

Towards Sequential Data Clustering via GBTM and LSTM

Yirui Hu, *Ph.D.*

Department of Population Health Sciences, *Geisinger*

Symposium on Data Science and Statistics

Machine Learning 3, June 05, 2020

Outline

- Electronic health records (EHR) contain longitudinal data with valuable phenotypic information on disease progression of patients.
 - Weight change trajectories in Roux-en-Y Gastric Bypass (RYGB) patients over 7 years
- Group-based trajectory modeling (GBTM)
 - A specialized application of finite mixture modeling designed to identify clusters of individuals who follow similar trajectories in sequential data
- Long Short Term Memory (LSTM)
 - A special kind of recurrent neural network that can learn the order dependence in sequential data

Weight change trajectories in RYGB patients





Surgery for Obesity and Related Diseases

Volume 14, Issue 11, November 2018, Pages 1680-1685



Original article

Demographic, clinical, and behavioral determinants of 7-year weight change trajectories in Roux-en-Y gastric bypass patients

Michelle R. Lent Ph.D. ^{a, b}  , Yirui Hu Ph.D. ^c, Peter N. Benotti M.D. ^a, Anthony T. Petrick M.D. ^a, G. Craig Wood M.S. ^a, Christopher D. Still D.O. ^a, H. Lester Kirchner Ph.D. ^c

 [Show more](#)

<https://doi.org/10.1016/j.soard.2018.07.023>

[Get rights and content](#)

- This work was supported by Pennsylvania Department of Health (#SAP 4100070267)

Weight change trajectories in RYGB patients

- To assess long-term weight trajectories in a large cohort of RYGB patients over 7 years
 - Primary outcome: Weight changes (% percent change from baseline) post-operatively over time
- Hypothesis: Patients respond differently to the same surgical intervention
 - Identify heterogenous subgroups
 - Identify high-risks patients / guide clinical care
 - Inform patient-provider conversations about treatment expectations

