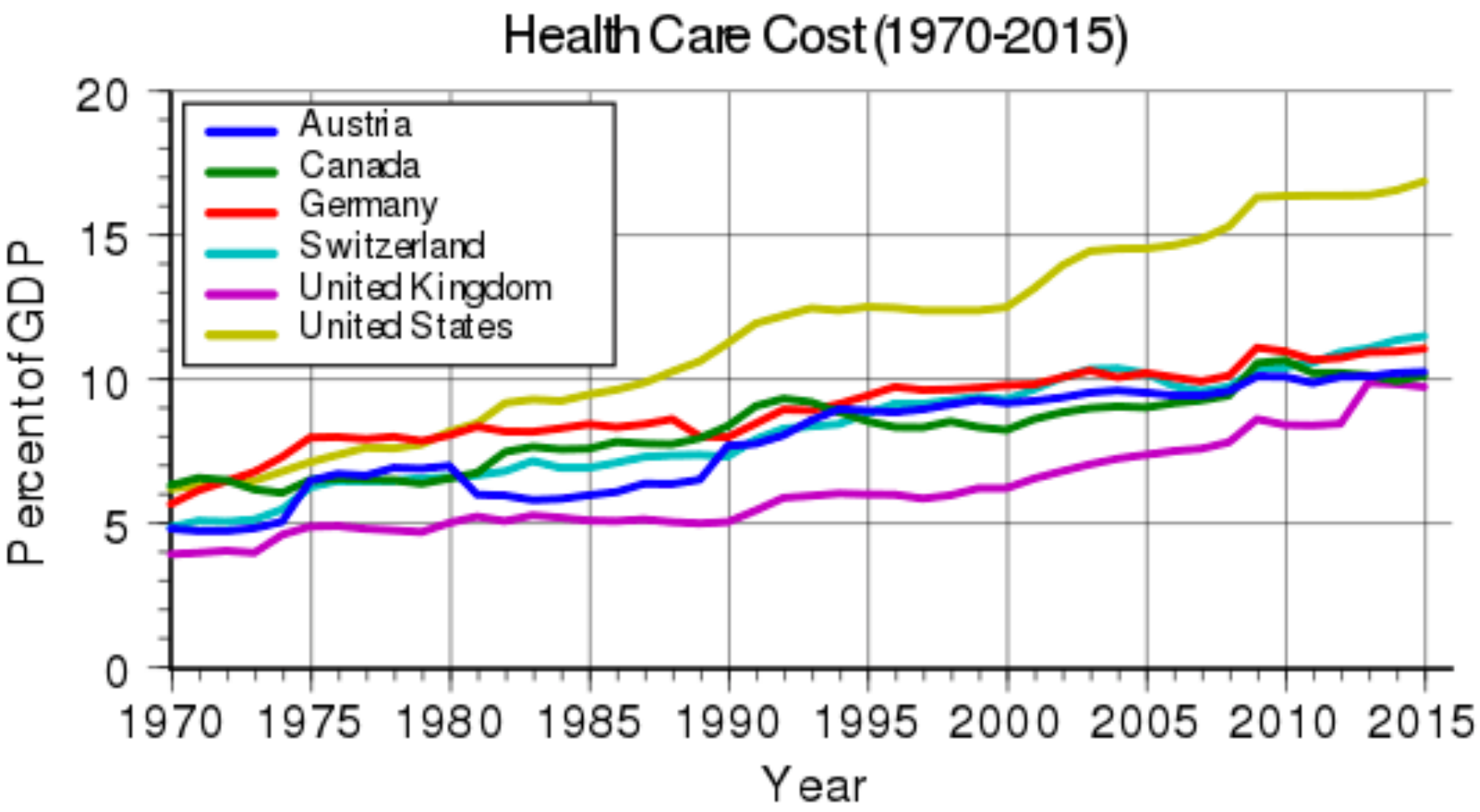
- Medical data is becoming more accessible
- Alternative sources of data are also available to help form diagnoses
- Technology allows us to combine qualitative and quantitative sources of information
- These trends suggest that medical students will need to become more proficient in data analysis

Healthcare spending



Soure: OECD Health Data 2013.
Data note: PPP = purchasing power parity.
Produced by Veronique de Rugy, Mercatus Center at George Mason University.

Health spending by % GDP

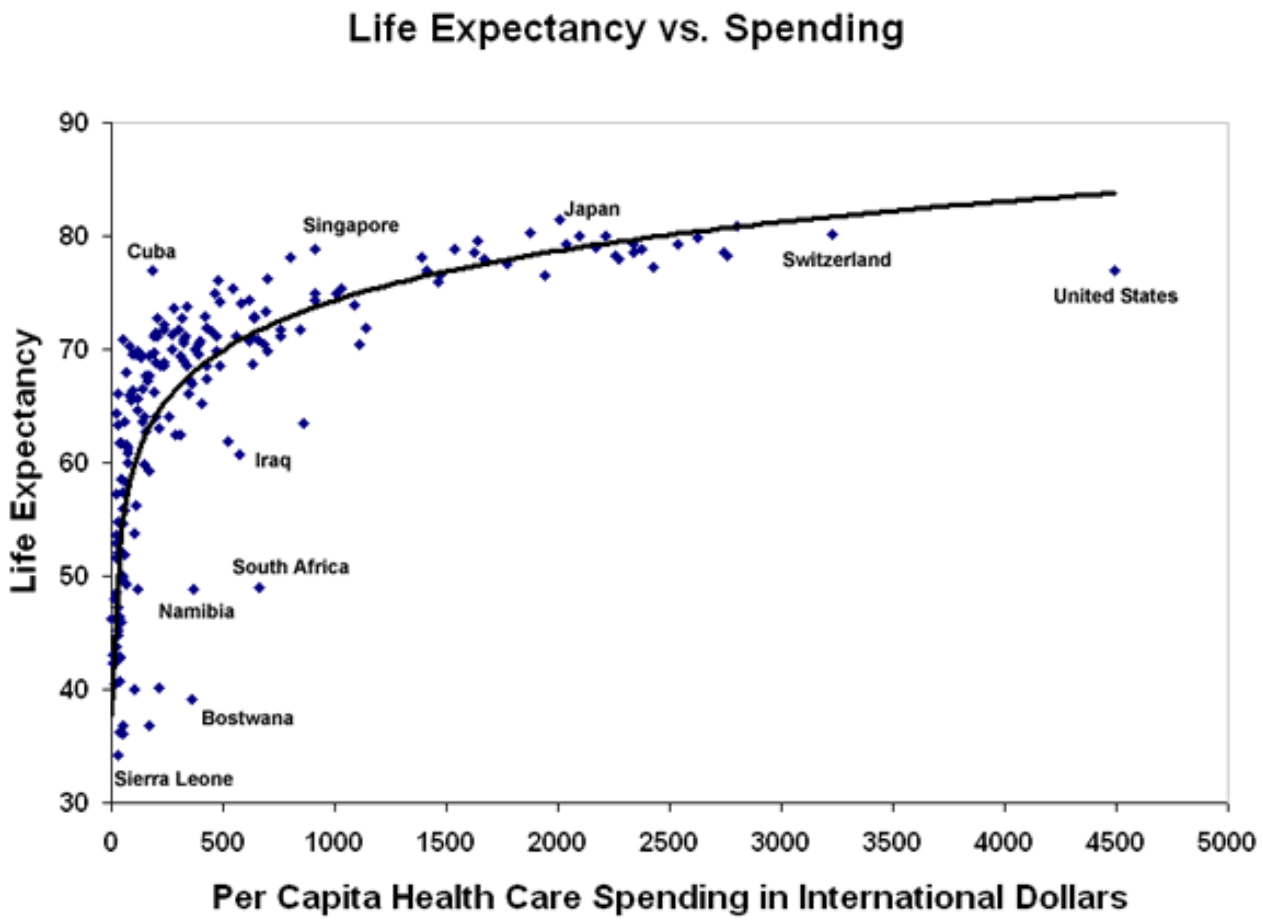


Poll

- To assess the value of healthcare in different countries, one could study which variable versus per capita health care spending?

- **Slido.com #36986**

Life expectancy versus spending



Malnutrition

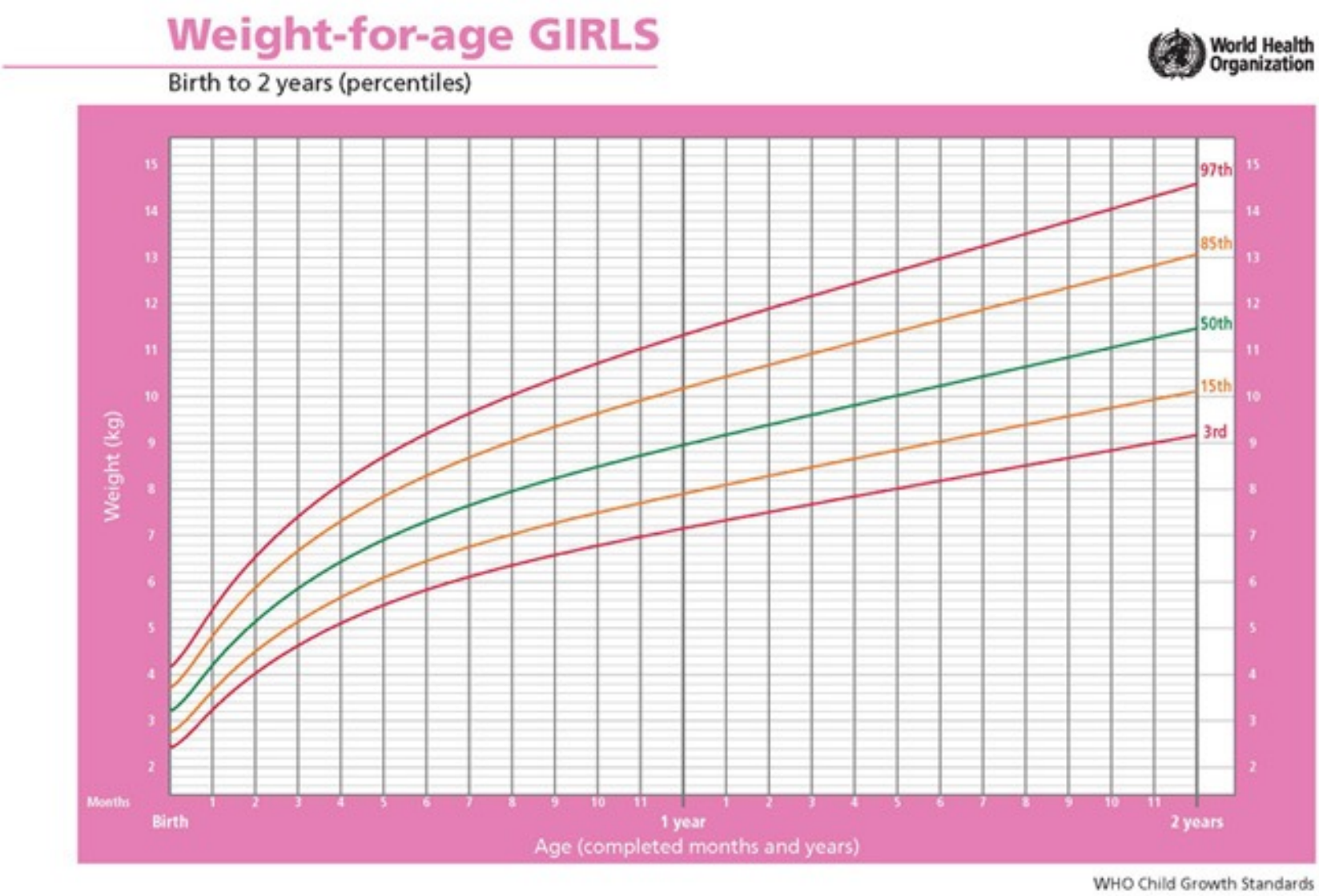
- Malnutrition is a serious condition that occurs when a person's diet doesn't contain the right amount of nutrients.
- Malnutrition is the largest single contributor to disease in the world, according to the UN's Standing Committee on Nutrition (SCN).
- A malnourished person finds that their body has difficulty doing normal things such as growing and resisting disease. Physical work becomes problematic and even learning abilities can be diminished (WFP).

Quiz

- How is a growth chart for babies (0 to 2 years) best described?
  a) Linear
  b) S-shaped
  c) Convex (slow and then rapid)
  d) Concave (rapid and then slow)

- **Slido.com #36986**

WHO growth charts



Growth charts

- Babies grow nonlinearly with time.
- Advanced Paediatric Life Support (APLS) estimates weight in Kg using age in years as:

  $$\text{weight} = [\text{age} + 4] \times 2.$$

- Used to estimate the weight of children in paediatric emergencies in order to administer correct drug dosages.
- But neither APLS nor WHO are accurate for malnutrition in developing countries.

Leffler formula

- However, as weight growth curves display evidence of nonlinearity, especially for young babies, more complex models have been developed.
- The Leffler formula provides distinct relationships for babies below and above the age of one year:

$$\text{weight} = 0.5a_m + 4$$

$$\text{weight} = 2a_y + 10$$

where a_m and a_y are age in month and years.

Nonlinear formula

- More complex models include the model of Theron and the non-parametric approach used by the WHO.
- Theron's formula (2005) is given by:

$$\text{weight} = \exp(0.175571 a_y + 2.197099)$$

- This is a log-linear relationship:

$$\log(\text{weight}) = b + c \cdot \text{age}$$

Nonlinear structures

- There are many examples of nonlinear relationships:

  $\log(\text{weight}) = a + b\ \text{age}$

  $\log(\text{weight}) = a + b \log(\text{age})$

  $\text{weight} = a + b\ \text{age} + c\ \text{age}^2$

- An alternative approach is to use a sigmoid relationship.

Community health workers

- The Rwandan Ministry of Health (MoH) created a program of community health workers (CHWs) in an initiative to improve the population's health status by ensuring access to preventative and curative healthcare services.
- Those CHWs are elected by their community and work as volunteers.
- Their activities focus mostly on children's health, vaccination, and malnutrition, as well as community-based activities around hygiene and sanitation.
- They also report disease epidemics in their area of coverage, and presently there are total of 40,000 CHWs in Rwanda.
- The population of Rwanda is 12 million which represent a ratio of 333 CHWs per 100,000 people.

