Machine Learning for Healthcare: Time-series

David Sontag
Clinical Machine Learning Group
MIT
Time-series in healthcare

In these applications, data is recorded at high frequency and to predict well we need to model trajectories to identify trends and outliers.

Patient with multiple myeloma:
Time-series in healthcare

1. Physiological monitoring
   - Monitoring babies in neonatal ICUs
   - Detecting atrial fibrillation

2. Disease progression
   - Predicting Alzheimer’s disease status
   - Learning about disease from data
     • New subtypes of asthma
     • New staging system for Alzheimer’s
Time-series in healthcare

1. Physiological monitoring
   – Monitoring babies in neonatal ICUs
   – Detecting atrial fibrillation

2. Disease progression
   – Predicting Alzheimer’s disease status
   – Learning about disease from data
     • New subtypes of asthma
     • New staging system for Alzheimer’s
Physiological time-series

Fig. 4. Probes used to collect vital signs data from an infant in intensive care. 1) Three-lead ECG, 2) arterial line (connected to blood pressure transducer), 3) pulse oximeter, 4) core temperature probe (underneath shoulder blades), 5) peripheral temperature probe, 6) transcutaneous probe.

(Quinn et al., TPAMI 2008)
Physiological time-series

• Typical use cases:
  1. Infer true physiological signal from noisy observations
  2. Risk stratification, e.g. predict clinical deterioration, or diagnosis

• Approach taken depends on:
  – Is labeled data available?
  – Do we have a good mechanistic/statistical model?
  – How much training data is there?
Two very different trajectories

(Quinn et al., TPAMI 2008)
Problem: measurements confounded by interventions & measurement errors

(Blood pressure)
- Sys. BP
- Dia. BP

(Oxygen uptake)
- TcPO$_2$
- TcPCO$_2$

BS
TR

Drop out
Blood sample

Transcutaneous probe recalibration

(Quinn et al., TPAMI 2008)
Can we identify the artifactual processes?

- Once identified, can remove for use in downstream predictive tasks (must deal with missing data)
- Can help mitigate *alarm fatigue* by not alerting the clinicians when unnecessary
- More broadly, can we maintain beliefs about the true physiological values of a patient?
(Switching) linear dynamical systems

- Conditioned on $s_t$, linear Gaussian state-space models (Kalman filters):

$$x_t \sim \mathcal{N}\left(\mathbf{A}^{(s_t)}x_{t-1} + \mathbf{d}^{(s_t)}, \mathbf{Q}^{(s_t)}\right)$$

$$y_t \sim \mathcal{N}\left(\mathbf{C}^{(s_t)}x_t, \mathbf{R}^{(s_t)}\right)$$
(Switching) linear dynamical systems

- **Full model:**

  ![Diagram](image.png)

  - **State**
  - **Confounding factors (e.g. artifactual events)**
  - **Observations**
Learning SLDS models

- Assume some labeled training data \( \{s, y\} \)
- True state \( x \) assumed to never be observed
- Learn using expectation maximization
Parameterizing model

- Normal heart rate dynamics are well-modeled using an autoregressive process, e.g.

\[ x_t - b_t \sim \mathcal{N} \left( \sum_{k=1}^{p_1} \alpha_k (x_{t-k} - b_{t-k}), \eta_1 \right) \]

\[ b_t \sim \mathcal{N} \left( \sum_{k=1}^{p_2} \beta_k b_{t-k}, \eta_2 \right), \]

Baseline process (smooth)  
Zero-mean, high frequency  
(Quinn et al., TPAMI 2008)
Parameterizing model

- One can use domain knowledge to specify parts of the artifacts model
  - Probe dropouts modeled by removing dependence of observation \( y_t \) on patient state \( x_t \)
  - Temperature probe disconnection: exponential decay to room temperature

\[
\begin{align*}
\text{State} & : x_{t-1} \rightarrow x_t \rightarrow x_{t+1} \\
\text{Confounding factors (e.g. artifactual events)} & : s_{t-1} \leftarrow s_t \rightarrow s_{t+1} \\
\text{Observations} & : y_{t-1} \rightarrow y_t \rightarrow y_{t+1}
\end{align*}
\]

(Quinn et al., TPAMI 2008)
Evaluation

• 3-fold cross validation, where for each fold train on 10 babies and test on 5
• 24-hours of data for each baby
• Normal dynamics refit for test babies using a 30-minute section near the start

(Quinn et al., TPAMI 2008)
**Evaluation**

GS = Gaussian-sum approximation (used for inference)

RBPF = Rao-Blackwellized particle filtering approximation (used for inference)

FHMM = Factorial HMM (simpler model which does not model normal physiological dynamics)

(Quinn et al., TPAMI 2008)
Inference of physiological state

Blood sample draw

Temperature probe disconnection

(Quinn et al., TPAMI 2008)
Inferred switch settings

TD = core temperature probe disconnection
TR = recalibration
Predicting morbidity in preterm newborns

Saria et al.,
Science Translational Medicine 2010
Can we predict major complications?

• Preterm neonates 34 weeks gestational age or less and <2000 g in weight

• Goal: estimate probability infant would have high morbidity (HM), using data in first 3 hours of life
  – Includes death, sepsis, hemorrhage, pulmonary hypertension, acute hemodynamic instability, and retinopathy of prematurity
  – Outcomes can manifest days or weeks later

• A benefit of using only first 3 hours is that data not typically confounded by medical intervention
  – Models may generalize better across NICUs
APGAR
Test Scoring

A - Appearance
Blue all over | Blue only at extremities | No blue coloration

P - Pulse
No pulse | <100 beats/min. | >100 beats/min.

G - Grimace
No response to stimulation | Grimace or feeble cry when stimulated | Sneezing, coughing, or pulling away when stimulated

A - Activity
No movement | Some movement | Active movement

R - Respiration
No breathing | Weak, slow, or irregular breathing | Strong cry

Figure from: http://www.medicinehack.com/2010/05/apgar-scoring.html
Goal of study

• “Electronic” Apgar score
• Better inform decisions regarding
  – Aggressive use of intensive care
  – Need for transport to tertiary centers
  – Resource allocation (currently $26 billion per year in US spent because of preterm birth)
Machine learning setup

- Binary classification
- Features:
  - Mean heart rate (+ base and residual variability); mean respiratory rate (+base and residual variability); mean oxygen saturation and cumulative hypoxia time
  - Gestational age and birth weight
- 138 preterm neonates (35 with HM complications)
- Leave-one-out cross-validation – no need for nested cross-validation since no hyperparameter tuning

HM = high morbidity
LM = low morbidity
Deriving the features: variability

(Saria et al., Science Translational Medicine 2010)
Prediction using probabilistic model

- L2-regularized logistic regression used to learn predict whether baby will be “high morbidity” (HM):
  \[ P(\text{HM}|v_1,v_2,\ldots,v_n) = \left(1 + \exp(b + w_0c + \sum_{i=1}^{n}w_i f(v_i))\right)^{-1} \]

- Non-linear transformation applied to the features:
  - Estimate \(\Pr(v_i \mid C)\) for each class of patient \(C=\{\text{HM or LM}\}\) using parametric models: exponential, Weibull, lognormal, gamma
  - Use log odds ratio of observed value as feature if observed, 0 if the value is missing:
    \[ f(v_i) = \log \frac{\Pr(v_i \mid \text{HM})}{\Pr(v_i \mid \text{LM})} \]
  - No need to do imputation with this approach!
  - Also use missingness indicators given that it is often informative
Prediction using probabilistic model

Maximum likelihood fit of a log-Normal distribution

Distribution of heart rate variability for patients with HM (high morbidity)

Distribution of heart rate variability for patients with LM (low morbidity)
Mean heart rate