BMI > 40 kg/m² or > 35 kg/m² + comorbidity

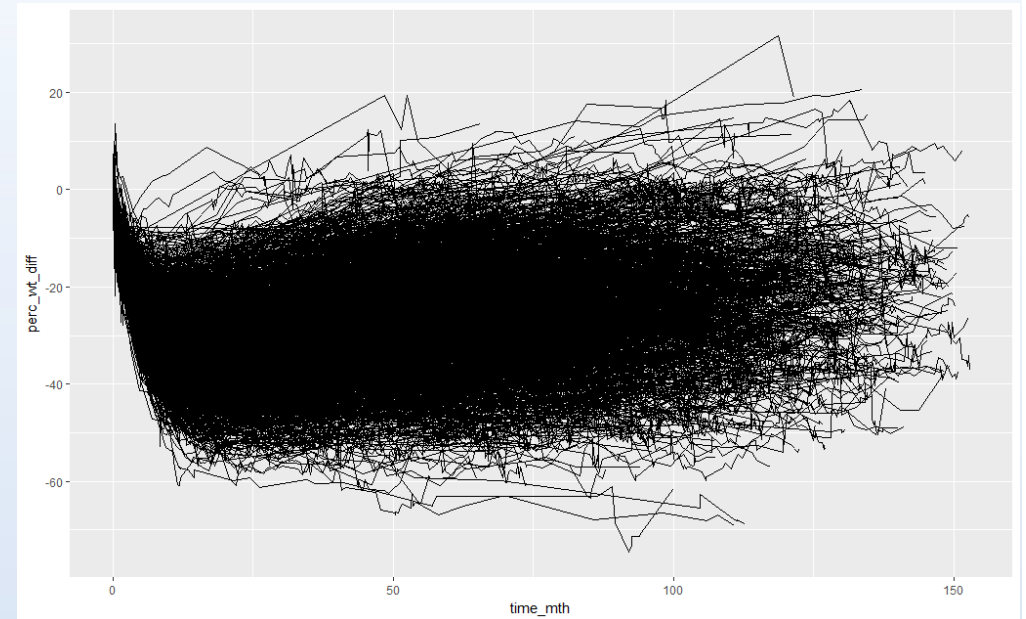
RYGB

Up to 7 years post-surgery

Weight data in the EMR pre and 7 years postop

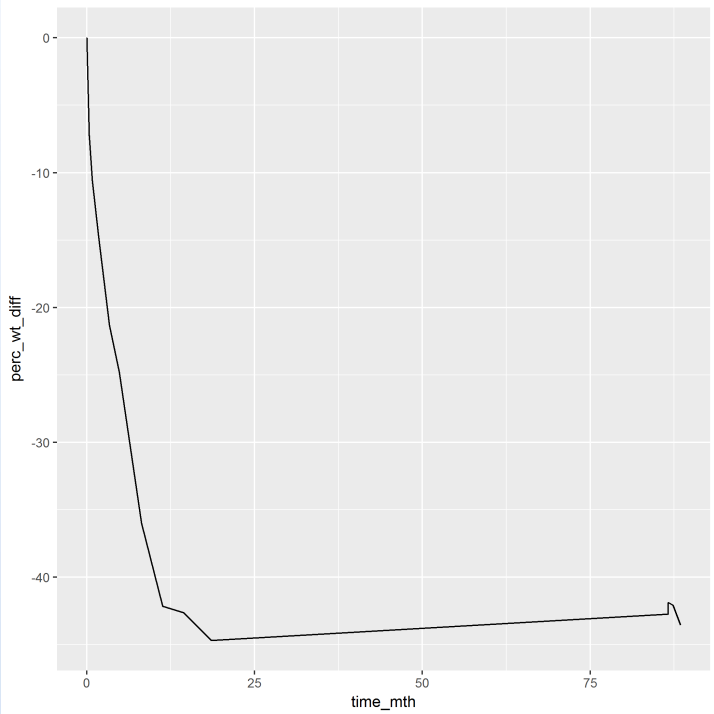
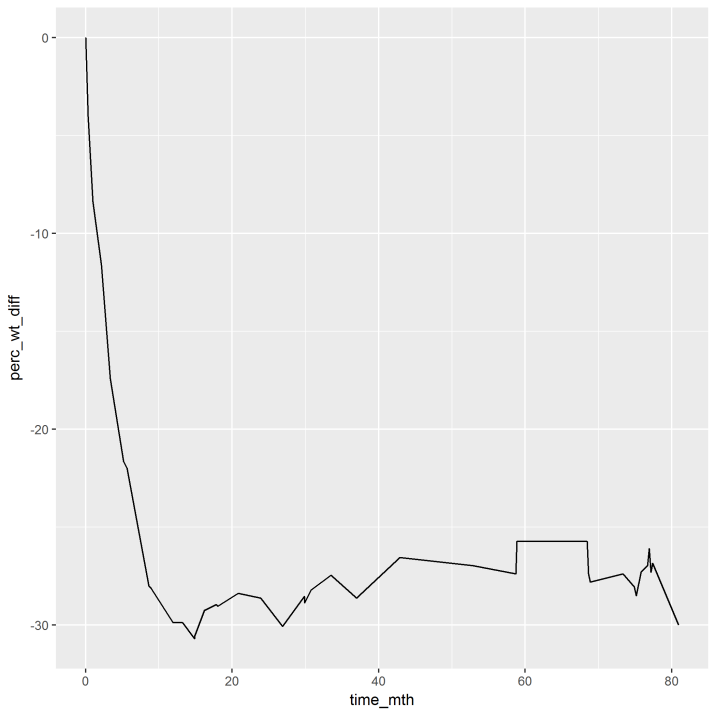
Study Population

- RYGB bariatric surgery (2004-2016)
 - N=3,215
 - 80% female, >95% White
 - Preoperative BMI = 49.4 ± 8.8 kg/m²
- Data visualization
 - X-axis: Months post-surgery
 - Y-axis: Weight changes (% from baseline)