The case study

- The case study was performed in cooperation with the Rwandan Ministry of Health (MoH) and Kibagabaga hospital in Gasabo district.
- Several approvals were acquired: Ethical approval from the Rwandan National Ethics Committee; and Data collection approval from the MoH.
- The liaison with the hospital was a coordinator of CHWs, and 24 participants were randomly selected from CHWs in her jurisdiction.
- Participating CHWs had a typical education, age and sex with the sample of approximately 1500 CHWs in Gasabo district, one of the most populated districts in Rwanda with approximately half a million people.

Brown, S. and McSharry, P. (2016). Improving accuracy and usability of growth charts: case study in Rwanda. British Medical Journal Open 2016;6: e009046.

Research goals

- Test the hypothesis that CHWs could use smart phones for health data collection.
- A mobile application was installed on a Tecno P3 smart phone, and given to the randomly selected Rwandan CHWs to monitor children's growth and development.
- Evaluation is carried out in both urban and rural locations, with 12 CHWs in each.
- The application is designed to be tolerant to delays in the mobile network and optimized for low-resource settings.
- The primary goals of this study it to assess the quality of health data collected by electronic tools.

Data collection

- CHWs collected electronic data for nine months (Mar-Dec 2014) on weight gain and Middle Upper Arm Circumference (MUAC), which are established measures for detecting malnutrition by UNICEF.
- Rwandan CHWs routinely collect data on those two indicators from children in their village once a month.
- For the purpose of analysis we also collected data recorded on paper notebooks from the participating CHWs.
- By the end of the study two CHWs, one from each location, dropped out because of equipment loss.
- Each group of CHWs was separately trained regarding smart phone and mobile application use

Weighing babies

Description of datasets

Data name	Description
Paper data	Data was obtained by recording children's weight from the CHWs' books in an excel spread sheet. Data contains information on weight and age for 320 boys and 380 girls. Those individual records are kept by the same CHWs participating in the case study.
Electronic data	In the case study 24 participating CHWs collected data using a custom made smart phone mobile application. By the end of the study they collected data for 922 girls and 886 boys who live in their district. The data is collected over 9 months, and most of the data is cross-sectional, with about 330 children with time series data for 3-6 consecutive months.
WHO data	The WHO study was carried out in six different countries: Brazil, Ghana, India, Norway, Oman and the USA in year 2006. The WHO standards are based on a longitudinal study of 882 children aged 0–24 months and on cross-sectional studies of 6669 children aged 18–71 months.

Poll

- Which of the following is likely best for representing the growth of babies in Rwanda?
  a) Rwandan paper records
  b) Rwandan electronic records
  c) World Health Organisation (WHO) growth curve

- **Slido.com #36986**

Data quality

Data quality dimension	Sub dimension	Meaning
Intrinsic	Accuracy Reputation	Data correct Trusted source
Contextual	Completeness	Values present
Representational	Interoperability Consistency	Language and unit correct Ease of understanding
Accessibility	Accessibility Security	Easy to retrieve Access restricted
Additional digital data timeliness capability	Feedback Trends Timeliness	Two-way communication Visual presentation Instant availability

Adopted from Baesens' "Analytics in a Big data World"

Rwanda data collection

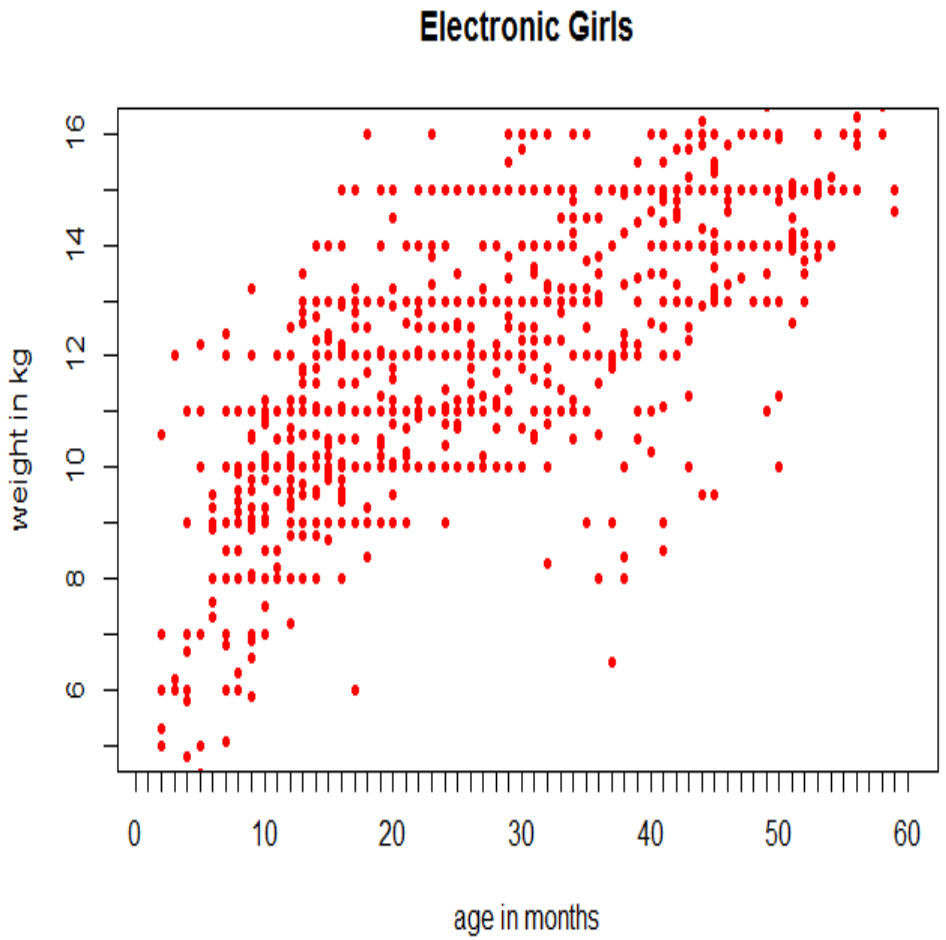
- **Accuracy**: the present paper system does not have consistency checks while the EHR approach incorporates those features in the custom mobile application.
- **Consistency**: by comparing the values that are input with a model based on previous measurements, it is possible to immediately detect errors and ensure self-consistency by prompting the data collector to revise the measurement.
- **Accessibility**: electronic reporting is timelier than paper reporting since the present paper system is submitted only once a month. The web application designed for the case study displays the data as soon as the network synchronizes the mobile app with the web app.
- **Feedback**: in mobile collection system it is possible to get a real time feedback and have a two-way communication channel.

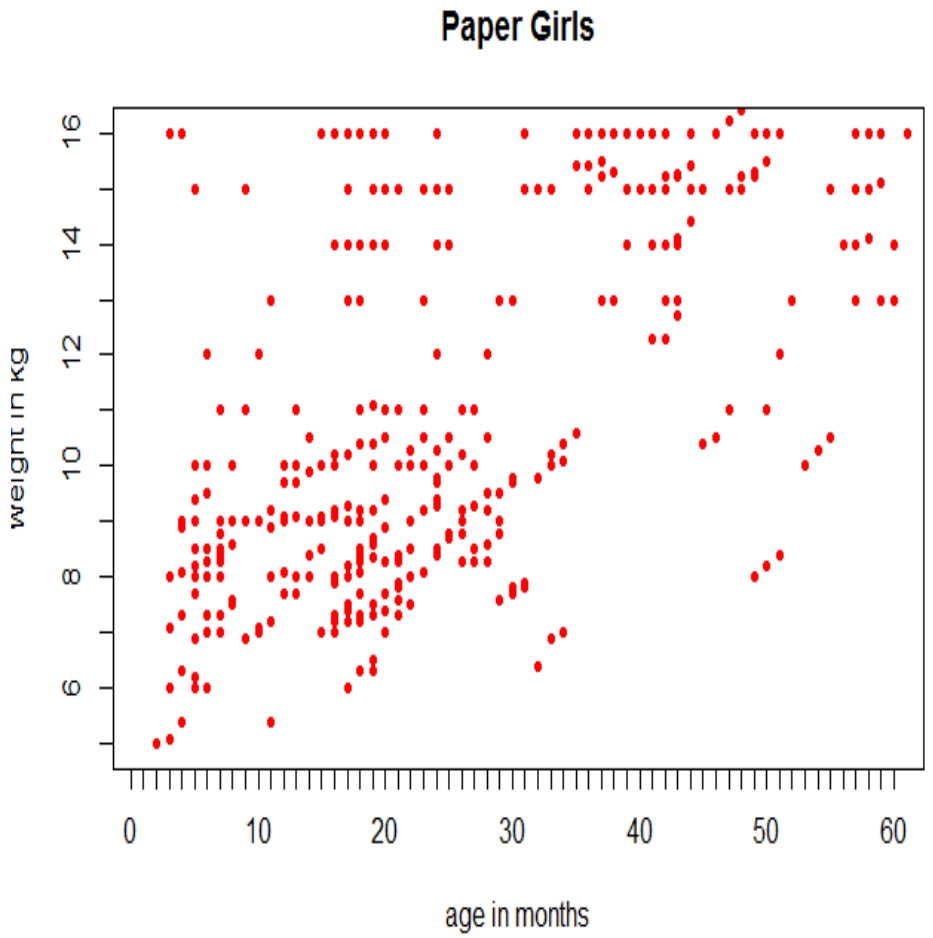
Poll

- When attempting to identify a pattern, it is always recommended to try visualization techniques first.
  a) Yes
  b) No

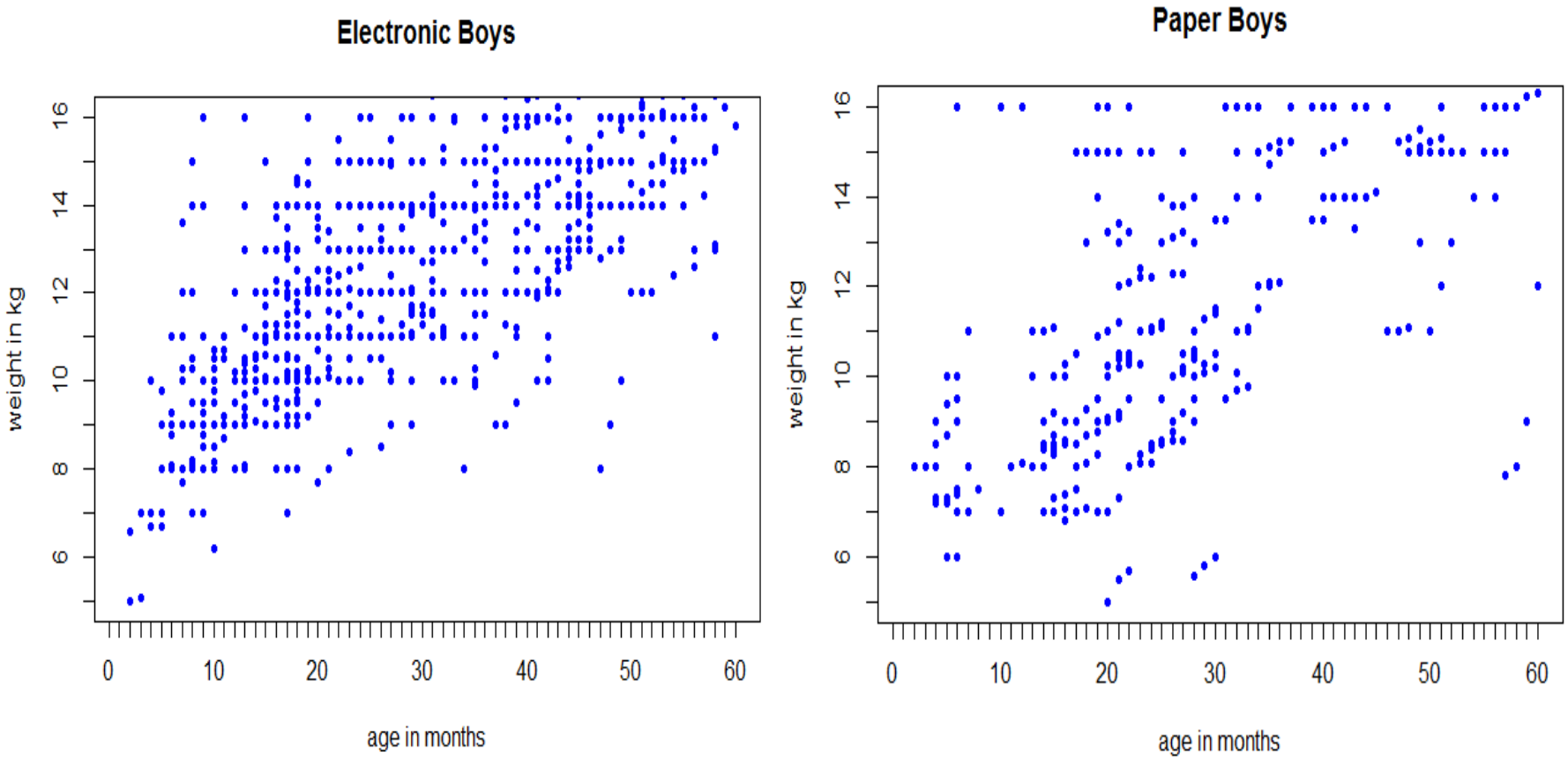
- **Slido.com #36986**

Electronic versus paper (girls)





Electronic versus paper (boys)



Evaluation metrics

- Denote y and y* as the actual and predicted values respectively and < > represents an average over all such evaluation pairs
- Mean Absolute Error (MAE):

  $MAE = \langle |y-y^*| \rangle$

- Mean Absolute Percentage Error (MAPE):

  $MAPE = \langle |y-y^*|/y \rangle$

- Coefficient of determination (R^2):

  R^2 = square of correlation coefficient

Model selection

- We used to data to select an optimal model structure.
- A log-log model was found to be most appropriate after comparing results with a log-linear model such as Theron's formula.
- The model structure is given by

$$\log(\text{weight}) = a + b \log(\text{age})$$

Evaluations

Evaluation	Training Data	Testing Data	Model	R^2
A	Paper records	Paper records	log(weight) = 0.70 + 0.26*log(age)	0.37
B	Electronic records	Electronic records	log(weight) = 0.69 + 0.28*log(age)	0.56
C	Simulated data from WHO chart	Simulated data from WHO chart	log(weight) = 0.56 + 0.37*log(age)	0.92
D	Simulated data from WHO chart	Electronic records	log(weight) = 0.56 + 0.37*log(age)	0.54

Comparing electronic and paper

R^2 values for the model, log(weight) = a +b log(age), for boys and girls using Electronic and Paper collection.