Short-term variability of heart rate

Long-term variability of heart rate

Mean respiratory rate

Short-term variability of respiratory rate

Long-term variability of respiratory rate

Mean oxygen saturation

% of time spent below 85% oxygen saturation

\[ \Pr(HM \mid v_i) \]
the receiver operating characteristic (ROC) curve (Fig. 1A) and associated area under the curve (AUC) values (Table 2) shows that PhysiScore exhibits good discriminative ability for prediction of morbidity and mortality risk and compares it to other risk assessment tools. Specifically, PhysiScore was compared to the Apgar score, long used as an indicator for the base physiological state of the newborn, as well as to extensively validated neonatal scoring systems that require invasive laboratory measurements [Score for Neonatal Acute Physiology-II (SNAP-II), SNAP Perinatal Extension-II (SNAPPE-II), and Clinical Risk Index for Babies (CRIB)]. For making predictions with the Apgar score, we constructed a model as in Eq. 1 using the 1- and 5-min Apgar scores as the only two inputs; this combined model outperformed either of the two Apgar scores when used in isolation. PhysiScore (AUC 0.9197) performed well across the entire range of the ROC curve and significantly better (P < 0.003) than all four of the other comparison scores (Table 2). PhysiScore’s largest performance gain occurred in the high-sensitivity/specificity region of the ROC curve. Setting a user-defined threshold based on desired sensitivity and specificity allows optimization for individual settings. For example, in our neonatal intensive care unit (NICU), a threshold of 0.5 achieves a sensitivity of 86% at a specificity of 95% for HM, as seen in Fig. 1A (inset). Alternatively, the use of a lower threshold would improve sensitivity at the expense of specificity.

We added the values obtained from laboratory tests to determine the magnitude of their contribution to risk prediction beyond the PhysiScore alone (Fig. 1B), incorporating parameters included in standard risk prediction scores (for example, SNAPPE-II): white blood cell count, band neutrophils, hematocrit, platelet count, and initial blood gas measurement of P_aO_2 (partial pressure of oxygen, arterial), P_aCO_2 (partial pressure of carbon dioxide, arterial), and pH (if available at <3 hours of age). No additional discriminatory power was achieved, suggesting that laboratory information is largely redundant with the patient’s physiological characteristics.

To further assess the performance of PhysiScore, we analyzed prediction performance for infants in major morbidity categories. Specifically, we extracted two categories: infection (NEC, culture-positive sepsis, urinary tract infection, etc.) and major cardiopulmonary complications (respiratory distress syndrome, chronic lung disease, etc.).
Feature importance

Error bars = variation over folds in cross-validation
Time-series in healthcare

1. Physiological monitoring
   - Monitoring babies in neonatal ICUs
   - Detecting atrial fibrillation

2. Disease progression
   - Predicting Alzheimer’s disease status
   - Learning about disease from data
     - New subtypes of asthma
     - New staging system for Alzheimer’s
Detecting atrial fibrillation

AliveCore ECG device

ECG = electrocardiogram
What type of heart rhythm?

- Normal rhythm
- AF rhythm
- Other rhythm
- Noisy recording

[Clifford, Liu, Moody, Mark. PhysioNet Computing in Cardiology Challenge 2017]
Abstract

ECG Feature Extraction plays a significant role in diagnosing most of the cardiac diseases. One cardiac cycle in an ECG signal consists of the P-QRS-T waves. This feature extraction scheme determines the amplitudes and intervals in the ECG signal for subsequent analysis. The amplitudes and intervals value of P-QRS-T segment determines the functioning of heart of every human. Recently, numerous research and techniques have been developed for analyzing the ECG signal. The proposed schemes were mostly based on Fuzzy Logic Methods, Artificial Neural Networks (ANN), Genetic Algorithm (GA), Support Vector Machines (SVM), and other Signal Analysis techniques. All these techniques and algorithms have their advantages and limitations. This proposed paper discusses various techniques and transformations proposed earlier in literature for extracting feature from an ECG signal. In addition this paper also provides a comparative study of various methods proposed by researchers in extracting the feature from ECG signal.

Keywords — Artificial Neural Networks (ANN), Cardiac Cycle, ECG signal, Feature Extraction, Fuzzy Logic, Genetic Algorithm (GA), and Support Vector Machines (SVM).

I. INTRODUCTION

The investigation of the ECG has been extensively used for diagnosing many cardiac diseases. The ECG is a realistic record of the direction and magnitude of the electrical commotion that is generated by depolarization and re-polarization of the atria and ventricles. One cardiac cycle in an ECG signal consists of the P-QRS-T waves. Figure 1 shows a sample ECG signal. The majority of the clinically useful information in the ECG is originated in the intervals and amplitudes defined by its features (characteristic wave peaks and time durations). The improvement of precise and rapid methods for automatic ECG feature extraction is of chief importance, particularly for the examination of long recordings [1].

The ECG feature extraction system provides fundamental features (amplitudes and intervals) to be used in subsequent automatic analysis. In recent times, a number of techniques have been proposed to detect these features [2] [3] [4]. The previously proposed method of ECG signal analysis was based on time domain method. But this is not always adequate to study all the features of ECG signals. Therefore the frequency representation of a signal is required. The deviations in the normal electrical patterns indicate various cardiac disorders. Cardiac cells, in the normal state are electrically polarized [5].

In recent year, several research and algorithm have been developed for the exertion of analyzing and classifying the ECG signal. The classifying method which have been proposed during the last decade and under evaluation includes digital signal analysis, Fuzzy Logic methods, Artificial Neural Network, Hidden Markov Model, Genetic Algorithm, Support Vector Machines, Self-Organizing Map, Bayesian and other method with each approach exhibiting its own advantages and disadvantages. This paper provides an overview on various techniques and transformations used for extracting the feature from ECG signal. In addition the future enhancement gives a general idea for improvement and development of the feature extraction techniques.

The remainder of this paper is structured as follows. Section 2 discusses the related work that was earlier proposed in literature for ECG feature extraction. Section 3 gives a general idea of further improvements of the earlier approaches in ECG.
2. Common structure of the QRS detectors.

3. Peak detector proposed in [41].

Fig. 1  Time series showing RR intervals from subject 202 from MIT-BIH arrhythmia database. (——) Assessment of atrial fibrillation (AF) or non-atrial fibrillation (N) as reported in database

[Tateno & Glass, Automatic detection of atrial fibrillation using the coefficient of variation and density histograms of RR and ΔRR intervals. MBEC, 2001]
Cardiac Arrhythmia Classification:
A Heart-Beat Interval-Markov Chain Approach *

WILL GERSCH,† DAVID M. EDDY,‡ AND EUGENE DONG, JR.§

Division of Cardiovascular Surgery, Department of Surgery, Stanford University Medical Center, Stanford, California 94305

Received March 2, 1970

A sequence of heart-beat intervals (R-R wave intervals) is automatically transformed into a three-symbol Markov chain sequence. For convenience the symbols used may be thought of as S-R-L for short, regular, and long heart-beat intervals, respectively. The probability that the observed sequence was generated by each of a set of prototype models characteristic of different cardiac disorders is computed. That prototype corresponding to the largest probability of observed sequence generation is designated as the disorder. This procedure is the equivalent of Kullback's classification by the minimization of directed divergence procedure.

In a preliminary experiment primarily using data sequences of 100 heart-beat intervals, 35 different known cases were automatically classified into six cardiac disorders without error. The disorders considered were atrial fibrillation, APC and VPC, bigeminy, sinus tachycardia with occasional bigeminy, sinus tachycardia, and ventricular tachycardia.

An automatic procedure to classify cardiac arrhythmias using a Markov chain interpretation of heart-beat interval data is reported. A sequence of heart-beat
Detection of Atrial Fibrillation Using Artificial Neural Networks

SG Artis, RG Mark, GB Moody

Harvard-MIT
Division of Health Sciences and Technology, Cambridge, MA

Abstract

Artificial neural networks (ANNs) were used as pattern detectors to detect atrial fibrillation (AF) in the MIT-BIH Arrhythmia Database. ECG data was represented using generalized interval transition matrices, as in Markov model AF detectors[1]. A training file was developed, using these transition matrices, for a backpropagation ANN. This file consisted of approximately 15 minutes each of AF and non-AF data. The ANN was successfully trained using this data. Three standard databases were used to test network performance. Post-processing of the ANN output yielded an AF sensitivity of 92.86% and an AF positive predictive accuracy of 92.34%.

