15th September 2020

Lee Kiat Kai

Presentation of findings

Contents

1. Study Protocol

Title:

An mHealth app using machine learning to increase physical activity in diabetes and depression: clinical trial protocol for the DIAMANTE Study

(using student dataset)

Background:

- Depression and diabetes are major causes of global disability, which can lead to an increased risk of mortality
- Increasing physical activities may be effective in reducing the risk of both depression and diabetes
- Use of mobile phone applications to help increase physical activities
- But this has to be tailored to increase effectiveness

Aim:

To test whether there will be improvements in physical activities, measured by the daily step counts in 6 months between 3 different groups

1) Group receiving adaptive messaging

Will receive the daily messages but message categories, timing and frequency will be determined by reinforcement learning algorithm

2) Group receiving uniform random messaging

Will receive 2 messages per day within 4 randomly selected time intervals

3) Control group receiving only a weekly mood message

Will not receive feedback messages

- All patients in every group will receive a mood message once a week, requiring them to rate their mood on a scale from 1 to 9

Variables (features)

- 120 total features
- 12 Response features:
  - for e.g. step_change_today
- 28 General features:
  - for e.g. education, marital status, gender, date of baseline visit
- 44 Health features:
  - for e.g. loneliness_1, start_gad_1
- 4 Sociability features:
  - for e.g. how often do you call, how often do you meet
- 11 Summary features:
  - for e.g. GAD_sum, GAD_cat
- 21 Messaging features:
  - for e.g. which time period of the day is the message sent, days since receiving no feedback messaging

2. Digital Phenotyping

Introduction to Digital Phenotyping

- Digital Phenotyping is the moment-by-moment quantification of the individual-level human phenotype in situ using personal digital devices
- Phenotype refers to the observable physical properties of an organism (for e.g. height, blood type, behavior)
- With a large number of population owning smartphones, it generates unique opportunities for quantifying human behavior where it can be carried out anywhere without needing to be confined to clinics or research laboratories

Making Sense of Smartphone Data:

Active Data:
Requires active participation such as surveys (mood message survey in the study)

Passive Data:
Generated without any participation or action (such as GPS traces)

Technology can capture data continuously and monitor for a subtle social or behavioral red flag (for e.g. time spent reading messages is zero, time spent on mHealth app is zero)

Paper 1: Mood Prediction of Patients with Mood Disorders by Machine Learning Using Passive Digital Phenotypes based on the Circadian Rhythm

Chul-Hyun Cho, MD, PhD; Taek Lee, PhD; Min-Gwan Kim, MS; Hoh Peter In, PhD; Leen Kim, MD, PhD; Heon-Jeong Lee, MD, PhD (17 August 2019)

- **Objective**: To evaluate the mood state during a period of 2 years from data collected by smartphones and conventional clinical assessment

- **Method**: Prospective observational cohort study on 55 patients with mood disorders
    - Passive digital phenotypes were processed into 130 features based on circadian rhythms, and a mood prediction was developed by random forests

Methods:

- 55 patients:
  - Diagnosed with MDD: 18
  - Diagnosed with BD I: 18
  - Diagnosed with BD II: 19

Assessment:

- Everyday 9pm, patients record their daily mood state (-3 to +3) on their smartphone app
- Dailymood scores were converted to absolute mood score (AMS: 0 to +3)
  - Higher the AMS, mood can be regarded as worse and unstable
  - Lower the AMS, mood can be regarded as stable

Datasets

Total of 17,542 sample days data but only used 2003 days without any single missing variables

Basic features that causes mood state to be affected by the disruption of circadian rhythms:

- Light exposure, 2
- Steps, 2
- Sleep, 4
- Heart rate, 5

130 features and class label of the mood state
= 131 total columns

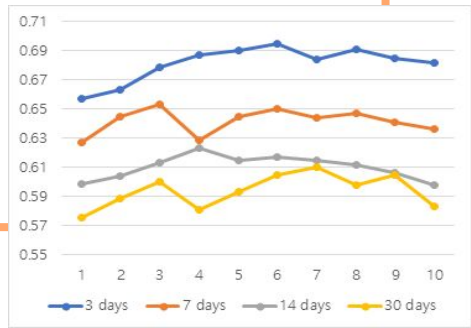
130 features =
13 basic + 13 basic x 3 types of past periods (previous consecutive 3,6,12 days)
x 3 types of statistics (mean, stdev, gradient coefficient)

Development and verification of the mood state and episode prediction algorithm

- Used a supervised learning algorithm, random forest, constructing multiple decision trees and outputting a class that is the mode of the classes of those trees
- Performance of the model was evaluated by assessing model's accuracy, sensitivity, specificity, and the area under the curve (AUC)

Development and verification of the mood state and episode prediction algorithm

- Data were sorted as such:
  - Prediction model was trained using data on days d[t-p,t] and tested using data on days d[t+1,t+q]
  - Repeated and experimented on p from 3 to 300 and q from 3 to 30
  - **p = 18 and q = 3 were found as the best combination**
  - **Suggest that 3 days** is the **most reasonable setting** in terms of predicting any distant future mood in the experiment