Gender	Rwanda Electronic	Rwanda Paper	Gain in R^2 due to electronic approach
Girls	0.56	0.37	51%
Boys	0.58	0.35	66%

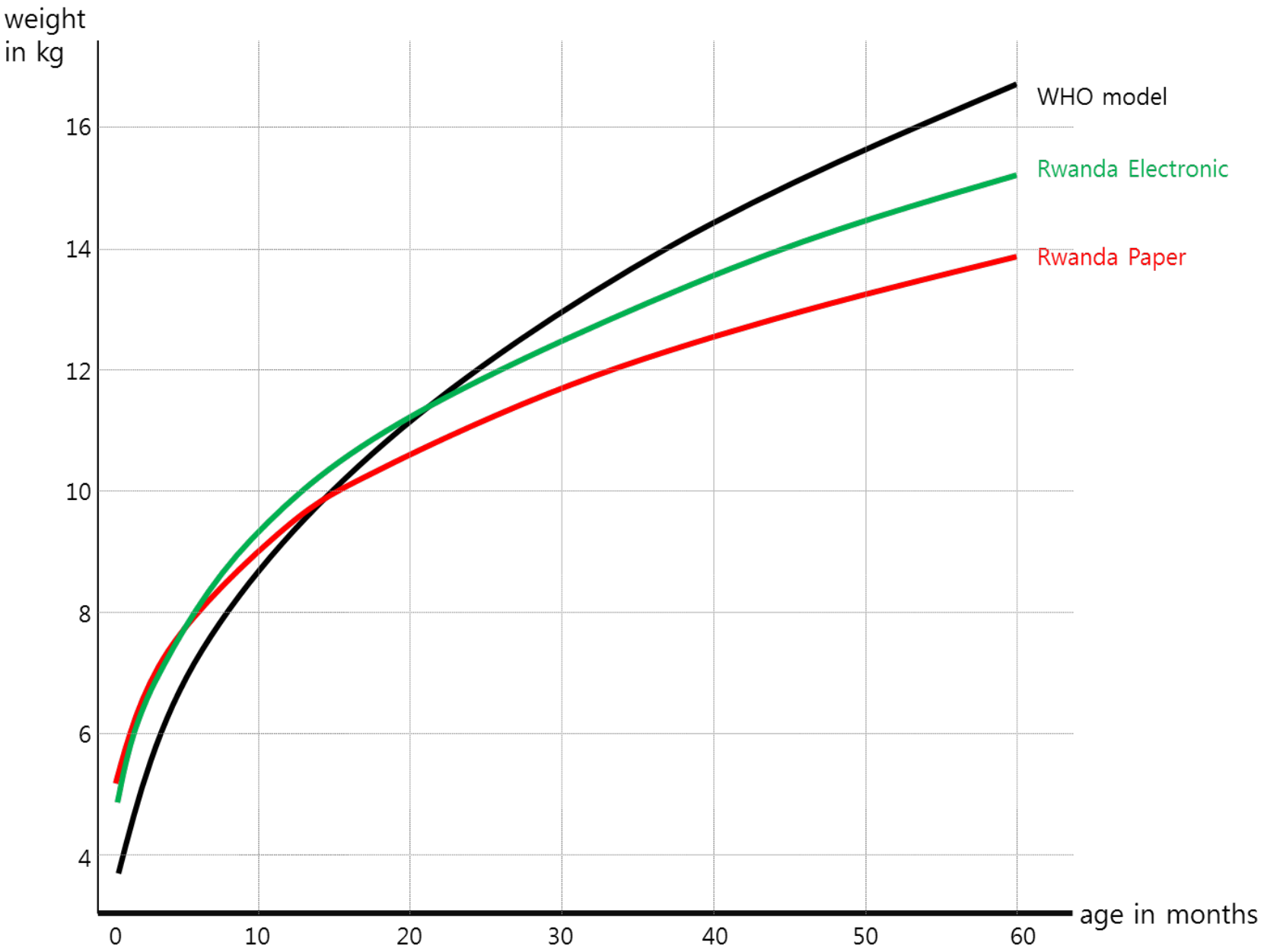
Assessment for Rwandan girls

A direct comparison of traditional forecast evaluation criteria, such as Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) provides a way to quantify the magnitude of the error in using a given model to make a prediction.

A model of the form, log(weight) = a + b log(age), allows us to compare the paper and electronic records, giving errors in units of kilograms.

Electronic records provide considerable improvements over paper records, with at least a 40% reduction in error in both performance metrics

Performance Metric	Rwanda Electronic (Evaluation B)	Rwanda Paper (Evaluation A)	Reduction in error of Electronic over Paper data
MAE	1.4 kg	2.4 kg	40%
MAPE	12%	21%	42%


WHO, Rwanda Electronic and Rwanda Paper for Girls under 5
weight
in kg
16
14
12
10
8
6
4
0
10
20
30
40
50
60
age in months
WHO model
Rwanda Electronic
Rwanda Paper

WHO versus Rwanda Electronic

Another comparison is made between Rwanda Electronic (Evaluation B) and model based on electronic records but coefficient estimates from the WHO regression (Evaluation D).

The two forecast performance metrics, MAE and MAPE, for the Rwanda Electronic (Evaluation B) and the WHO model (evaluation D), for Rwandan girls.

The Rwanda Electronic model is outperforming the WHO model. The gain in performance of the Rwanda Electronic model is 10% for MAPE and 7% for MAE. This gain was statistically significant ($p<0.01$) based on a Kolmogorov-Smirnov test.

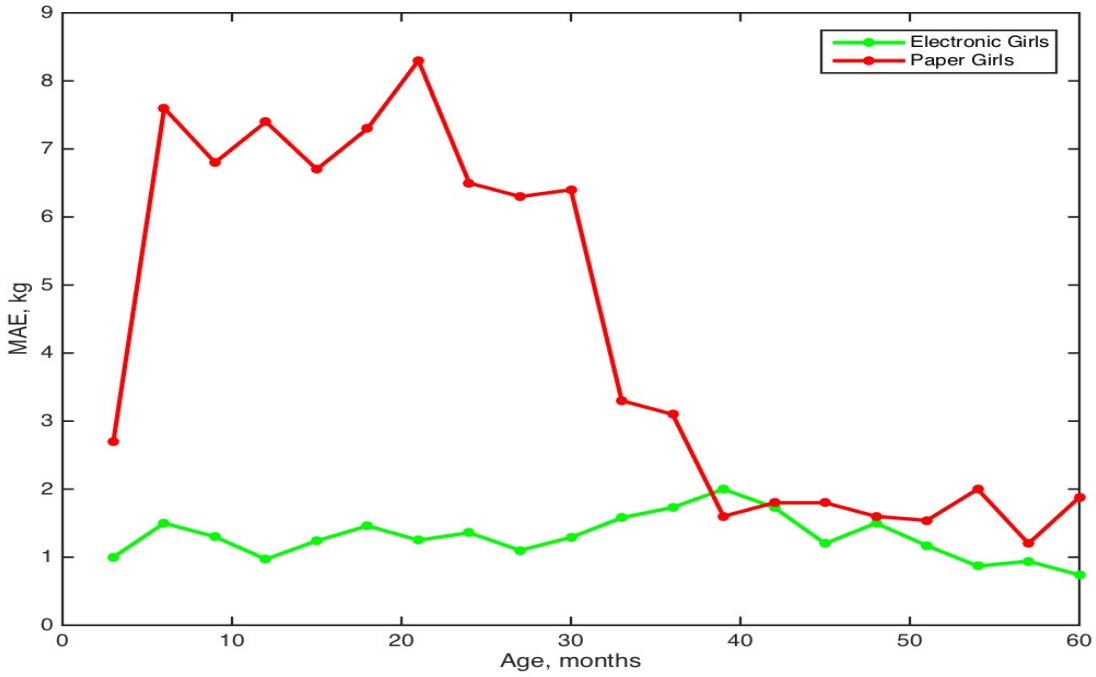
Metric	WHO model (Evaluation D)	Rwanda Electronic (Evaluation B)	Reduction in error of Electronic model over WHO model
MAE	1.5 kg	1.4 kg	7%
MAPE	13.2 %	12%	10%

Non-parametric comparison

The MAE for the electronic recording does not vary significantly, and ranges between 1 and 2 kg.

Alternatively the MAE for paper recordings varies significantly over the growth chart with a major decrease in MAE taking place after three years of age.

The MAE for paper recording fluctuates around 7 kg for the first three years, only dropping to a comparable level of error for ages three to five



Conclusions

- Electronic data records for Rwandan children have a goodness of fit, measured by R^2, which is more than twice that of the Paper data records for both boys and girls.
- Electronic data and Paper data collected from Rwandan children differ from WHO growth curves, raising doubts about applicability of WHO growth charts to developing countries with considerable malnutrition.
- Comparisons of the electronic and paper recording methods with the standard WHO model show that electronic data is closer to WHO model.
- Electronic data improves performance over the WHO model by 10% in mean absolute percentage error and 7% in mean absolute error. Results are statistically significant using Kolmogorov-Smirnov test at $p < 0.01$.

Data Analytics

WEEK 5B

Assignment 3

- Intra-day 15-min electricity demand
- Ten days equals 10 x 24 x 4 samples
- Seasonality expected: diurnal and weekly
- Watch out for daylight saving
- Time of year variable: 0 to 1
- Use weekday to identify particular days
- MAE used in Q9: error in Watts
- MAPE used in Q10: error as a percentage

Course outline

Week	Lecture A	Lecture B
1	Data Analytics	Weather forecasting
2	Renewable energy	Wind energy
3	Solar energy	Demand forecasting
4	Risk	Extreme events
5	Health	Biomedicine
6	Early warning systems	Economic forecasting

Today’s Lecture

No.	Activity	Description	Time
1	Challenge	Detecting disorders	10
2	Discussion	Epilepsy	10
3	Case study	ECG analysis	10
4	Analysis	Parkinson’s disease	20
5	Demo	Voice disorders	20
6	Q&A	Questions and feedback	10

Quiz

- Which part of the human body is associated with the biomedical signal known as the electroencephalogram (EEG)?
  a) Feet
  b) Heart
  c) Brain
  d) Stomach

- **Slido.com #46920**

Healthcare Applications

- Epilepsy – Electroencephalogram (EEG)
- Cardiac disorders – Electrocardiogram (ECG)
- Voice disorders – Speech signals
- Parkinson’s Disease – Gait signals

Epileptic seizures

- Serious neurological disorder
- Affects 1% of the population at some time
- Identification of seizure precursors could facilitate clinical intervention
- Require robust and reliable detection
- Automated detection using scalp electroencephalogram (EEG) assists clinicians
- Prediction may also be possible

McSharry et al., Med. Biol. Eng. Comput., 2002, 40, 447–461

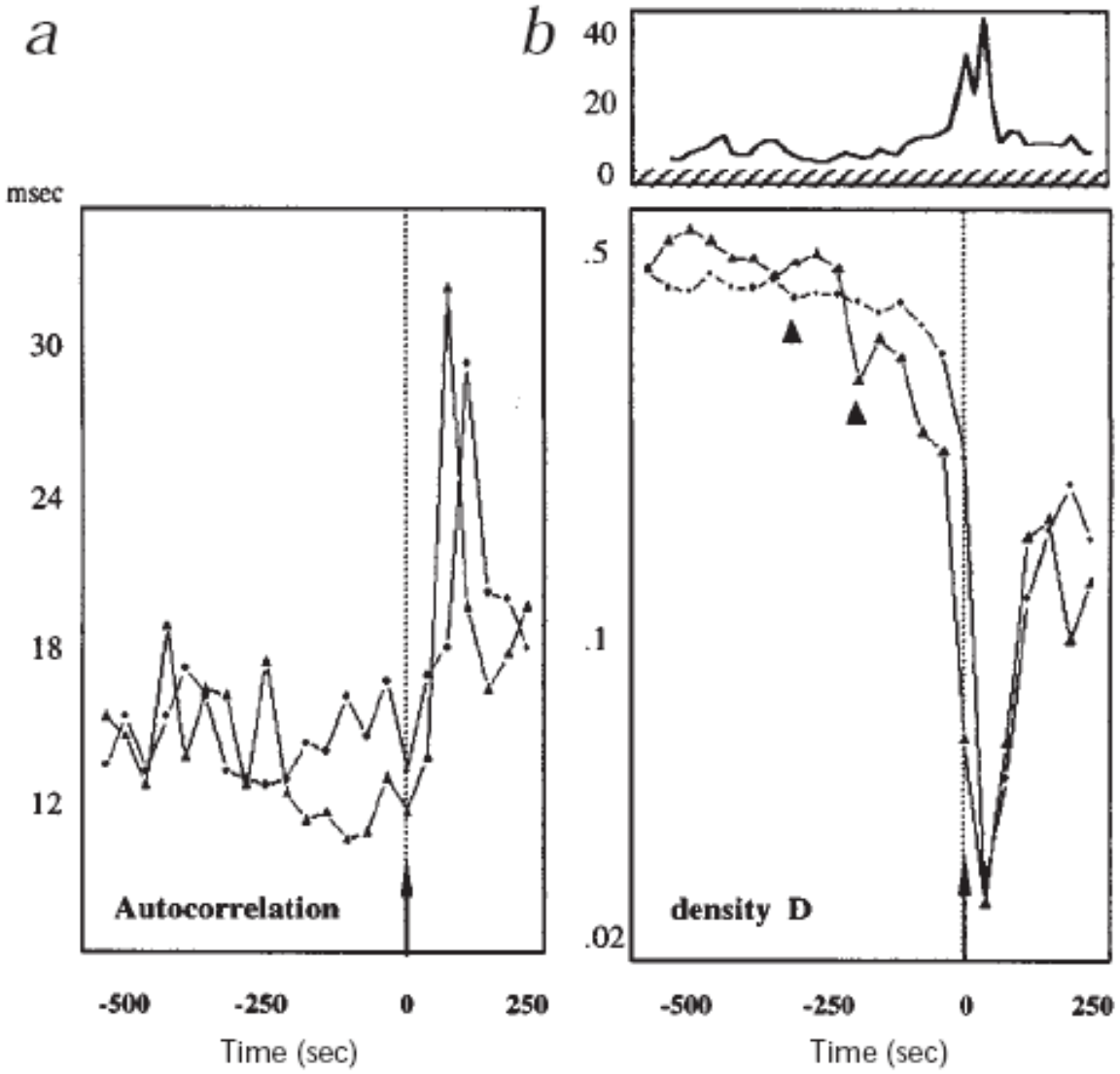
McSharry et al., IEEE Trans. Biomed. Eng., 2003, 50(5):628-623