1 Introduction

on R-R interval sequences using a variety of statistical methods [1] but there is room for improvement in these techniques.

Pattern classifiers exist in many forms, and artificial neural networks (ANNs) represent an important subset of these classifiers. ANNs are attractive for solving pattern recognition problems because few assumptions about the underlying data need to be made. The task of the operator of an ANN is to separate the data into subsets. The network will be able classify these subsets according to type as long as they are distinct. Neural network training requires appropriate training data, pre-processing and post-processing algorithms, an appropriate network topology, and a training algorithm, as well as evaluation databases. This document will present the design and evaluation of a technique which detects AF in the presence of other cardiac arrhythmias using a backpropagation artificial neural network.
Winning approach

- Training data in 2017 Physionet challenge: ~8500 ECGs
- Best algorithms use a combination of expert-derived features and machine learning

Abductive reasoning for Atrial Fibrillation identification

Table 1: Set of features used to train the global classifier

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(tSR)</td>
<td>Proportion of the record length interpreted as a regular rhythm (Normal rhythm, tachycardia or bradycardia).</td>
</tr>
<tr>
<td>(t1b)</td>
<td>Number of milliseconds from the beginning of the record to the first interpreted heartbeat.</td>
</tr>
<tr>
<td>(tOR)</td>
<td>Number of milliseconds interpreted as a non-regular rhythm.</td>
</tr>
<tr>
<td>(longTch)</td>
<td>Longest period of time with heart rate over 100bpm.</td>
</tr>
<tr>
<td>(RR)</td>
<td>Median RR interval of regular rhythms.</td>
</tr>
<tr>
<td>(RRd)</td>
<td>Median Absolute Deviation (MAD) of the RR interval in regular rhythms.</td>
</tr>
<tr>
<td>(RR_{MIrr})</td>
<td>Max. RR irregularity measure.</td>
</tr>
<tr>
<td>(PNN(10,50,100))</td>
<td>Global PNNx measures.</td>
</tr>
<tr>
<td>(mRR)</td>
<td>Min. RR interval of regular rhythms.</td>
</tr>
<tr>
<td>(o_PNN50)</td>
<td>PNN50 of non-regular rhythms.</td>
</tr>
<tr>
<td>(n_nP)</td>
<td>Proportion of heartbeats with detected P-wave inside regular rhythms.</td>
</tr>
<tr>
<td>(n_aT)</td>
<td>Median of the amplitude of the T waves inside regular rhythms.</td>
</tr>
<tr>
<td>(n_{PR})</td>
<td>Median PR duration inside regular rhythms.</td>
</tr>
<tr>
<td>(P_{smooth})</td>
<td>Median of the ratio between the standard deviation and the mean value of P-waves' derivative signal.</td>
</tr>
<tr>
<td>(Pdistd)</td>
<td>MAD of the measure given by the P wave delineation method.</td>
</tr>
<tr>
<td>(pw)</td>
<td>Profile of the full signal.</td>
</tr>
<tr>
<td>(xcorr)</td>
<td>Median of the maximum cross-correlation between QRS complexes interpreted in regular rhythms.</td>
</tr>
<tr>
<td>(PRd)</td>
<td>Global MAD of the PR durations.</td>
</tr>
<tr>
<td>(TP)</td>
<td>Median of the prevailing frequency in the TP intervals.</td>
</tr>
<tr>
<td>(pw_{prof})</td>
<td>Profile measure of the signal in the P-wave area.</td>
</tr>
<tr>
<td>(n_Txcorr)</td>
<td>Median of the maximum cross-correlation between T-waves inside regular rhythms.</td>
</tr>
<tr>
<td>(baseline)</td>
<td>Profile of the baseline in regular rhythms.</td>
</tr>
<tr>
<td>(wQRS)</td>
<td>Proportion of wide QRS complexes (duration longer than 110ms).</td>
</tr>
<tr>
<td>(wQRS_{prof})</td>
<td>Median of the signal profile in the 300ms before each wide QRS complex.</td>
</tr>
<tr>
<td>(x_{xc})</td>
<td>Median of the maximum cross-correlation between ectopic beats.</td>
</tr>
</tbody>
</table>

Not enough data for deep learning? Wrong architectures?

“However, the fact that a standard random forest with well chosen features performed as well as more complex approaches, indicates that perhaps a set of 8,528 training patterns was not enough to give the more complex approaches an advantage. With so many parameters and hyperparameters to tune, the search space can be enormous and significant overtraining was seen...”

[Clifford et al. AF Classification from a Short Single Lead ECG Recording: the PhysioNet/Computing in Cardiology Challenge, Computing in Cardiology 2017]
Cardiologist-Level Arrhythmia Detection With Convolutional Neural Networks

Pranav Rajpurkar*, Awni Hannun*, Masoumeh Haghpanahi, Codie Bourn, and Andrew Ng

A collaboration between Stanford University and iRhythm Technologies

We develop a model which can diagnose irregular heart rhythms, also known as arrhythmias, from single-lead ECG signals better than a cardiologist.

Key to exceeding expert performance is a deep convolutional network which can map a sequence of ECG samples to a sequence of arrhythmia annotations along with a novel dataset two orders of magnitude larger than previous datasets of its kind.
Differences with previous work

• Sensor is a Zio patch – conceivably much less noisy:

• ~60K ECG records annotated (from ~30K patients)

• Identify 12 heart arrhythmias, sinus rhythm and noise for a total of 14 output classes
<table>
<thead>
<tr>
<th>Class</th>
<th>Description</th>
<th>Example</th>
<th>Train + Val Patients</th>
<th>Test Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFIB</td>
<td>Atrial Fibrillation</td>
<td></td>
<td>4638</td>
<td>44</td>
</tr>
<tr>
<td>AFL</td>
<td>Atrial Flutter</td>
<td></td>
<td>3805</td>
<td>20</td>
</tr>
<tr>
<td>AVB_TYPE2</td>
<td>Second degree AV Block Type 2 (Mobitz II)</td>
<td></td>
<td>1905</td>
<td>28</td>
</tr>
<tr>
<td>BIGEMINY</td>
<td>Ventricular Bigeminy</td>
<td></td>
<td>2855</td>
<td>22</td>
</tr>
<tr>
<td>CHB</td>
<td>Complete Heart Block</td>
<td></td>
<td>843</td>
<td>26</td>
</tr>
<tr>
<td>EAR</td>
<td>Ectopic Atrial Rhythm</td>
<td></td>
<td>2623</td>
<td>22</td>
</tr>
<tr>
<td>IVR</td>
<td>Idioventricular Rhythm</td>
<td></td>
<td>1962</td>
<td>34</td>
</tr>
<tr>
<td>Class</td>
<td>Description</td>
<td>Example</td>
<td>Train + Val Patients</td>
<td>Test Patients</td>
</tr>
<tr>
<td>----------------------</td>
<td>--------------------------------------</td>
<td>---------------</td>
<td>----------------------</td>
<td>---------------</td>
</tr>
<tr>
<td>JUNCTIONAL</td>
<td>Junctional Rhythm</td>
<td></td>
<td>2030</td>
<td>36</td>
</tr>
<tr>
<td>NOISE</td>
<td>Noise</td>
<td></td>
<td>9940</td>
<td>41</td>
</tr>
<tr>
<td>SINUS</td>
<td>Sinus Rhythm</td>
<td></td>
<td>22156</td>
<td>215</td>
</tr>
<tr>
<td>SVT</td>
<td>Supraventricular Tachycardia</td>
<td></td>
<td>6301</td>
<td>34</td>
</tr>
<tr>
<td>TRIGEMINY</td>
<td>Ventricular Trigeminy</td>
<td></td>
<td>2864</td>
<td>21</td>
</tr>
<tr>
<td>VT</td>
<td>Ventricular Tachycardia</td>
<td></td>
<td>4827</td>
<td>17</td>
</tr>
<tr>
<td>WENCKEBACH</td>
<td>Wenckebach (Mobitz I)</td>
<td></td>
<td>2051</td>
<td>29</td>
</tr>
</tbody>
</table>
Deep convolutional network