Results

- Activity and light exposure during bedtime showed a higher tendency in the HAMS groups than the LAMS groups
- Total sleep time and sleep quality did not show much differences
- Regularity of the sleep-wake cycle were disrupted in the HAMS group, indicating that regularity of sleep-wake cycle is closely related to mood state
- Misaligned or shifted heart rate acrophase (time period in a cycle during which the cycle crests or peaks) could be a useful feature for determining mood state as there was a remarkable difference between the HAMS and LAMS groups

Paper 2: Beyond smartphones and sensors: choosing appropriate statistical methods for the analysis of longitudinal data

Ian Barnett, John Torous, Patrick Staples, Matcheri Keshavan and Jukka-Pekka Onnela (23 August 2018)

- **Objective:** To review dimension reduction for correlated behavioral covariates for longitudinal data

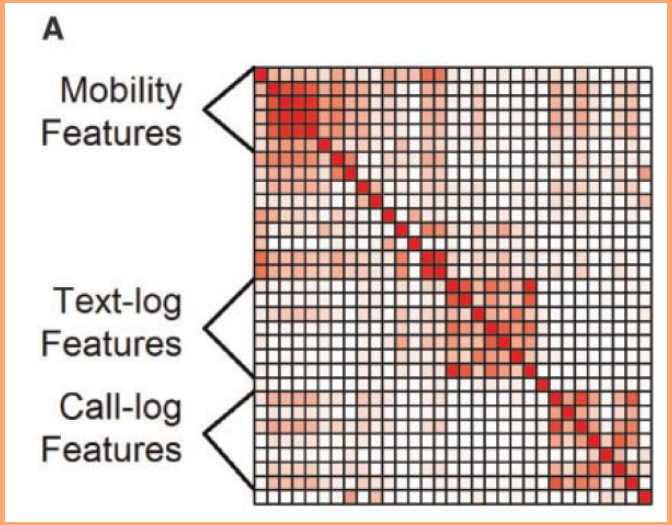
- **Method:**
  - Cohort Study of 17 patients with schizophrenia for up to 3 months through their smartphone use
  - 16 summary features of mobility were estimated from GPS (for e.g. distance travelled, fraction of time spent at home, etc)
  - 15 summary features of sociability (for e.g. number of texts sent and number of texts received)
  - Phone surveys given to patients, measuring their anxiety, depression, sleeping habits, psychosis and warning signs of psychosis

Dimension reduction:

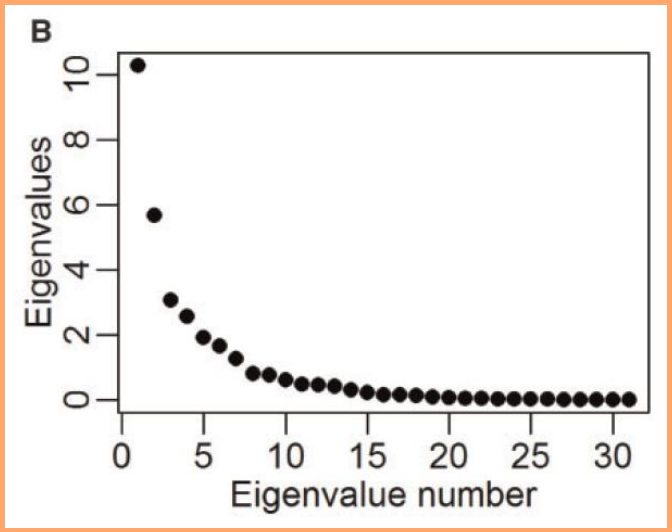
Redundant predictors can greatly hamper a wide variety of statistical analyses by

- Increasing variability in estimation
- Increasing degrees of freedom
- Obfuscating model interpretations

Heat Map:



Scree plot:



Applying to the DIAMANTE dataset

Types of clustering methods that we can try:

- Random forest
  - Constructs multiple decision trees and outputs the majority output
- K-Means
  - Iterative clustering algorithm where we have to specify the number of clusters K, at the start
- Hierarchical clustering
  - Starts with all the data points assigned to a cluster of their own. Then two nearest clusters are merged into the same cluster. In the end, this algorithm terminates when there is only a single cluster left.

Applying to the DIAMANTE dataset

Types of clustering methods that we can try:

- Latent class analysis
  - identifying unmeasured class membership among subjects using categorical and/or continuous observed variables
  - For example, we may want to categorize participants into different level of fitness (latent class) based on the number of steps they take daily and other useful observations.

Applying to the DIAMANTE dataset

- Along with the different clustering methods, we will also have to perform dimension reduction to remove redundant and irrelevant features without incurring much loss of information
- Lastly, we have to evaluate the performance of all our models through to find the best

Benefits/use of digital phenotyping on DIAMANTE

Can determine which messages is the best in stimulating participants to walk	Can determine which timing is the best for stimulating participants to walk	Being able to determine the best message and timing allows for a intervention that may work well with the general public
Reduce the risk of depression and diabetes in patients	Allows for a cost-efficient intervention method	

Questions:

- For example start_ipaq only has 1 to 3 but for instance start_phq8 has 1 to 8, what does the numbering behind represent? (Days, weeks, intervals?)
- What defines being successful in taking more steps? (Do we look at if the change is more than 1? or do we compared with predictions?)
- What does this variables represent?

Questions

- basics_challenges_r
- years_us
- pain
- start_soc_active, start_soc_months
- individual
- Fcorrection
- Mcorrection
- days.since.ind, days.since.soc
- arm