Ranking Poll

- When publishing a study about a new data-driven technique, we should aim to:
  a) Understand how and why it works
  b) Be fully transparent about the approach
  c) Compare with existing benchmarks
  d) Make the data publicly available

- **Slido.com #46920**

Nature Medicine – Martinerie et al. 1999

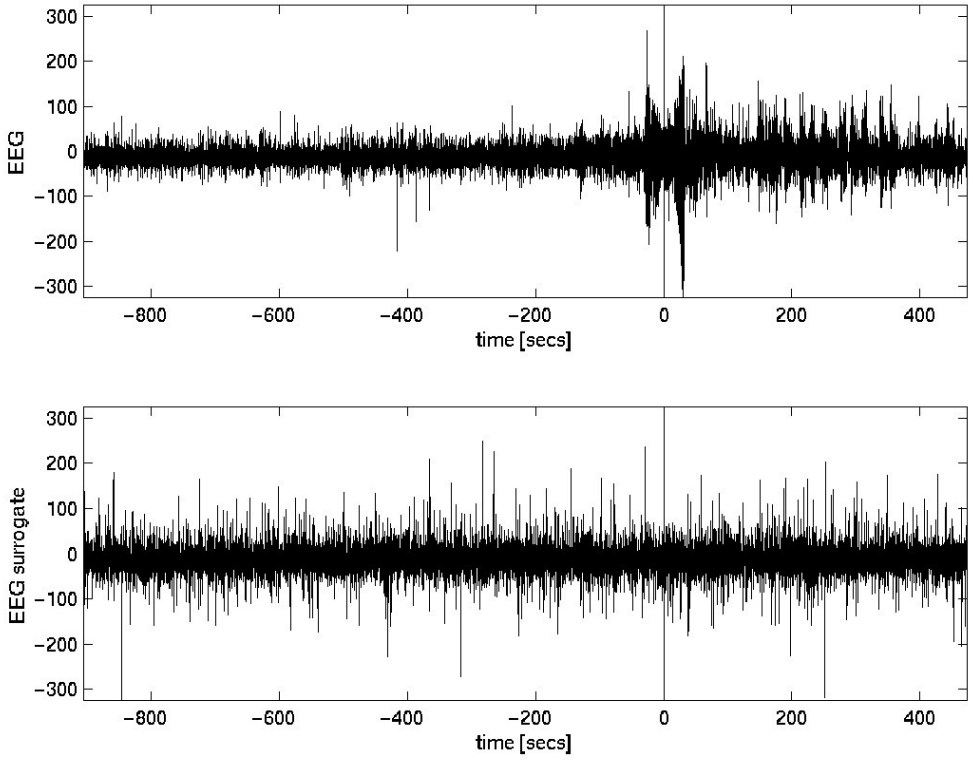


Surrogates

- How to be sure that the changes you detect are not simply due to linearity?
- Monte Carlo hypothesis testing
- Preserve power spectrum and autocorrelation by fixing the amplitudes of the FFT and scrambling the phases
- Polished surrogates also attempt to maintain the same distribution as the original data
- Compute sample distribution of nonlinear statistic using surrogates
- Test for a significant difference with respect to the original data

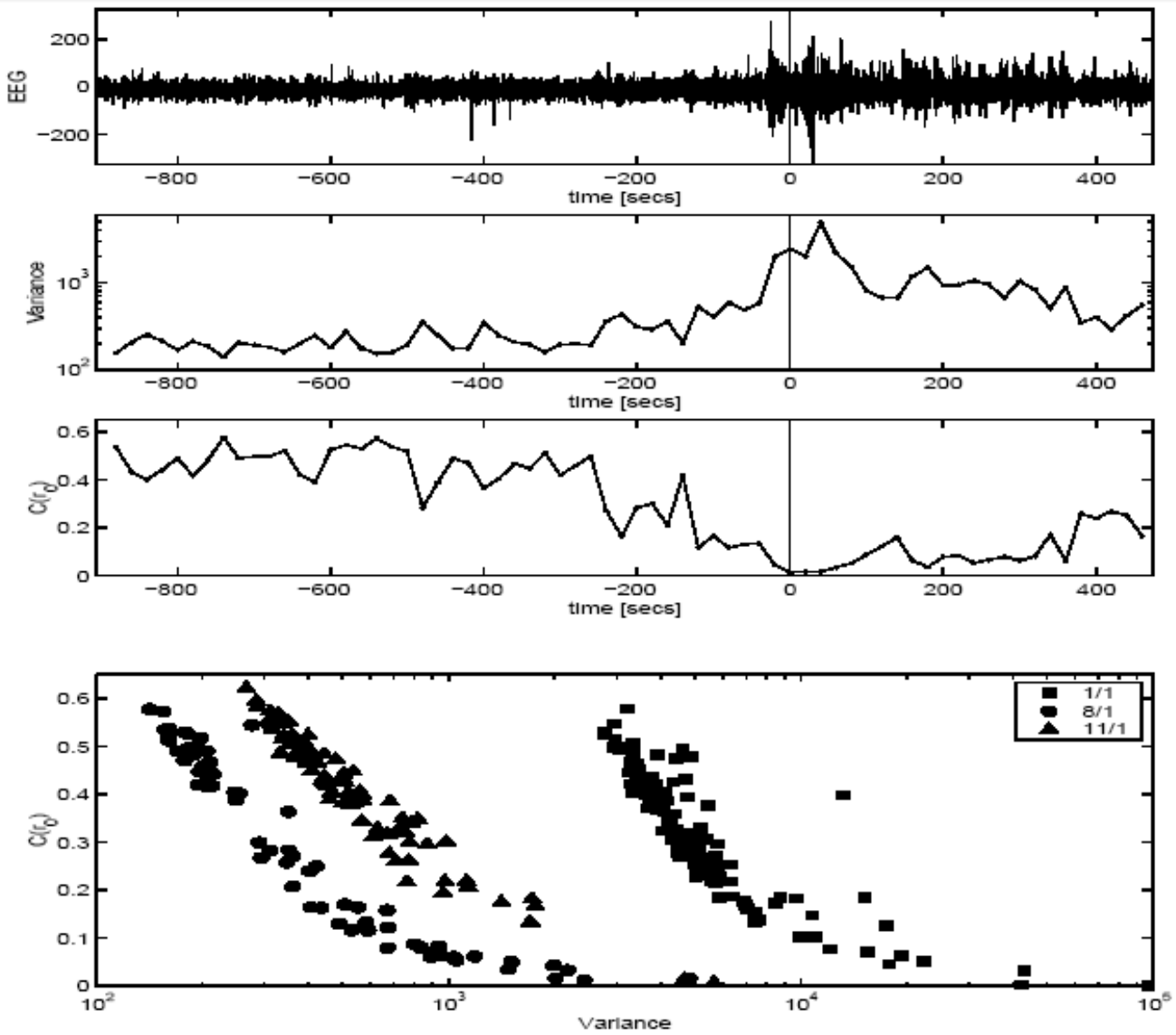
Surrogate data

- Aim: to measure significance against a specific null hypothesis



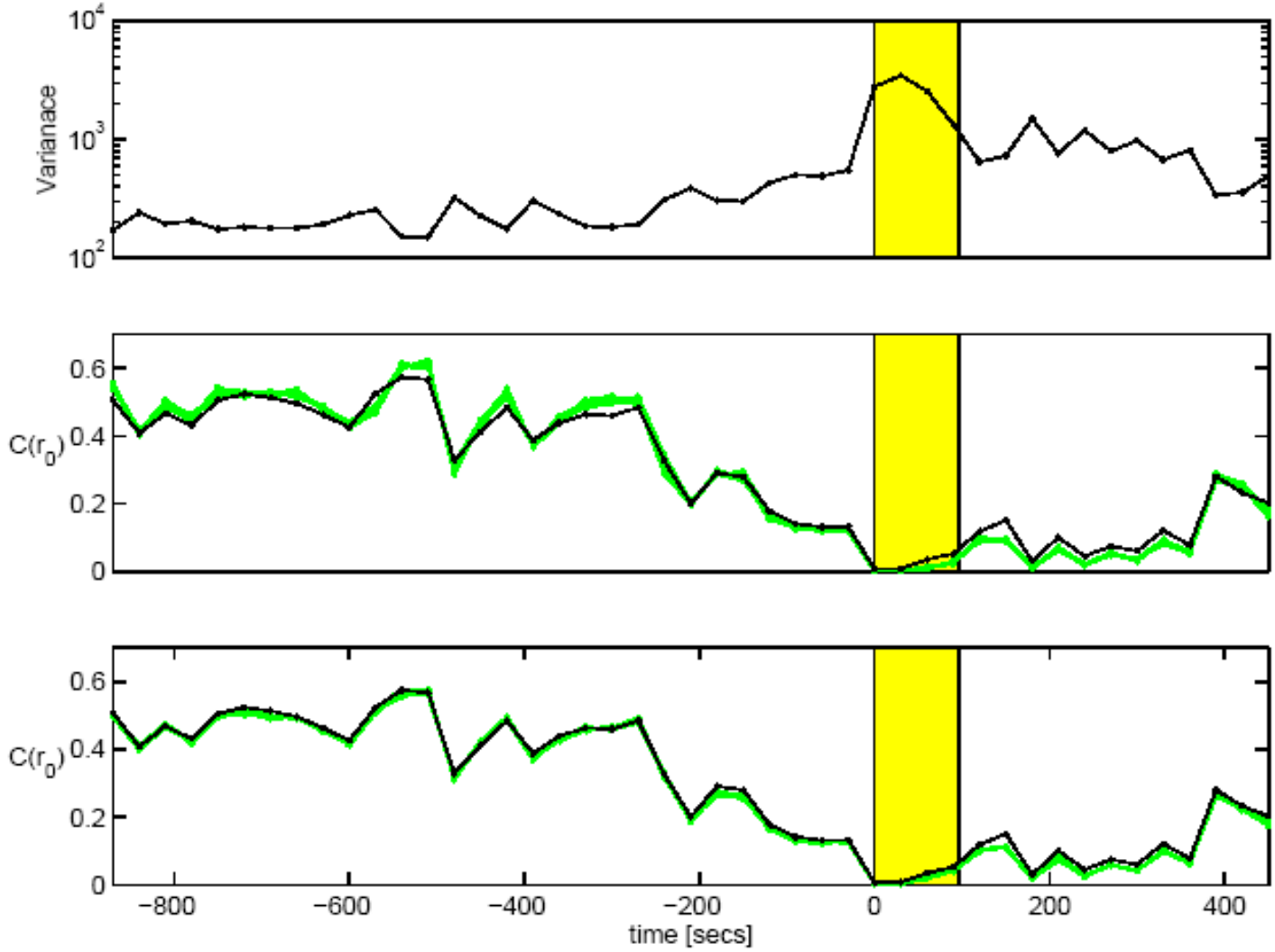
- Choice of a relevant null is critical when developing a clinical technique!

Correlation density versus variance



McSharry *et al.* Nature Medicine 9(3):241-242 (2003)

Block Surrogate analysis preserving heteroskedasticity



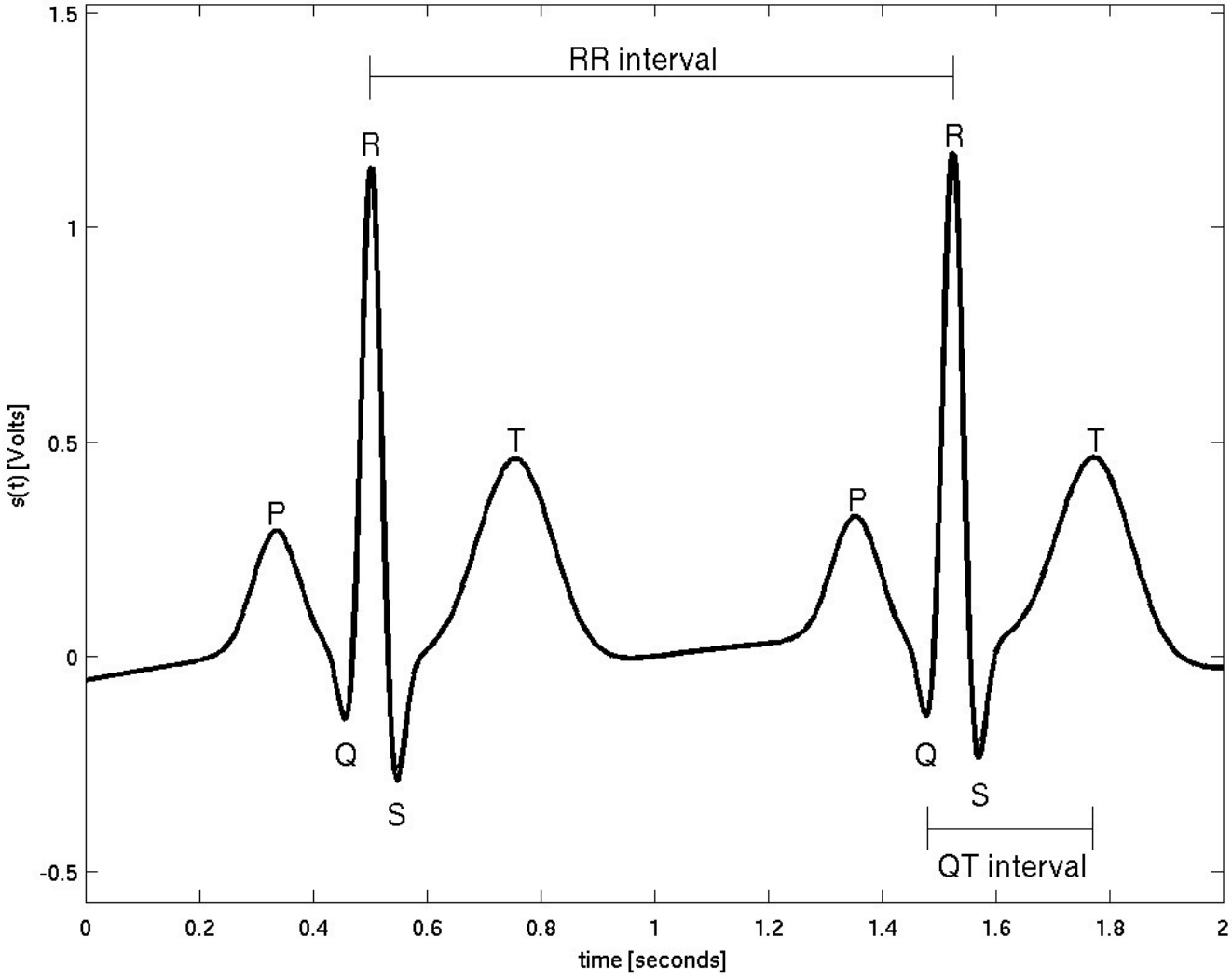
McSharry *et al.* IEEE Trans. Biomed. Eng. 50(5): 628-633 (2003)