- 1-D signal sampled at 200Hz, labeled at 1 sec intervals
- 34 layers
- Shortcut connections (ala residual networks) with max-pooling
- Subsampled every other layer ($2^8$ in total)

[Rajpurkar et al., arXiv:1707.01836, 2017]
Example of 1D convolution

\[
<1,0,1> \ast <2,3,1> = 1 \times 2 + 0 \times 3 + 1 \times 1 = 3.
\]

Input

Filter

Output
## Evaluation

<table>
<thead>
<tr>
<th>Class-level F1 Score</th>
<th>Seq</th>
<th></th>
<th>Set</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model</td>
<td>Cardiol.</td>
<td>Model</td>
<td>Cardiol.</td>
</tr>
<tr>
<td>AFIB</td>
<td>0.604</td>
<td>0.515</td>
<td>0.667</td>
<td>0.544</td>
</tr>
<tr>
<td>AFL</td>
<td>0.687</td>
<td>0.635</td>
<td>0.679</td>
<td>0.646</td>
</tr>
<tr>
<td>AVB_TYPE2</td>
<td>0.689</td>
<td>0.535</td>
<td>0.656</td>
<td>0.529</td>
</tr>
<tr>
<td>BIGEMINY</td>
<td>0.897</td>
<td>0.837</td>
<td>0.870</td>
<td>0.849</td>
</tr>
<tr>
<td>CHB</td>
<td>0.843</td>
<td>0.701</td>
<td>0.852</td>
<td>0.685</td>
</tr>
<tr>
<td>EAR</td>
<td>0.519</td>
<td>0.476</td>
<td>0.571</td>
<td>0.529</td>
</tr>
<tr>
<td>IVR</td>
<td>0.761</td>
<td>0.632</td>
<td>0.774</td>
<td>0.720</td>
</tr>
<tr>
<td>JUNCTIONAL</td>
<td>0.670</td>
<td>0.684</td>
<td>0.783</td>
<td>0.674</td>
</tr>
<tr>
<td>NOISE</td>
<td>0.823</td>
<td>0.768</td>
<td>0.704</td>
<td>0.689</td>
</tr>
<tr>
<td>SINUS</td>
<td>0.879</td>
<td>0.847</td>
<td>0.939</td>
<td>0.907</td>
</tr>
<tr>
<td>SVT</td>
<td>0.477</td>
<td>0.449</td>
<td>0.658</td>
<td>0.556</td>
</tr>
<tr>
<td>TRIGEMINY</td>
<td>0.908</td>
<td>0.843</td>
<td>0.870</td>
<td>0.816</td>
</tr>
<tr>
<td>VT</td>
<td>0.506</td>
<td>0.566</td>
<td>0.694</td>
<td>0.769</td>
</tr>
<tr>
<td>WENCKEBACH</td>
<td>0.709</td>
<td>0.593</td>
<td>0.806</td>
<td>0.736</td>
</tr>
</tbody>
</table>

### Aggregate Results

<table>
<thead>
<tr>
<th></th>
<th>Seq</th>
<th></th>
<th>Set</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision (PPV)</td>
<td>0.800</td>
<td>0.723</td>
<td>0.809</td>
<td>0.763</td>
</tr>
<tr>
<td>Recall (Sensitivity)</td>
<td>0.784</td>
<td>0.724</td>
<td>0.827</td>
<td>0.744</td>
</tr>
<tr>
<td>F1</td>
<td>0.776</td>
<td>0.719</td>
<td>0.809</td>
<td>0.751</td>
</tr>
</tbody>
</table>
probability distributions have been applied to the task of classification, Hidden Markov models with Gaussian observation. Drawing inspiration from automatic speech recognition is a common sub-problem of heart-arrhythmia classification. Much work has been done to automate the extraction of features pertaining to the QRS morphology are needed. However, such features alone are not sufficient to distinguish between most heart arrhythmias since features based on wavelet transformations to compute features from the raw ECG followed by finely-tuned threshold based classification, including detecting tachycardias (fast heart rate), bradycardias (slow heart rate), and irregular rhythms. Algorithms are often used for coarse-grained heart rhythm classification, including detecting tachycardias (fast heart rate), bradycardias (slow heart rate), and irregular rhythms. Accurate estimates of heart rate and heart rate variability can be extracted from R-peak features, feature-engineered algorithms are the most common approach.

Figure 4. A confusion matrix for the model predictions on the test set. Many of the mistakes the model makes are not surprising. For example, confusing second degree A V Block (Type 2) with Wenckebach makes sense given the often similar expression of these rhythms. For example, confusing second degree A V Block (Type 2) with Wenckebach makes sense given the often similar expression of these rhythms. The most common dataset used to design and evaluate ECG rhythm recognition is the MIT-BIH arrhythmia database which consists of 48 half-hour strips of ECG data. Other commonly used datasets include the MIT-BIH Atrial Fibrillation dataset and Sinus Rhythm on the MIT-BIH dataset. While the neurovascular network can distinguish between these two classes with high-accuracy, it does not generalize to noisier single-lead recordings or classify among the full range of human agreement rates when large annotated datasets are available. Machine learning models based on deep neural networks trained on these models are tractable. Many of the mistakes the model makes are not surprising. For example, confusing second degree A V Block (Type 2) with Wenckebach makes sense given the often similar expression of these rhythms. The most common dataset used to design and evaluate ECG rhythm recognition is the MIT-BIH arrhythmia database which consists of 48 half-hour strips of ECG data. Other commonly used datasets include the MIT-BIH Atrial Fibrillation dataset and Sinus Rhythm on the MIT-BIH dataset. While the neurovascular network can distinguish between these two classes with high-accuracy, it does not generalize to noisier single-lead recordings or classify among the full range of human agreement rates when large annotated datasets are available. Machine learning models based on deep neural networks trained on these models are tractable.
Time-series in healthcare

1. Physiological monitoring
   - Monitoring babies in neonatal ICUs
   - Detecting atrial fibrillation

2. Disease progression
   - Predicting Alzheimer’s disease status
   - Learning about disease from data
     • New subtypes of asthma
     • New staging system for Alzheimer’s
Nature of chronic diseases

- Disease starts
- Complications start
- Symptoms
- Diagnosis
- Disease severity

-10 - 5 - 0

Medical intervention and treatments

[Image credit: Farzad Kamalzadeh]
Disease progression modeling: predictive (supervised)

• When will the disease progress, e.g. from chronic kidney disease stage 3 to 4?
• What is the future trajectory of a biomarker?
• When will complications arise, what types and how severe?
• How will a patient respond to a treatment?
Predicting disease progression in Alzheimer’s disease

[Image credit: Wikipedia; "Alzheimer's Disease Education and Referral Center, a service of the National Institute on Aging."]
### MINI MENTAL STATE EXAMINATION (MMSE)

One point for each answer

<table>
<thead>
<tr>
<th>DATE:</th>
<th>5</th>
<th>5</th>
<th>5</th>
</tr>
</thead>
</table>

#### ORIENTATION
- **Year**, **Season**, **Month**, **Date**, **Time**
- **Country**, **Town**, **District**, **Hospital**, **Ward/Floor**

#### REGISTRATION
Examiner names three objects (e.g. apple, table, penny) and asks the patient to repeat (1 point for each correct. THEN the patient learns the 3 names repeating until correct).

#### ATTENTION AND CALCULATION
Subtract 7 from 100, then repeat from result. Continue five times: 100, 93, 86, 79, 65. (Alternative: spell “WORLD” backwards: DLROW).

#### RECALL
- Ask for the names of the three objects learned earlier.

#### LANGUAGE
- Name two objects (e.g. pen, watch).
- Repeat “No ifs, ands, or buts”.
- Give a three-stage command. Score 1 for each stage. (e.g. “Place index finger of right hand on your nose and then on your left ear”).
- Ask the patient to read and obey a written command on a piece of paper. The written instruction is: “Close your eyes”.
- Ask the patient to write a sentence. Score 1 if it is sensible and has a subject and a verb.

#### COPYING: Ask the patient to copy a pair of intersecting pentagons

<table>
<thead>
<tr>
<th>TOTAL:</th>
<th>30</th>
<th>30</th>
<th>30</th>
</tr>
</thead>
</table>

**MMSE scoring**
- 24-30: no cognitive impairment
- 18-23: mild cognitive impairment
- 0-17: severe cognitive impairment
Predicting disease progression in Alzheimer’s disease

• Goal: Predict disease status in 6, 12, 24, 36, and 48 months
• Five different regression tasks?
• Challenge: data sparsity
  – Total number of patients is small
  – Labels are noisy
  – Due to censoring, fewer patients at later time points

[Zhou et al., KDD ’12]
Predicting disease progression in Alzheimer’s disease

- Goal: Predict disease status in 6, 12, 24, 36, and 48 months
- Five different regression tasks?
- Challenge: data sparsity

<table>
<thead>
<tr>
<th></th>
<th>M06</th>
<th>M12</th>
<th>M24</th>
<th>M36</th>
<th>M48</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of patients</td>
<td>648</td>
<td>642</td>
<td>569</td>
<td>389</td>
<td>87</td>
</tr>
</tbody>
</table>

M06 = 6 months after baseline

[Zhou et al., KDD ’12]
Multi-task learning

• Goal: Predict disease status in 6, 12, 24, 36, and 48 months

• Rather than learn several independent models, view as multi-task learning
  – Select common set of biomarkers for all time points
  – Also allow for specific set of biomarkers at different time points
  – Incorporate temporal smoothness in models

[Zhou et al., KDD ’12]
Convex fused sparse group lasso

• Simultaneously learn all 5 models by solving the following convex optimization problem:

\[
\min_W L(W) + \lambda_1 \|W\|_1 + \lambda_2 \left\| RW^T \right\|_1 + \lambda_3 \|W\|_{2,1}
\]

• Squared loss: \( L(W) = \| S \odot (XW - Y) \|_F^2 \) 
  (\( S \) is a mask to account for labels missing in subset of tasks)

• Group Lasso penalty \( \|W\|_{2,1} \) given by \( \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{t} W_{ij}^2} \)

• \( R = \begin{bmatrix}
1 & -1 & 1 & -1 & 1 \\
1 & -1 & 1 & -1 & 1 \\
1 & -1 & 1 & -1 & 1
\end{bmatrix} \)

[Zhou et al., KDD ’12]
Features

MRI scans (white matter parcellation volume, etc.) +

<table>
<thead>
<tr>
<th>Demographic</th>
<th>age, years of education, gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genetic</td>
<td>ApoE-ɛ4 information</td>
</tr>
<tr>
<td>Baseline cognitive scores</td>
<td>MMSE, ADAS-Cog, ADAS-MOD, ADAS subscores, CDR, FAQ, GDS, Hachinski, Neuropsychological Battery, WMS-R Logical Memory</td>
</tr>
<tr>
<td>Lab tests</td>
<td>RCT1, RCT11, RCT12, RCT13, RCT14, RCT1407, RCT1408, RCT183, RCT19, RCT20, RCT29, RCT3, RCT392, RCT4, RCT5, RCT6, RCT8</td>
</tr>
</tbody>
</table>

371 in total

[Zhou et al., KDD ’12]
Results (averaged over 5 time points)

<table>
<thead>
<tr>
<th></th>
<th>Baseline – independent regressors</th>
<th>Temporal smoothing helps!</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ridge</td>
<td></td>
</tr>
<tr>
<td>nMSE</td>
<td>0.548 ± 0.057</td>
<td>0.428 ± 0.052</td>
</tr>
<tr>
<td>R</td>
<td>0.689 ± 0.030</td>
<td>0.772 ± 0.030</td>
</tr>
<tr>
<td></td>
<td>cFSGL1</td>
<td>0.400 ± 0.053</td>
</tr>
<tr>
<td></td>
<td>cFSGL2</td>
<td>0.790 ± 0.032</td>
</tr>
<tr>
<td></td>
<td>cFSGL3</td>
<td>0.796 ± 0.031</td>
</tr>
<tr>
<td>Target: MMSE</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

nMSE – normalized mean squared error. Smaller is better
R – average $R^2$ (correlation coefficient). Larger is better

$$\min_W L(W) + \lambda_1 \|W\|_1 + \lambda_2 \left\| RW^T \right\|_1 + \lambda_3 \|W\|_{2,1}$$
is a potential factor that contributes to the lower predictive performance of MMSE than that of ADAS-Cog in our study. 90 percent of data is used as training data. TGL) on longitudinal MMSE and ADAS-Cog prediction using MRI+META features (M+E) in terms of Table 3: Comparison of our proposed approaches (cFSGL and nFSGL) and existing approaches (Ridge and M36 MSE). ADAS-Cog may be a better cognitive measurement for long-
Time-series in healthcare

1. Physiological monitoring
   - Monitoring babies in neonatal ICUs
   - Detecting atrial fibrillation

2. Disease progression
   - Predicting Alzheimer’s disease status
   - Learning about disease from data
     • New subtypes of asthma
     • New staging system for Alzheimer’s
Disease progression modeling: descriptive (unsupervised)

- Find markers of disease stage and progression, statistics of what to expect when
- Discover new disease subtypes
- Key challenges:
  - Never directly observe disease stage, but rather only indirect observations (e.g. symptoms)
  - Data often observes patients only at *one* stage in their disease, not beginning to end
In 1-D, might assume that low values correspond to an early disease stage (or vice-versa)

Assume samples were all taken today
What about in higher dimensions?
What about in higher dimensions?

Insight #1: with enough data, may be possible to recognize structure

[Bendall et al., Cell 2014 (human B cell development)]
What about in higher dimensions?

Insight #2: sequential observations from the same patient can also help.

Each color is a different patient.
What about in higher dimensions?

Biomarker A

Biomarker B

Early disease

Late disease
May also seek to discover disease subtypes
Summary – two tasks

1. Find subtypes (clustering)
2. Sort into early-late disease or severity, i.e. discover the trajectory
Time-series in healthcare

1. Physiological monitoring
   – Monitoring babies in neonatal ICUs
   – Detecting atrial fibrillation

2. Disease progression
   – Predicting Alzheimer’s disease status
   – Learning about disease from data
     • New subtypes of asthma
     • New staging system for Alzheimer’s
K-Means

- An iterative clustering algorithm
  - Initialize: Pick $K$ random points as cluster centers
  - Alternate:
    1. Assign data points to closest cluster center
    2. Change the cluster center to the average of its assigned points
  - Stop when no points’ assignments change
K-means clustering: Example

- Pick $K$ random points as cluster centers (means)

Shown here for $K=2$
K-means clustering: Example

Iterative Step 1
- Assign data points to closest cluster center
K-means clustering: Example

Iterative Step 2
- Change the cluster center to the average of the assigned points
K-means clustering: Example

- Repeat until convergence
Asthma: the problem

• 5 to 10% of people with severe asthma remain poorly controlled despite maximal inhaled therapy

[Holgate ST, Polosa R. The mechanisms, diagnosis, and management of severe asthma in adults. Lancet. 2006; 368:780–793]
Asthma: the question

“It is now recognised that there are distinct asthma phenotypes and that distinct therapeutic approaches may only impinge on some aspects of the disease process within each subgroup”

• What are the processes (genetic or environmental) that underlie different subtypes of asthma?
• Which aspects of airway remodelling are important in disease subtypes?
• What are the best biomarkers of disease progression or treatment response?
• Why are some patients less responsive to conventional therapies than others?