Poll

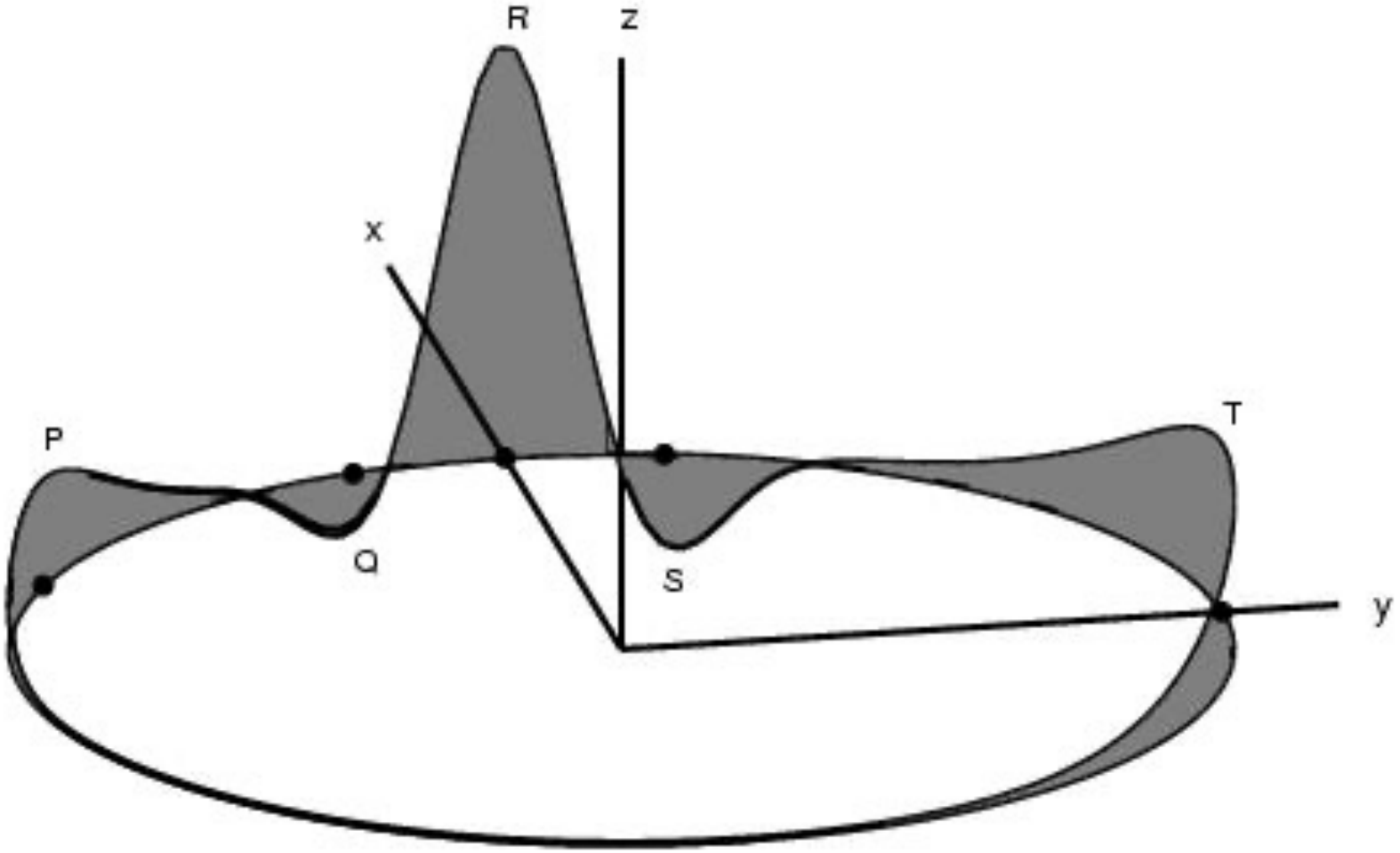
- Identify the correct statement about the frequencies of these two human body oscillations:
  a) Respiration < Heart rate
  b) Respiration = Heart rate
  c) Respiration > Heart rate

- **Slido.com #46920**

The electrocardiogram (ECG)



3D dynamical model of the ECG



www.physionet.org/physiotools/ecgsyn/

McSharry PE, Clifford GD, Tarassenko L, Smith L.

IEEE Transactions on Biomedical Engineering **50**(3): 289-294 (2003)

3D equations of motion

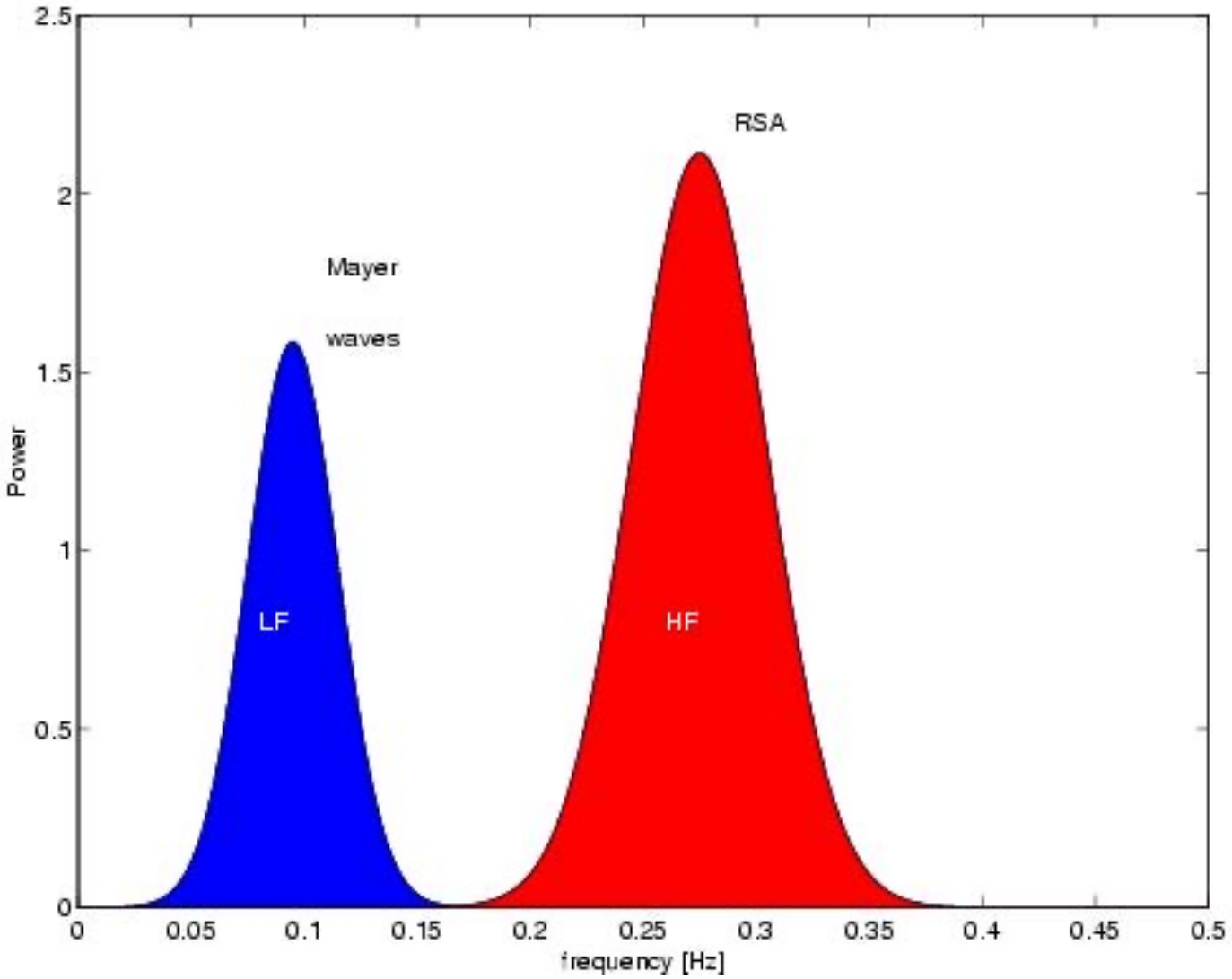
$$\dot{x} = \alpha x - \omega y$$

$$\dot{y} = \alpha y + \omega x$$

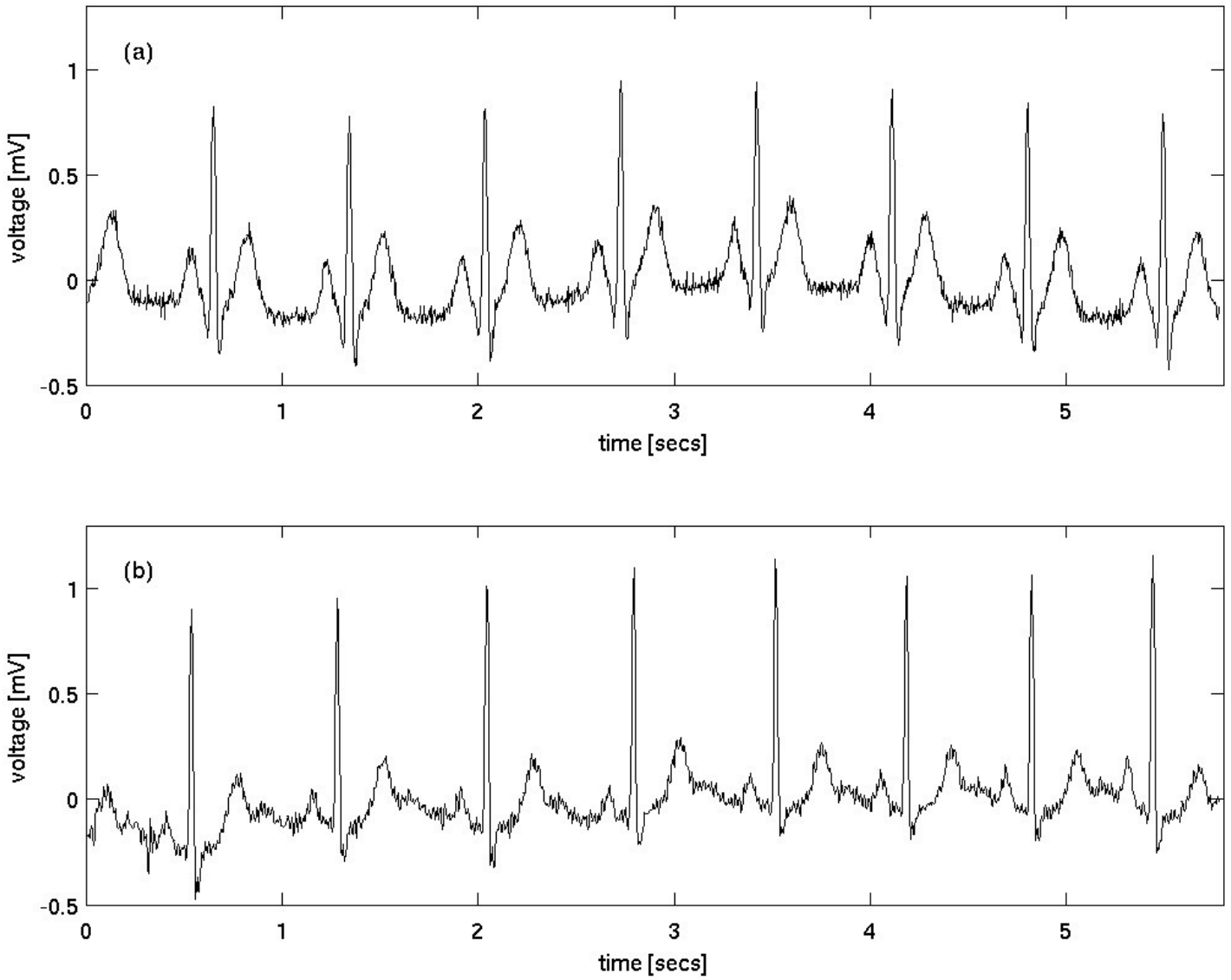
$$\dot{z} = -\sum_i a_i \Delta\theta_i \exp\left(-\frac{\Delta\theta_i^2}{2b_i^2}\right) - (z - z_0)$$

- where $\alpha = 1 - (x^2 + y^2)^{1/2}$, $\Delta\theta_i = \theta - \theta_i \bmod 2\pi$, $\theta = \text{atan2}(y,x)$, and ω is the angular velocity
- a_i govern the magnitude of the peaks
- b_i define the width (time duration) of each peak
- Baseline wander may be introduced by coupling the baseline value z_0 to the respiratory frequency.

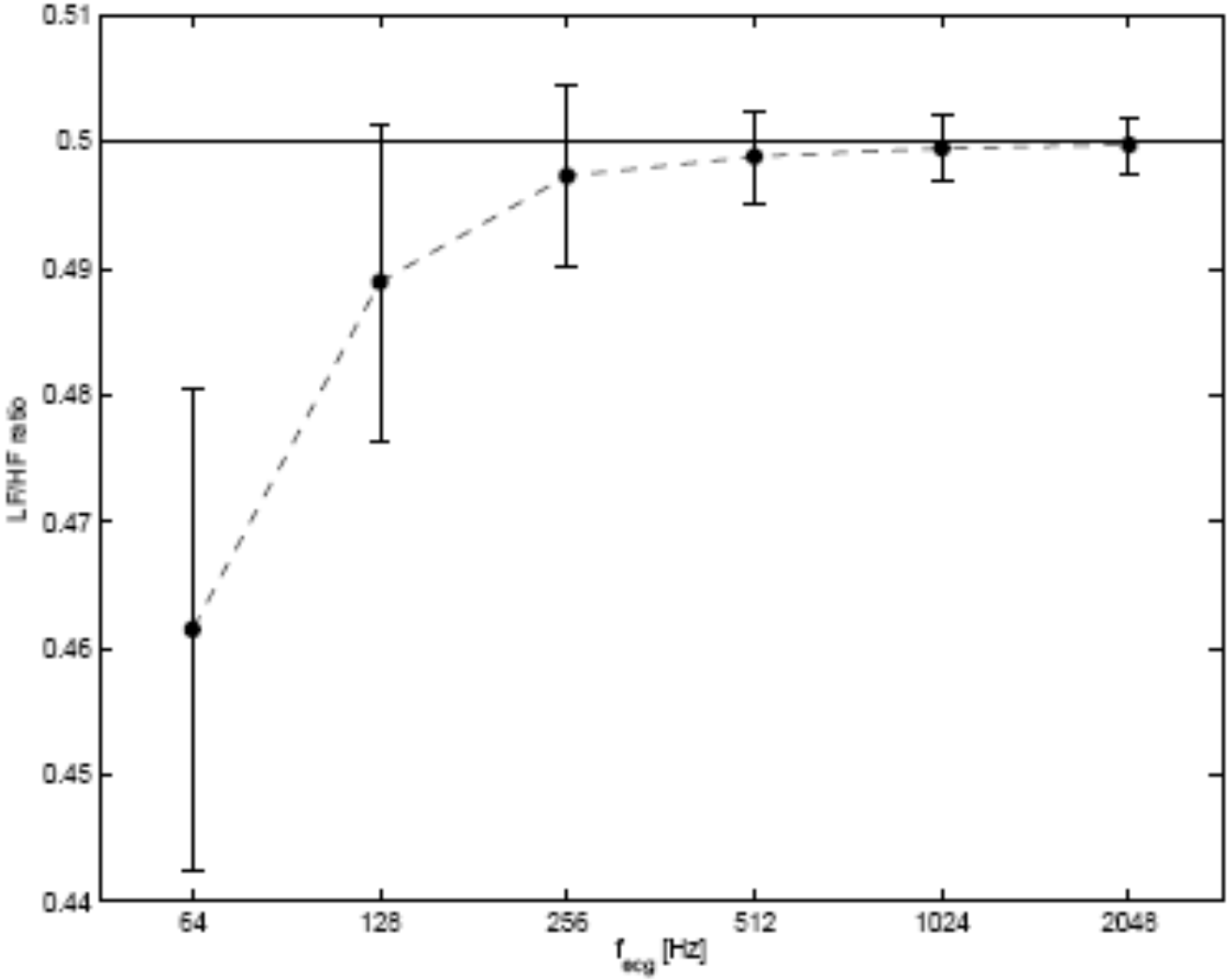
Spectral properties of RR intervals used to drive the angular frequency ω



Comparison of artificial and real ECGs



ECG sampling frequency



ECGSYN applications

- Testing biomedical signal processing techniques (Physionet, Harvard-MIT)
- Teaching medical students by producing arrythmias and morphological changes
- Measuring QT intervals; drug development
- Dynamical model for noise reduction (Fetal ECG extraction, collaborators in MIT, Paris and Liverpool)
- Racing heart music: http://www.mcsharry.net

Poll

- Would we expect AI/ML to improve a classifier constructed by an experienced human expert?
  a) Yes
  b) No

- **Slido.com #46920**

Parkinson's Disease

- Parkinson's disease (PD) claims lives at an epidemic rate (affecting ~20/100,000 people every year).
- There is no treatment, but drugs can alleviate some of the symptoms.
- Clinical metric used to quantify average symptom severity: Unified Parkinson's Disease Rating Scale (UPDRS).
- Currently, UPDRS is estimated by clinical raters (subjective, inter-rater variability).
- PD affects speech, and there is empirical evidence of degrading speech performance with disease progression.
- We propose objective mapping of features extracted from speech signals (dysphonia measures) to UPDRS.

Parkinson's Disease Rating

- Parkinson's disease (PD) symptom severity is typically quantified using clinical metrics, where a medical rater assesses the subject's condition and ability to perform a range of tasks.
- Two of the most commonly used PD metrics are the Unified Parkinson's Disease Rating Scale (UPDRS) and the Hoehn and Yahr (H&Y) scale.

Tsanas, Little, McSharry, Scanlon, Papapetropoulos (2012). Parkinsonism and Related Disorders 18: 697–699.

Scanlon's formula

- A mathematical formula was proposed to express H&Y as a function of UPDRS using intuitive rules based upon H&Y evaluation guidelines [Scanlon et al.,, 2008].
- This formula uses a decision tree applied to 27 clinical variables such as postural stability, gait and ability to rise from a chair.
- The formula outputs H&Y ratings of {1,1.5,2,2.5,3,4,5}.

Scanlon's formula

Scanlon's formulas with parameters optimized using a genetic algorithm.

Refined Scanlon's formula

if ((item18 = 0 AND item19 = 0) AND ((item20L + item21L + item22L + item23L + item24L + item25L + item26L = 0) OR
(item20R + item21R + item22R + item23R + item24R + item25R + item26R = 0))), **HY = 1**;
if ((item18 > 0 OR item19 > 0) AND ((item20L + item21L + item22L + item23L + item24L + item25L + item26L = 0) OR
(item20R + item21R + item22R + item23R + item24R + item25R + item26R = 0))), **HY = 1.5**;
if (item30 = 0 AND ((item20L + item21L + item22L + item23L + item24L + item25L + item26L > 0) AND
(item20R + item21R + item22R + item23R + item24R + item25R + item26R > 0))), **HY = 2**;
if ((item30 = 1) AND ((item20L + item21L + item22L + item23L + item24L + item25L + item26L > 0) AND
(item20R + item21R + item22R + item23R + item24R + item25R + item26R > 0))), **HY = 2.5**;
if ((item30 > 1 AND item30 < 4) AND ((item20L + item21L + item22L + item23L + item24L + item25L + item26L > 0) AND
(item20R + item21R + item22R + item23R + item24R + item25R + item26R > 0))), **HY = 3**;
if (((item29 > $\mathbf{X_1}$ AND item29 < 4) AND (item27 < 4) AND (item31 > $\mathbf{X_2}$ AND item31 <=4))), **HY = 4**;
if (item29 = 4) OR (item30 = 4), **HY = 5**;

Refined, modified Scanlon's formula

if ((item18 = 0 AND item19 = 0 AND item22_neck = 0) AND ((item20L + item21L + item22L + item23L + item24L + item25L + item26L = 0) OR
(item20R + item21R + item22R + item23R + item24R + item25R + item26R = 0))), **HY = 1**;
if ((item18 > 0 OR item19 > 0 OR item22_neck >0) AND ((item20L + item21L + item22L + item23L + item24L + item25L + item26L = 0) OR
(item20R + item21R + item22R + item23R + item24R + item25R + item26R = 0))), **HY = 1.5**;
if (item30 = 0 AND ((item20L + item21L + item22L + item23L + item24L + item25L + item26L > 0) AND
(item20R + item21R + item22R + item23R + item24R + item25R + item26R > 0))), **HY = 2**;
if ((item30 = 1) AND ((item20L + item21L + item22L + item23L + item24L + item25L + item26L > 0) AND
(item20R + item21R + item22R + item23R + item24R + item25R + item26R > 0))), **HY = 2.5**;
if ((item30 > 1 AND item30 < 4) AND ((item20L + item21L + item22L + item23L + item24L + item25L + item26L > 0) AND
(item20R + item21R + item22R + item23R + item24R + item25R + item26R > 0))), **HY = 3**;
if (((item29 > $\mathbf{X_3}$ AND item29 < 4) AND (item27 < 4) AND (item31 > $\mathbf{X_4}$ AND item31 <=4))), **HY = 4**;
if (((item29 = 4) OR (item30 = 4))), **HY = 5**;

$\mathbf{X_1}$ = 1(S), 2(M), 2(F), $\mathbf{X_2}$ = 2(S), 2(M), 1(F), $\mathbf{X_3}$ = 1(S), 1(M), 2(F), $\mathbf{X_4}$ = 2(S), 3(M), 1(F)

The parameters in Scanlon's formulas which have been refined using the genetic algorithm appear in the form $\mathbf{X_k}$, as explained in the last row of this Table. 'S' stands for Scanlon's original formula, and the genetic algorithm values are denoted 'M' for the male subset, and 'F' for the female subset.

Empirical study

- University of Rochester's Parkinson's Disease Data and Organizing Center provided a database of 566 subjects (1623 samples as multiple evaluations).
- Data from 525 (340 male) subjects, giving 1486 (979 male) samples with no missing values.
- Ages were (mean ± standard deviation) 62.7 ± 9.3 years, with 4.7 ± 1.9 years since PD diagnosis.
- UPDRS score: 32.7 ± 15.6, and H&Y: 2.0 ± 0.5.

Relationship between UPDRS and H&Y

Male (N = 979)			Female (N = 507)		
UPDRS	MI	Spearman	UPDRS	MI	Spearman
Postural Stability	0.282	0.627	Postural Stability	0.330	0.664
Gait	0.119	0.393	Gait	0.164	0.425
Arise from chair	0.118	0.398	Arise from chair	0.111	0.365
Bradykinesia	0.106	0.390	Bradykinesia	0.106	0.396
Posture	0.096	0.378	Walking	0.105	0.324
Walking	0.089	0.330	Posture	0.091	0.372
Left leg agility	0.086	0.363	Hygiene	0.088	0.244
Speech	0.081	0.366	Left leg agility	0.082	0.346
Rigid LLE	0.078	0.323	Left hand P/S	0.077	0.349
Left hand grip	0.077	0.362	Dressing	0.076	0.293
Rigid neck	0.074	0.330	Rigid neck	0.075	0.305
Freezing	0.074	0.314	Finger tap LH	0.074	0.343

Parameter Optimization

- We fixed the variables in Scanlon's formula and considered optimization of the Mean Absolute Error (MAE).
- Each parameter can take on five values {0,1,2,3,4}.
- Brute force optimization would require 5^{27} ~$7.5x10^{18}$ evaluations.
- We used a Genetic Algorithm with 200 chromosomes and 5000 iterations to explore the parameter space.

Cross-validation

- To avoid the risk of over-fitting, the genetic algorithm optimization was performed in a 10-fold cross validation setting: the dataset consisting of N samples (N = 979 for the male subset, and N = 507 for the female subset) was randomly permuted, and we used 90% of the data to determine the model parameters.
- The process was repeated 10 times, each time randomly permuting the initial dataset and using 90% of the samples.
- In total, we obtained 10 optimized combinations of parameter values (chromosomes) as a result of the 10 repetitions, and each model parameter was assigned to the most frequently-occurring, corresponding value of the parameter in the optimized chromosomes.

Performance

- Scanlon's formulas estimate H&Y with MAE around 0.176 points.
- Optimizing the formulas' parameters using the genetic algorithm leads to statistically significant ($p < 0.001$) improvement
- MAE drops to 0.161, approximately 9% improvement.
- Overall, the formulas mapping UPDRS to H&Y are intuitively attractive, clinically relevant, easily interpretable, and statistically accurate.

Voice Disorders

- Voice disorders affect patients profoundly, and acoustic tools can potentially measure voice function objectively.
- Disordered sustained vowels exhibit wide-ranging phenomena, from nearly periodic to highly complex, aperiodic vibrations, and increased "breathiness".
- Modelling and surrogate data studies have shown significant nonlinear and non-Gaussian random properties in these sounds.

Speech Processing

- Existing tools are limited to analysing voices displaying near periodicity, and do not account for this inherent biophysical nonlinearity and non-Gaussian randomness, often using linear signal processing methods insensitive to these properties.
- They do not directly measure the two main biophysical symptoms of disorder: complex nonlinear aperiodicity, and turbulent, aeroacoustic, non-Gaussian randomness.
- Often these tools cannot be applied to more severe disordered voices, limiting their clinical usefulness.