[Adcock et al., “New targets for drug development in asthma”. The Lancet, 2008]
Discovering subtypes from data

[Primary Care Asthma]
Concordant Disease
Symptom-based approach to therapy titration may be sufficient.

[Secondary Care Asthma]

Discordant Symptoms
Monitoring inflammation allows down-titration of corticosteroids.

EARLY SYMPTOM PREDOMINANT
Early onset, atopic.
Normal BMI.
High symptom expression.

OBESE NON-EOSINOPHILIC
Later onset, female preponderance.
High symptom expression.

EARLY ONSET ATOPIC ASTHMA
Concordant symptoms, inflammation & airway dysfunction.

BENIGN ASThma
Mixed middle-aged cohort
Well controlled symptoms and inflammation. Benign prognosis.

INFLAMMATION PREDOMINANT
Late onset, greater proportion of males.
Few daily symptoms but active eosinophilic inflammation.

Monitoring inflammation allows targeted corticosteroids to lower exacerbation frequency.

[Haldar et al., Am J Respir Crit Care Med, 2008]
The data

- All patients had physician diagnosis of asthma and at least one recent prescription for asthma therapy
- All were current nonsmokers
- **Data set #1**: 184 patients recruited from primary-care practices in the UK
- **Data set #2**: 187 patients from refractory asthma clinic in the UK
- **Data set #3**: 68 patients from 12 month clinical study
- Features: z scores for continuous variables, 0/1 for categorical
  - Some of the continuous variables log-transformed to approximate a normal distribution

[Haldar et al., *Am J Respir Crit Care Med*, 2008]
Comparison of Baseline Characteristics in the three Asthma Populations

<table>
<thead>
<tr>
<th>Variable</th>
<th>Primary Care ($n = 184$)</th>
<th>Secondary Care ($n = 187$)</th>
<th>Longitudinal Cohort ($n = 68$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex, % female</td>
<td>54.4</td>
<td>65.8</td>
<td>47.1</td>
</tr>
<tr>
<td>Age, yr (SD)</td>
<td>49.2 (13.9)</td>
<td>43.4 (15.9)</td>
<td>52.4 (14.6)</td>
</tr>
<tr>
<td>Age of onset, yr (SD)</td>
<td>24.7 (19)</td>
<td>20.3 (18.4)</td>
<td>31.1 (23.7)</td>
</tr>
<tr>
<td>Atopic status, % positive</td>
<td>72.8</td>
<td>73.8</td>
<td>57.4</td>
</tr>
<tr>
<td>Body mass index, kg/m² (SD)</td>
<td>27.5 (5.4)</td>
<td>28.5 (6.5)</td>
<td>28.0 (5.9)</td>
</tr>
<tr>
<td>PC$_{20}$ methacholine †, mg/ml</td>
<td>1.04 (1.13)</td>
<td>†</td>
<td>0.67 (0.68)</td>
</tr>
<tr>
<td>Peak flow variability, amp % mean</td>
<td>17 (0.38)</td>
<td>32.2 (0.48)</td>
<td>13.8 (0.29)</td>
</tr>
<tr>
<td>FEV$_1$ change with bronchodilator, %</td>
<td>1.63 (1.16)</td>
<td>12.8 (0.41)</td>
<td>3.2 (1.04)</td>
</tr>
<tr>
<td>Post-bronchodilator FEV$_1$, % predicted</td>
<td>91.4 (21)</td>
<td>82.1 (21.1)</td>
<td>80.2 (20.6)</td>
</tr>
<tr>
<td>Sputum eosinophil count, %</td>
<td>1.32 (0.62)</td>
<td>2.9 (0.99)</td>
<td>2.4 (0.81)</td>
</tr>
<tr>
<td>FE$_{NO}$ ‡, ppb</td>
<td>31.6 (0.33)</td>
<td>43 (0.32)</td>
<td>4.32 (0.64)‡</td>
</tr>
<tr>
<td>Sputum neutrophil count, %</td>
<td>55.09 (0.31)</td>
<td>46.7 (0.32)</td>
<td>41.1 (0.35)</td>
</tr>
<tr>
<td>Modified JACS § (SD)</td>
<td>1.36 (0.74)</td>
<td>2.02 (1.16)</td>
<td>1.42 (1.26)</td>
</tr>
<tr>
<td>Dose of inhaled corticosteroid, BDP equivalent/µg (SD)</td>
<td>632 (579)</td>
<td>1,018 (539)</td>
<td>1,821 (1,239)</td>
</tr>
<tr>
<td>Long-acting bronchodilator use, %</td>
<td>40.2</td>
<td>93</td>
<td>86.7</td>
</tr>
</tbody>
</table>

*Significance figures are derived using one-way analysis of variance between the three populations for continuous variables or χ² test for proportions.

†Bronchial challenge testing is not routinely performed in secondary care for refractory asthma. The comparison given is between the primary-care asthma population and the longitudinal study cohort.

‡FE$_{NO}$ was measured using the NIOX (Aerocrine, Solna, Sweden) analyzer at 50 ml/second in the primary-care population and secondary-care population. The Logan (Logan Research, Ltd., Rochester, Kent, UK) analyzer was used at a flow rate of 250 ml/second in the longitudinal study cohort. A strong linear correlation of 0.97 exists between the two measurement protocols. The statistical comparison is between Fe$_{NO}$ levels in primary and secondary care using NIOX.

§The Juniper Asthma Control Score, modified to include the symptom domains only (see the online supplement).

**Definition of abbreviations:** amp = amplitude; BDP = beclomethasone dipropionate; JACS = Juniper Asthma Control Score

[Haldar et al., *Am J Respir Crit Care Med*, 2008]
### TABLE 2

Clusters in Primary Care (found by K-means)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Primary Care (n = 184)</th>
<th>Early-Onset Atopic Asthma (n = 61)</th>
<th>Cluster 2 Obese Noneosinophilic (n = 27)</th>
<th>Cluster 3 Benign Asthma (n = 96)</th>
<th>Significance (P Value)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex †, % female</td>
<td>54.4</td>
<td>45.9</td>
<td>81.5</td>
<td>52.1</td>
<td>0.006</td>
</tr>
<tr>
<td>Age, yr (SD)</td>
<td>49.2 (13.9)</td>
<td>44.5 (14.3)</td>
<td>53.9 (14)</td>
<td>50.8 (13)</td>
<td>0.003</td>
</tr>
<tr>
<td>Age of onset †, yr (SD)</td>
<td>24.7 (19)</td>
<td>14.6 (15.4)</td>
<td>35.3 (19.6)</td>
<td>28.2 (18.3)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Atopic status †, % positive</td>
<td>72.8</td>
<td>95.1</td>
<td>51.9</td>
<td>64.6</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Body mass index †, kg/m² (SD)</td>
<td>27.5 (5.4)</td>
<td>26.1 (3.8)</td>
<td>36.2 (5.5)</td>
<td>26 (3.6)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>PC₂₀ methacholine †‡, mg/ml</td>
<td>1.04 (1.13)</td>
<td>0.12 (0.86)</td>
<td>1.60 (0.93)</td>
<td>6.39 (0.75)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>PC₂₀ &gt;8 mg/ml, n (%)</td>
<td>64 (34.7)</td>
<td>2 (3.3)</td>
<td>6 (22.2)</td>
<td>56 (58.3)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Peak flow variability †‡, amp % mean</td>
<td>17 (0.38)</td>
<td>20 (0.47)</td>
<td>21.9 (0.32)</td>
<td>14.8 (0.32)</td>
<td>0.039</td>
</tr>
<tr>
<td>FEV₁ change with bronchodilator †, %</td>
<td>1.63 (1.16)</td>
<td>4.5 (0.91)</td>
<td>1.82 (1.16)</td>
<td>0.83 (1.22)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Post-bronchodilator FEV₁, % predicted</td>
<td>91.4 (21)</td>
<td>86.9 (20.7)</td>
<td>91.5 (21.4)</td>
<td>94.2 (20.7)</td>
<td>0.107</td>
</tr>
<tr>
<td>Sputum eosinophil count †‡, %</td>
<td>1.32 (0.62)</td>
<td>3.75 (0.64)</td>
<td>1.55 (0.51)</td>
<td>0.65 (0.44)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>FₑₑNO †‡§, ppb</td>
<td>31.6 (0.33)</td>
<td>57.5 (0.27)</td>
<td>25.8 (0.29)</td>
<td>22.8 (0.27)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Sputum neutrophil count †, %</td>
<td>55.09 (0.31)</td>
<td>45.87 (0.24)</td>
<td>72.71 (0.13)</td>
<td>57.56 (0.36)</td>
<td>0.038</td>
</tr>
<tr>
<td>Modified JACS † (SD)</td>
<td>1.36 (0.74)</td>
<td>1.54 (0.58)</td>
<td>2.06 (0.73)</td>
<td>1.04 (0.66)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Dose of inhaled corticosteroid, BDP equivalent/µg (SD)</td>
<td>632 (579)</td>
<td>548 (559)</td>
<td>746 (611)</td>
<td>653 (581)</td>
<td>0.202</td>
</tr>
<tr>
<td>Long-acting bronchodilator use, %</td>
<td>40.2</td>
<td>34.4</td>
<td>48.2</td>
<td>41.7</td>
<td>0.442</td>
</tr>
<tr>
<td>Previous hospital admission or emergency attendance, no. per patient</td>
<td>0.60 (1.57)</td>
<td>1.04</td>
<td>0.26</td>
<td>0.20</td>
<td>0.037</td>
</tr>
<tr>
<td>Previous outpatient attendance, % attended</td>
<td>15%</td>
<td>22%</td>
<td>19%</td>
<td>6%</td>
<td>0.121</td>
</tr>
<tr>
<td>Severe asthma exacerbations (requiring oral corticosteroids) in past 12 mo, no. per patient</td>
<td>1.25 (1.94)</td>
<td>1.86 (0.32)</td>
<td>1.07 (0.32)</td>
<td>0.39 (0.18)</td>
<td>0.002</td>
</tr>
</tbody>
</table>