Hoarseness Diagram

- Hoarseness diagram using two novels features:
- (1) Entropy of near returns from recurrence analysis to measure degree of periodicity
- (2) Fractal index given by detrended fluctuation analysis (DFA) to provide a measure of roughness of the signal
- Achieved classification performance of 91.8 ± 2.0%

MA Little, PE McSharry, S.J. Roberts, DAE Costello and IM Moroz. Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection. BioMedical Engineering OnLine 2007, 6:23

Entropy

- Entropy measures the average uncertainty in the value of the discrete-valued probability density.
- The recurrence probability density entropy (RPDE) for a probability of recurrences times, p(k), with k=1,…,K is given by:

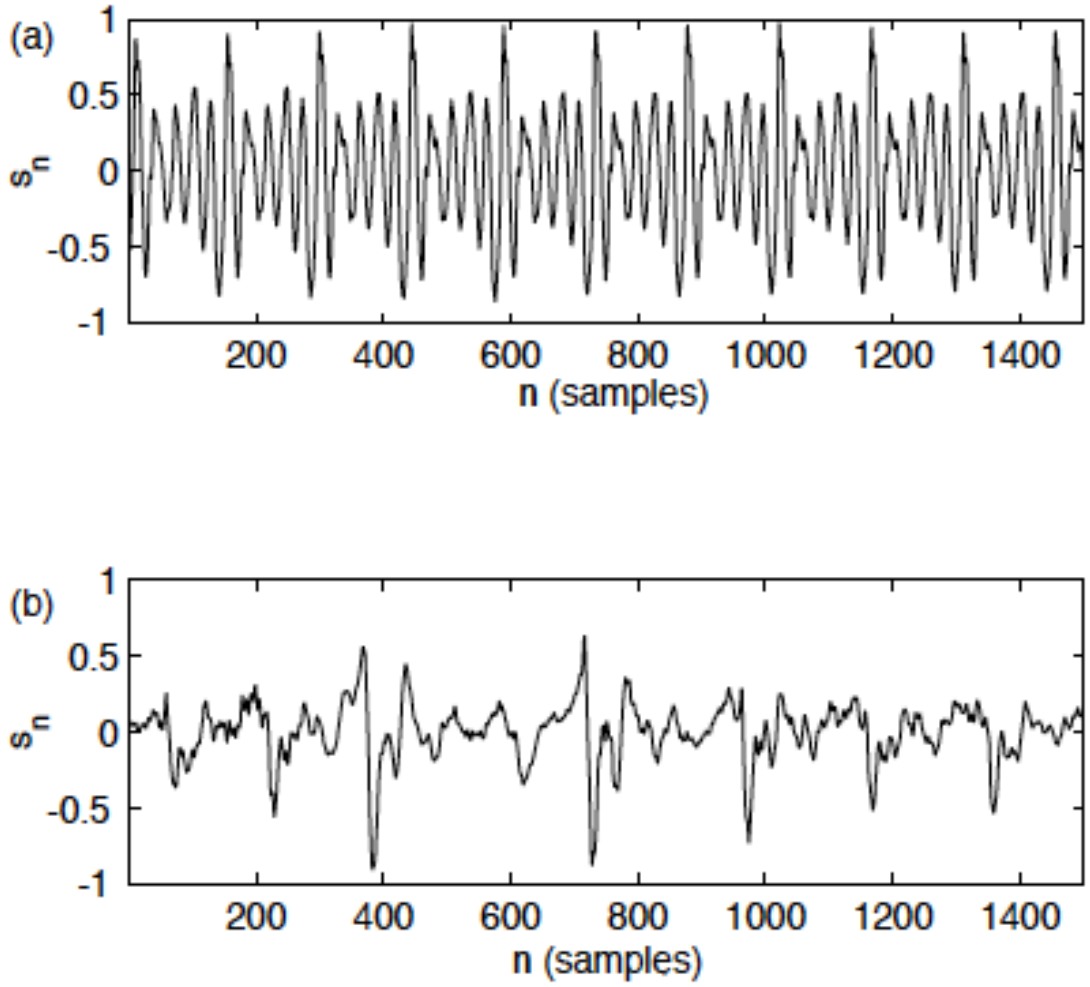
  $H = -\Sigma_k\, p(k) \ln p(k)$.
- For a perfectly periodic signal with p(k)=1 for a single value of k and zero otherwise: $H_{per} = 0$.
- For a random signal with uniform density, $H_{iid} = \ln K$.
- Normalized entropy $H_{norm} = H/H_{iid}$ lies between 0 and 1.

Fractal index

- Detrended Fluctuation Analysis (DFA) characterizes the self-similarity of the graph of a signal from a stochastic process and provides a scaling exponent.
- The logistic function is used to normalize this scaling exponent to provide α_{norm} between 0 and 1 with proximity to 1 being characteristic of general voice disorder:

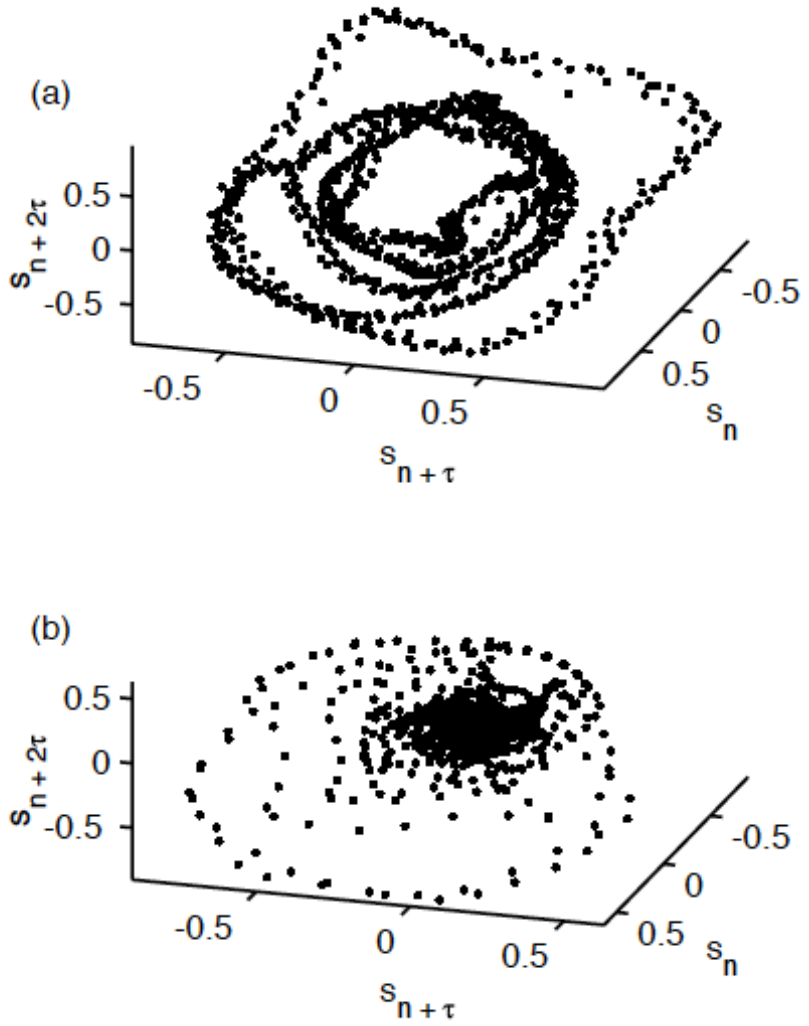
$$\alpha_{norm} = [1 + \exp(-\alpha)]^{-1}$$

Speech Signal Examples



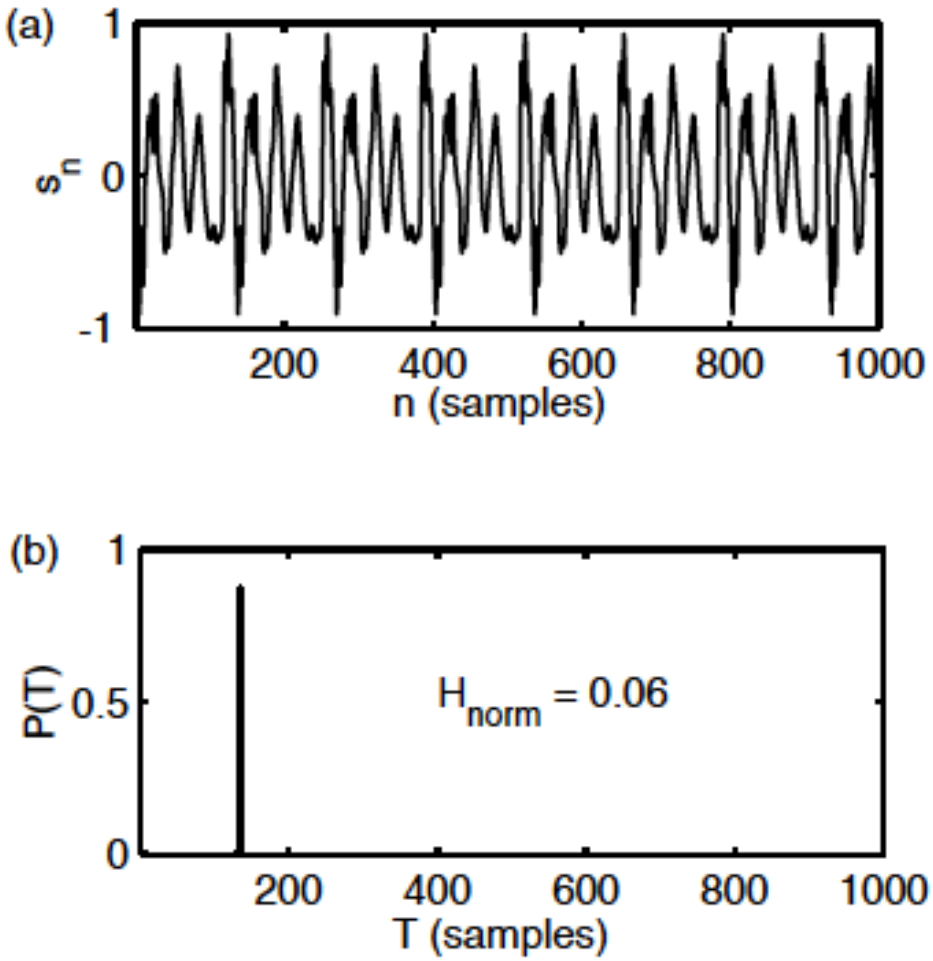
Normal (a) and (b) disordered speech signal examples.

Reconstructed State Space

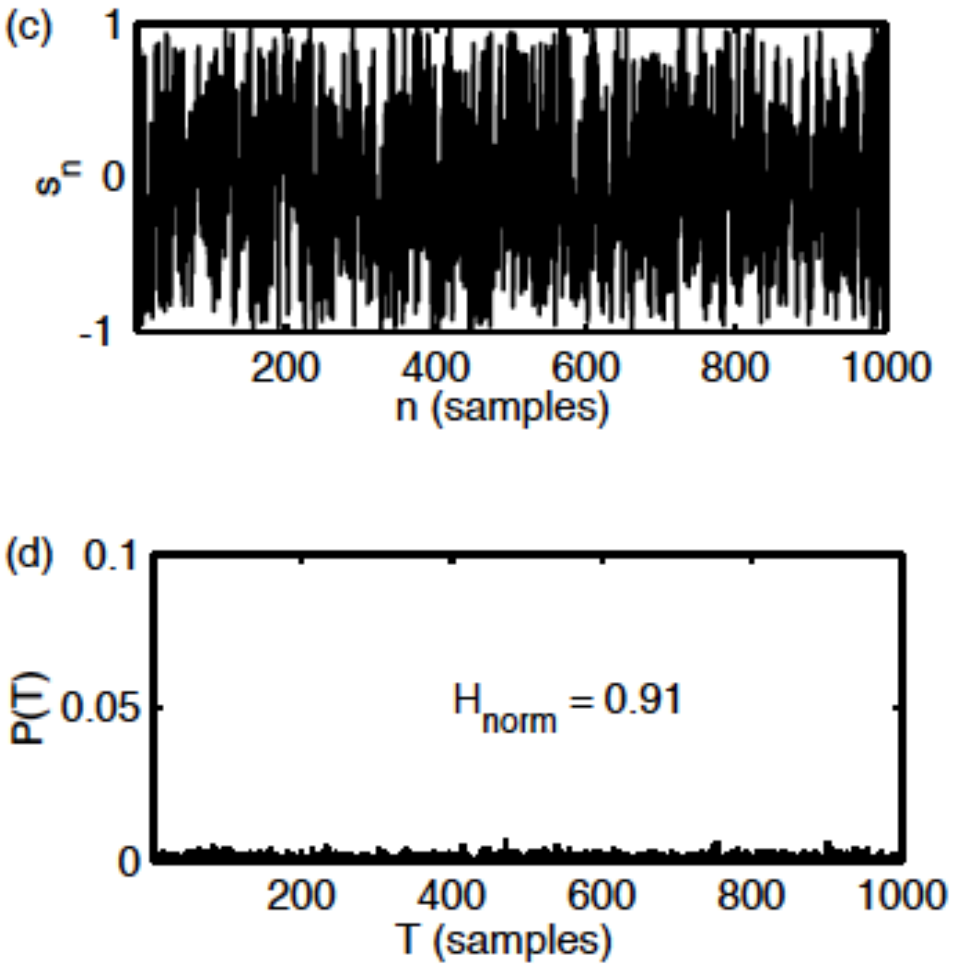


Normal (a) and (b) disordered speech signal examples using reconstruction dimension of *m* = 3 and the time delay is τ = 7 samples.