*Boldface type denotes population statistics. The column headed “Cluster 3” represents a cluster not observed in the secondary-care asthma population.

*Comparison between clusters using analysis of variance for continuous variables and χ² test for proportions. Significance values for variables included in the cluster analysis are a product of the cluster algorithm and are provided for illustrative purposes only.

†Variables included in the cluster analysis.

‡Geometric mean (log₁₀ SD)

§Measured with NIOX at a flow rate of 50 ml/second.
### Clusters in secondary care

<table>
<thead>
<tr>
<th>Variable</th>
<th>Secondary Care ( (n = 187) )</th>
<th>Cluster 1 Early Onset, Atopic ( (n = 74) )</th>
<th>Cluster 2 Obese, Noneosinophilic ( (n = 23) )</th>
<th>Cluster 3 Early Symptom Predominant ( (n = 22) )</th>
<th>Cluster 4 Inflammation Predominant ( (n = 68) )</th>
<th>Significance ( (P \text{ Value}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex ( \dagger, % \text{ female} )</td>
<td>65.8</td>
<td>68.2</td>
<td>47.1</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age, yr (SD)</td>
<td>43.4 (15.9)</td>
<td>35.5 (15.5)</td>
<td>50.6 (15.1)</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age of onset ( \dagger, ) yr (SD)</td>
<td>20.3 (18.4)</td>
<td>12.6 (15)</td>
<td>32.6 (19.1)</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Atopic status ( \dagger, % \text{ positive} )</td>
<td>73.8</td>
<td>81.8</td>
<td>63.2</td>
<td>0.024</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Body mass index ( \dagger, \text{ kg/m}^2 ) (SD)</td>
<td>28.5 (6.5)</td>
<td>23.6 (3.1)</td>
<td>27 (3.9)</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peak flow variability ( \dagger, % \text{ mean} )</td>
<td>32.2 (0.48)</td>
<td>24.2 (0.65)</td>
<td>27.6 (0.36)</td>
<td>0.002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FEV(_1) change with bronchodilator ( \dagger ) %,</td>
<td>12.8 (0.41)</td>
<td>24.5 (0.31)</td>
<td>9.3 (0.35)</td>
<td>4.5 (0.33)</td>
<td>9.8 (0.34)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Post-bronchodilator FEV(_1), % predicted (SD)</td>
<td>82.1 (21.1)</td>
<td>79.0 (21.9)</td>
<td>79.0 (18.5)</td>
<td>79.5 (26.1)</td>
<td>87.2 (18.5)</td>
<td>0.093</td>
</tr>
<tr>
<td>Sputum eosinophil count ( \dagger, % )</td>
<td>2.9 (0.99)</td>
<td>4.2 (0.76)</td>
<td>1.3 (1.01)</td>
<td>0.1 (0.9)</td>
<td>8.4 (0.64)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>FE(_{NO}), ppb ( \dagger )</td>
<td>43 (0.32)</td>
<td>51.2 (0.36)</td>
<td>24.2 (0.27)</td>
<td>22.6 (0.30)</td>
<td>53.1 (0.32)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Sputum neutrophil count, % ( \dagger )</td>
<td>46.7 (0.32)</td>
<td>45.4 (0.39)</td>
<td>49.3 (0.22)</td>
<td>51.3 (0.23)</td>
<td>45.9 (0.29)</td>
<td>0.892</td>
</tr>
<tr>
<td>Modified JACS ( \dagger ) (SD)</td>
<td>2.02 (1.16)</td>
<td>2.63 (0.93)</td>
<td>2.37 (1.09)</td>
<td>2.11 (1.11)</td>
<td>1.21 (0.95)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Dose of inhaled corticosteroid, BDP equivalent/µg (SD)</td>
<td>1,018 (539)</td>
<td>1,168 (578)</td>
<td>1,045 (590)</td>
<td>809 (396)</td>
<td>914 (479)</td>
<td>0.008</td>
</tr>
<tr>
<td>Long-acting bronchodilator use, %</td>
<td>93.0</td>
<td>91.9</td>
<td>95.4</td>
<td>90.9</td>
<td>94.1</td>
<td>0.999</td>
</tr>
</tbody>
</table>

*Resembled clusters from primary care – i.e., these are common across spectrum of severity*

**Objective measures of disease severity show more advanced disease**
How should we treat asthma?

• Now we use 3rd dataset – 68 patients over 12 months
• Randomized control trial with two arms:
  – Standard clinical care ("clinical")
  – Regular monitoring of airway inflammation using induced sputum, to titrate steroid therapy to maintain normal eosinophil counts ("sputum")
• Original study found no difference in corticosteroid usage
  – But, this could have been explained by heterogeneity in treatment response!

[Haldar et al., Am J Respir Crit Care Med, 2008]
Patients in different clusters respond differently to treatment! (analysis using 3\textsuperscript{rd} dataset from 12 month study)

<table>
<thead>
<tr>
<th>Cluster (found using baseline data)</th>
<th>Outcomes</th>
<th>Treatment strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Clinical ($n = 10$)</td>
</tr>
<tr>
<td>1: Obese female</td>
<td>$\Delta$ Inhaled corticosteroid dose $^\star$ /µg per day (SEM)</td>
<td>$-400$ (328)</td>
</tr>
<tr>
<td></td>
<td>Severe exacerbation frequency over 12 mo (SEM)</td>
<td>1.40 (0.78)</td>
</tr>
<tr>
<td></td>
<td>Number commenced on oral corticosteroids</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Clinical ($n = 15$)</td>
<td>Sputum ($n = 24$)</td>
</tr>
<tr>
<td>2: Inflammation predominant</td>
<td>$\Delta$ Inhaled corticosteroid dose $^\star$ /µg per day (SEM)</td>
<td>$+753$ (334)</td>
</tr>
<tr>
<td></td>
<td>Severe exacerbation frequency over 12 mo (SEM)</td>
<td>3.53 (1.18)</td>
</tr>
<tr>
<td></td>
<td>Number commenced on oral corticosteroids</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Clinical ($n = 7$)</td>
<td>Sputum ($n = 4$)</td>
</tr>
<tr>
<td>3: Early symptom predominant</td>
<td>$\Delta$ Inhaled corticosteroid dose $^\star$ /µg per day (SEM)</td>
<td>$+1,429$ (429)</td>
</tr>
<tr>
<td></td>
<td>Severe exacerbation frequency over 12 mo (SEM)</td>
<td>5.43 (1.90)</td>
</tr>
<tr>
<td></td>
<td>Number commenced on oral corticosteroids</td>
<td>6</td>
</tr>
</tbody>
</table>

[Haldar et al., *Am J Respir Crit Care Med*, 2008]
Time-series in healthcare

1. Physiological monitoring
   - Monitoring babies in neonatal ICUs
   - Detecting atrial fibrillation

2. Disease progression
   - Predicting Alzheimer’s disease status
   - Learning about disease from data
     • New subtypes of asthma
     • **New staging system for Alzheimer’s**
What if staging system is unknown, or incomplete?

- 3 currently defined clinical stages of Alzheimer’s disease:
  - Normal
  - MCI (Mild Cognitive Impairment)
  - AD (Alzheimer’s disease)

- But, are there really just 3 stages?

- **Goal:** using clinical data, learn a *new* 6 stage system

[Sukkar et al., IEEE EMBS ‘12]
Alzheimer’s disease neuroimaging dataset

- Alzheimer’s disease neuroimaging dataset:
  - 819 subjects
  - 229 “Normal” at beginning, 398 “MCI”, and 192 “AD”
  - Followed for up to 36 months with visits every 6 months

Brain ventricular and hippocampus volumes, as measured by MRI, correlated with AD diagnosis:

[Figure showing scatter plot of normalized hippocampus volume vs. ventricular boundary shift integral (VBSI)]

[Sukkar et al., IEEE EMBS ‘12]
Observations at each time point

• **Four features:**
  – Ventricular boundary shift integral (VBSI)
  – Hippocampus volume normalized by the skull volume
  – Change in VBSI between two successive visits
  – Change in normalized hippocampus volume between two successive visits

• (A modern version of this study would use a deep generative model directly on the images)

[Sukkar et al., IEEE EMBS ‘12]
Hidden Markov model (HMM)

- $X_t$ denotes disease stage at time $t$ (a discrete value from 1 through 6) – never observed
- Probabilistic model parameterized by:
  - Prior distribution $Pr(X_1)$
  - Transition distribution $Pr(X_t | X_{t-1})$
  - Emission distribution $Pr(Y_t | X_t)$: Gaussians
Parameterizing the HMM

• Each subject *regardless of clinical diagnosis at any of his/her visits* allowed to enter HMM at any state, end at any state
  – I.e., $\Pr(X_1)$ is unconstrained

• HMM restricted to only allow transitions between neighboring states, e.g. $1<->2$, $2<->3$, ...
  – I.e., $\Pr(X_t = s_t | X_{t-1} = s_{t-1}) = 0$ if $|s_t - s_{t-1}| \neq 1$

[Disease stage (unknown)]

[Observations at 6 month visit]

[Sukkar et al., IEEE EMBS ‘12]
Learning the HMM

• Have sequences $Y$ for each patient
• Learn model parameters using maximum likelihood estimation
  – We again use expectation maximization because $X$ is unobserved

[Disease stage (unknown)]

[E-Step]
Impute variables

[M-Step]
Update model

Observations at 6 month visit

$X_1, X_2, X_3, X_4, X_5, X_6$

$Y_1, Y_2, Y_3, Y_4, Y_5, Y_6$

[Sukkar et al., IEEE EMBS ‘12]
Results – new stages reflect a spectrum

Based on MAP inference on held-out data:

Results of HMM modeling disease progression, the HMM state correlates with Alzheimer’s disease progression [1]. Higher scores indicate higher dementia impairment and mental status examination and is included in the ADNI dataset for each visit. The score ranges from 0 to 8 where 1 represents no dementia, 2 represents very mild dementia, 3 represents mild dementia, 4 represents moderate dementia, 5 represents severe dementia, and 6 represents very severe dementia.

To see how each individual subject progresses through varying and more granular stages in disease progression, we can interpret these results as evidence that the states of the HMM capture meaningful clinical information.

The HMM state sequence for each subject includes the number of visits for subject and the actual CDR SB score over all visits. We computed the average CDR SB score over all visits for each subject. The CDR SB score is derived from patient interviews and is included in the ADNI dataset for each visit. The CDR SB score correlates with Alzheimer’s disease progression and dominates the Clinical Dementia Rating Scale Sum of Boxes (CDR SB). The CDR SB score ranges from 0 to 8 where 1 represents no dementia, 2 represents very mild dementia, 3 represents mild dementia, 4 represents moderate dementia, 5 represents severe dementia, and 6 represents very severe dementia.

To evaluate the model for its performance in diagnosing Alzheimer’s disease, we used a measurement called the Clinical Dementia Rating Scale Sum of Boxes (CDR SB). The CDR SB score ranges from 0 to 8 where 1 represents no dementia, 2 represents very mild dementia, 3 represents mild dementia, 4 represents moderate dementia, 5 represents severe dementia, and 6 represents very severe dementia.

To measure how well the HMM model captures the clinical diagnosis, we used the root mean square deviation between the CDR SB scores of the subject’s actual diagnosis and the predicted diagnosis. We computed the average CDR SB score over all visits for each subject. We can see that “Normal” diagnosis dominates in early (i.e., low index) states and diminishes with increasing state index, while “AD” diagnosis behaves in the opposite way monotonically increasing with state index. The “MCI” state index, while “AD” diagnosis behaves in the opposite way monotonically increasing with state index.

We computed the probability of clinical diagnosis over all visits for each subject given his/her biomarker measurements at the corresponding visit. As a result, we have, for each subject, an HMM state sequence and a corresponding state index. We computed the probability of clinical diagnosis for the three classes of diagnoses given the HMM state, $P[\text{Diagnosis}|\text{State}]$. Figure 3 shows the probability of clinical diagnosis over all visits for subject and the actual CDR SB score of each subject.

Figure 2. Scatter plot of the normalized hippocampus volume versus VBSI for Normal and AD subjects. Figure 3. Probability of clinical diagnosis over all visits - Results – new stages reflect a spectrum.
Results – stages correlate with dementia clinical score

Average Clinical Dementia Rating Scale Sum of Boxes (CDR-SB)

For some subjects, the progression path at the normal and early stages indicated by the CDR-SB scores. This suggests that the HMM and CDR-SB indicate similar progression paths. We showed that the optimal HMM state indexes are lower than at the high state indexes. This gives evidence of the progression paths which can give a different perspective on the progression of the disease increasingly different from the current stages. To gain more revealing aspects of disease progression, we trained an unsupervised Hidden Markov Model in the ADNI data set. Using the ADNI data set, we presented a model for disease progression based on a Hidden Markov Model framework. The results of Figure 5 give a different perspective on the progression of the disease.

We showed that the deviation at the low HMM state according to the HMM state index with the Root Mean Squared Deviation computed over all visits that dwelled in a given HMM state versus HMM state index.

CONCLUSIONS

In the HMM, the disease progression path gives evidence of the progression paths which can give a different perspective on the progression of the disease increasingly different from the current stages. To gain more revealing aspects of disease progression, we trained an unsupervised Hidden Markov Model in the ADNI data set. Using the ADNI data set, we presented a model for disease progression based on a Hidden Markov Model framework. The results of Figure 5 give a different perspective on the progression of the disease increasingly different from the current stages.

REFERENCES


Summary

• We are nearly always in realm of “not enough data”
• Modeling and incorporating prior knowledge is critical to good performance
• Design principles
  – Model the distribution of physiological dynamics
  – Derive features using existing clinical knowledge
  – Start from the simplest possible model
  – Share statistical strength across tasks
  – Discover structure using unsupervised learning