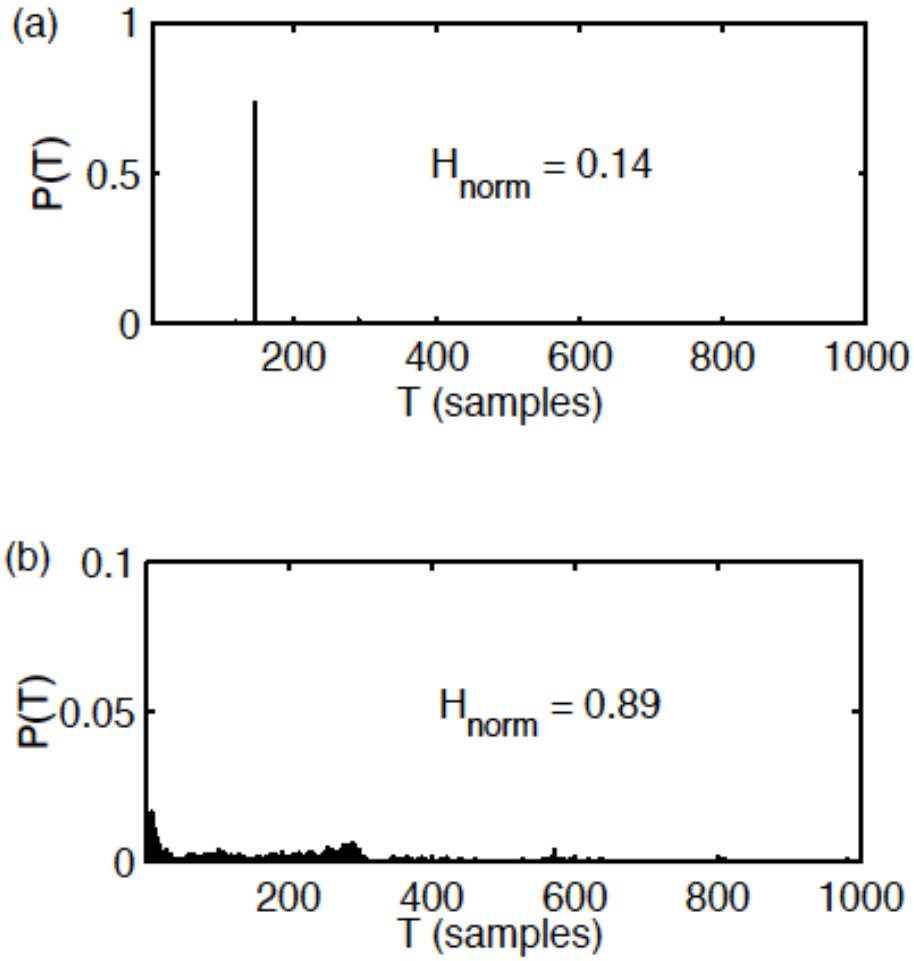
RDPE of periodic signal



RDPE for random signal

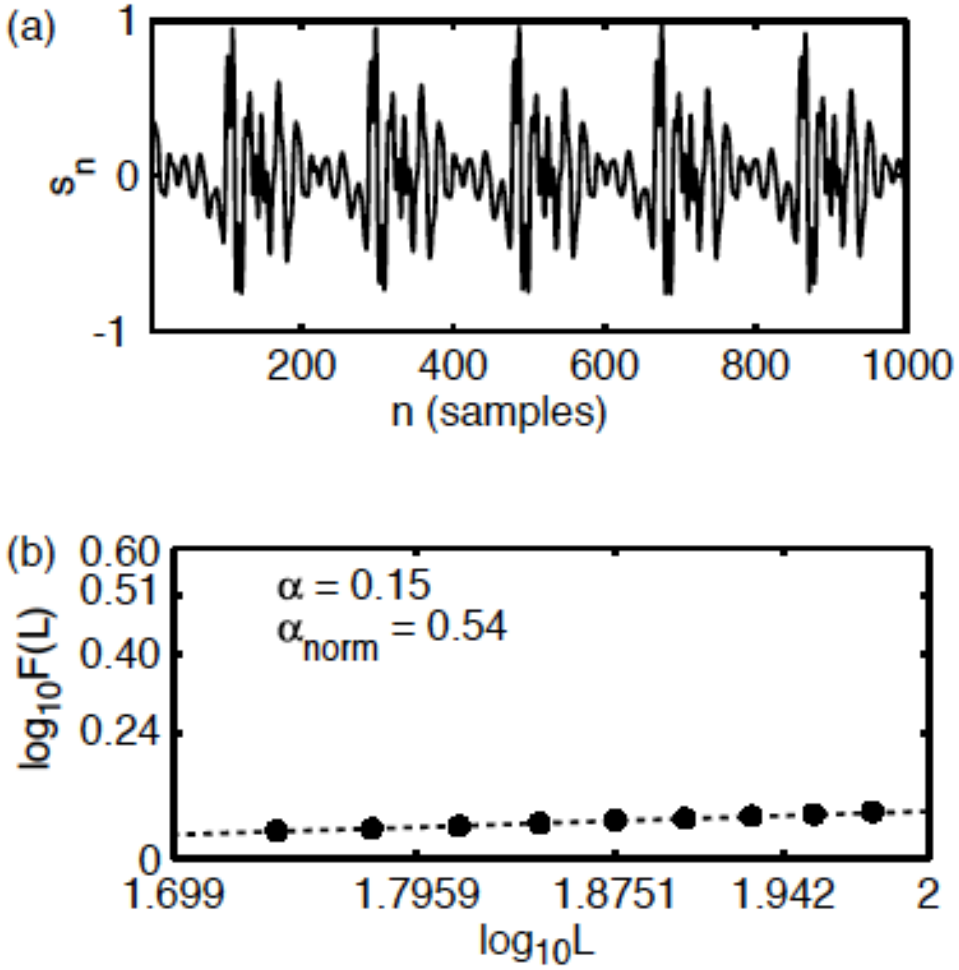


RDPE of speech signals

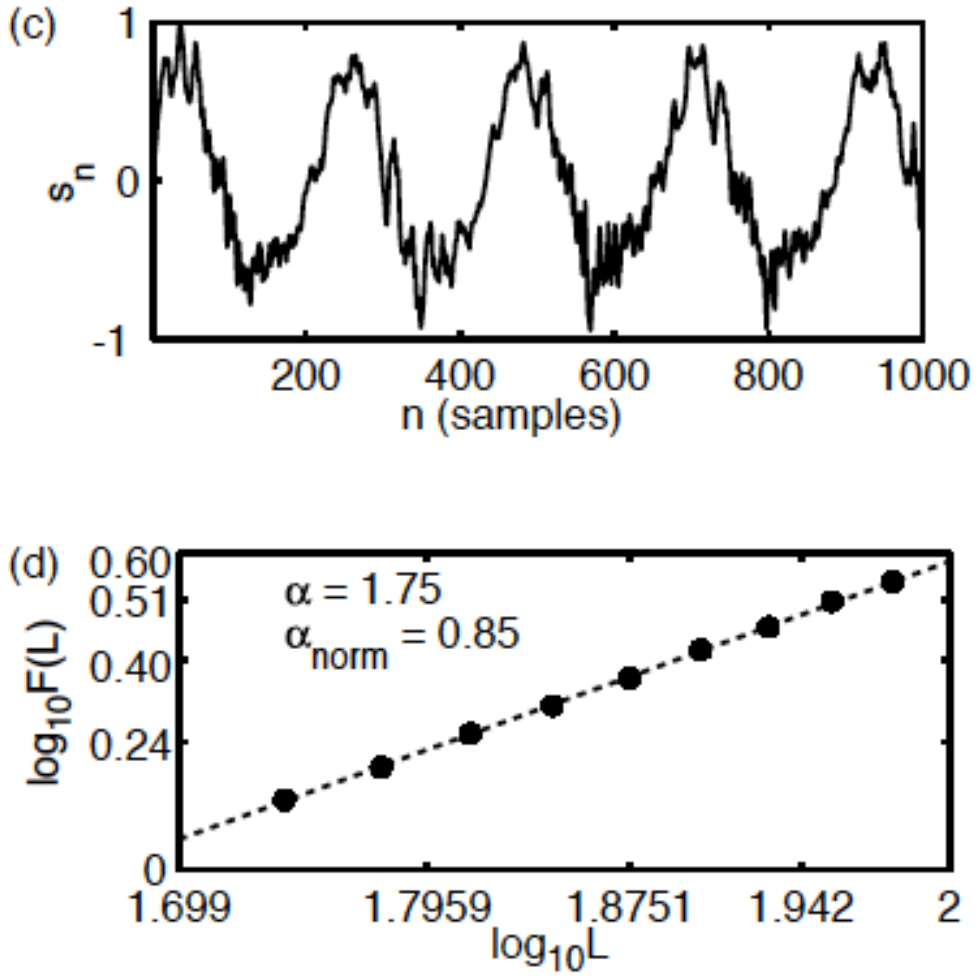


Normal (a) and (b) disordered speech signal examples.

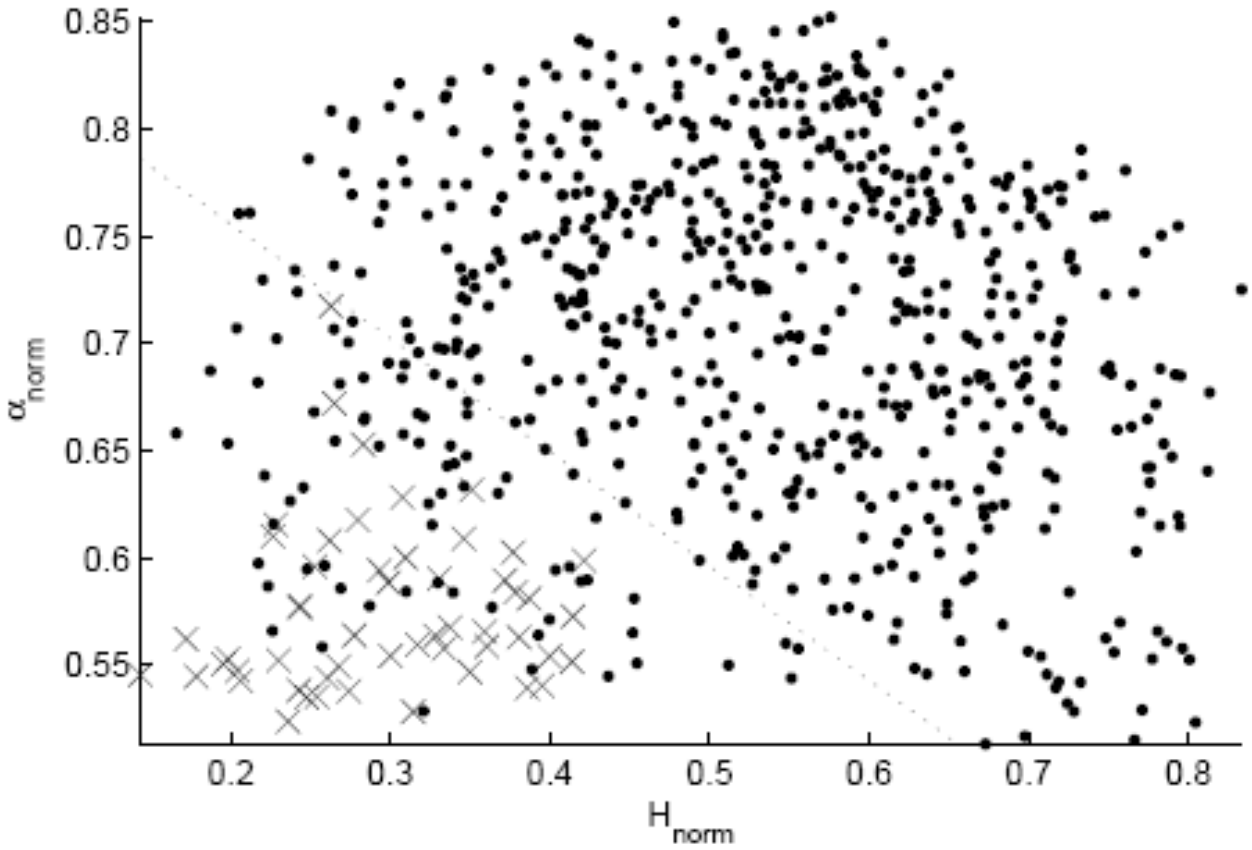
DFA of normal speech



DFA of disordered speech



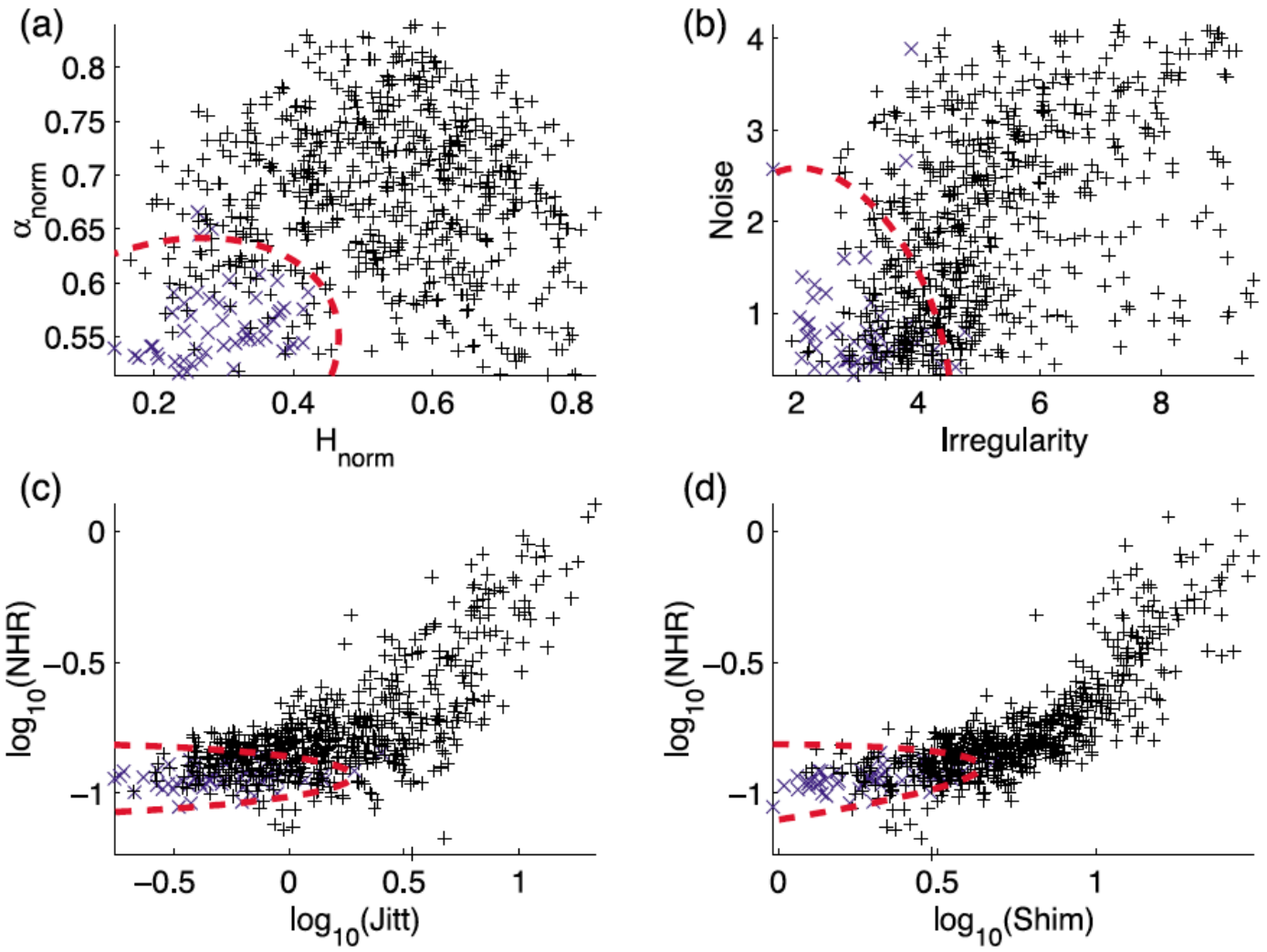
Classification of sustained vowels



Normal (x) and voice disorders (.).

Little, McSharry, Roberts, Costello & Moroz. Biomedical Engineering Online 2007, **6**:23

Classification using QDA



Classification Results

Features	Subjects	Performance
RPDE/DFA	707	91.8 ± 2.0%
Jitt/Shim	685	81.4 ± 3.9%
Shim/NHR	684	80.7 ± 4%
Irreg/Noise	707	79.3 ± 5.5%
Jitt/NHR	684	76.4 ± 4.8%

Performance using our two newly defined features was 91.8% an improvement of 12.8% beyond what was achieved using the best combination of traditional speech features.

Feature engineering has the potential to dramatically improve performance!