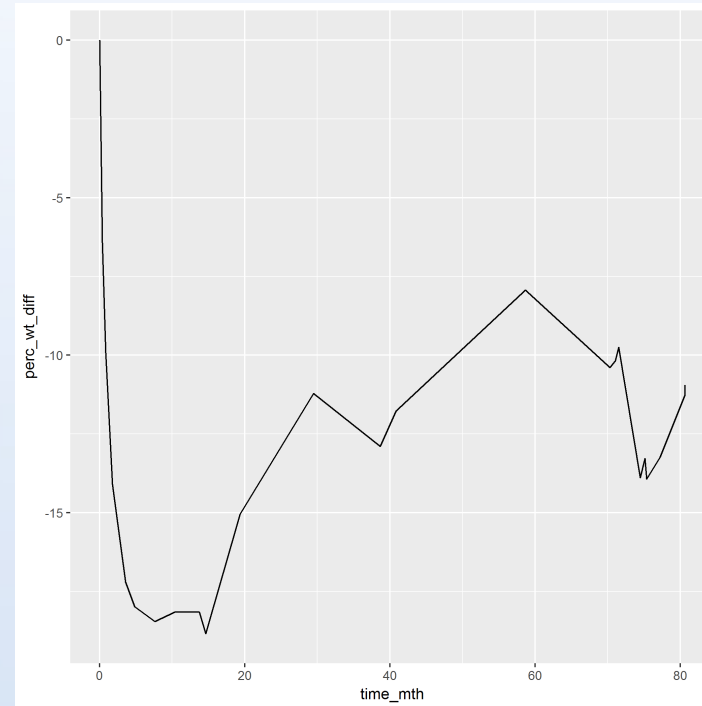
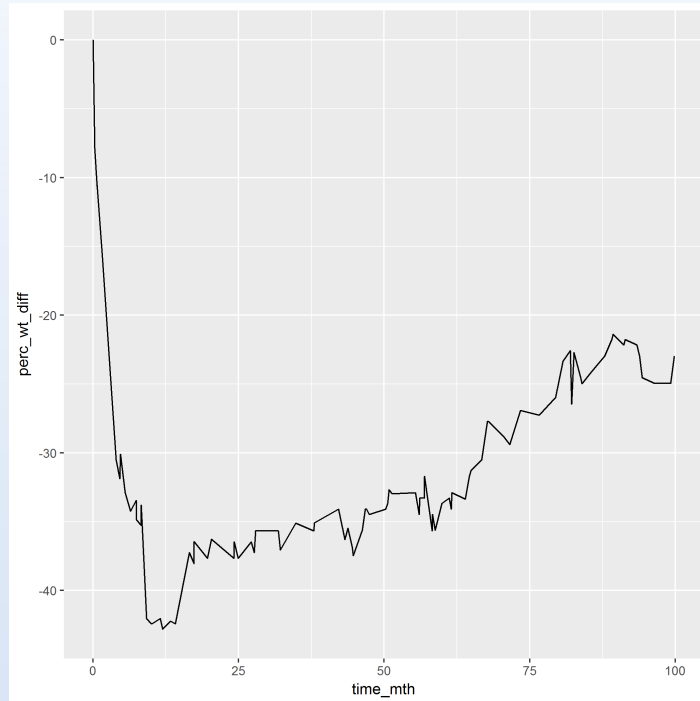
Individual-Level Data Visualization

- Pattern 1: Remain stable in weight change after the first major weight loss



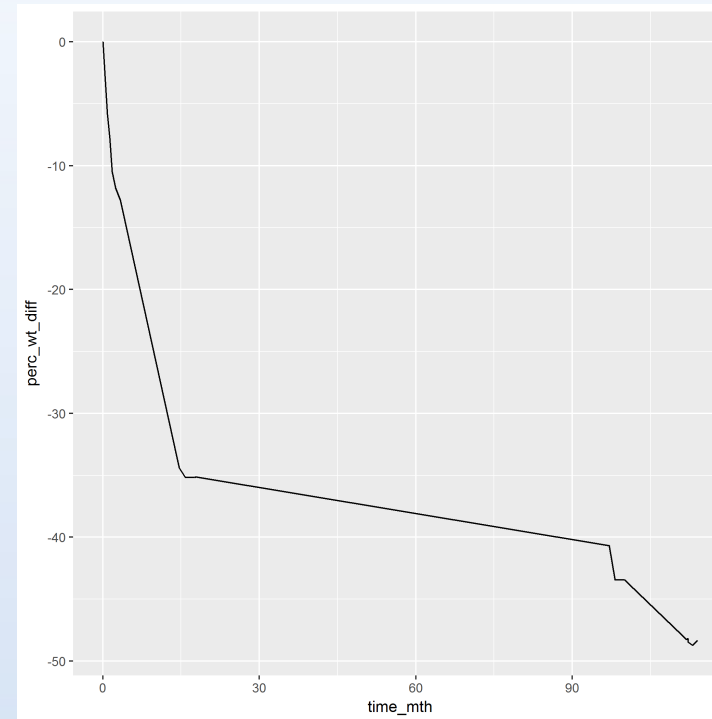
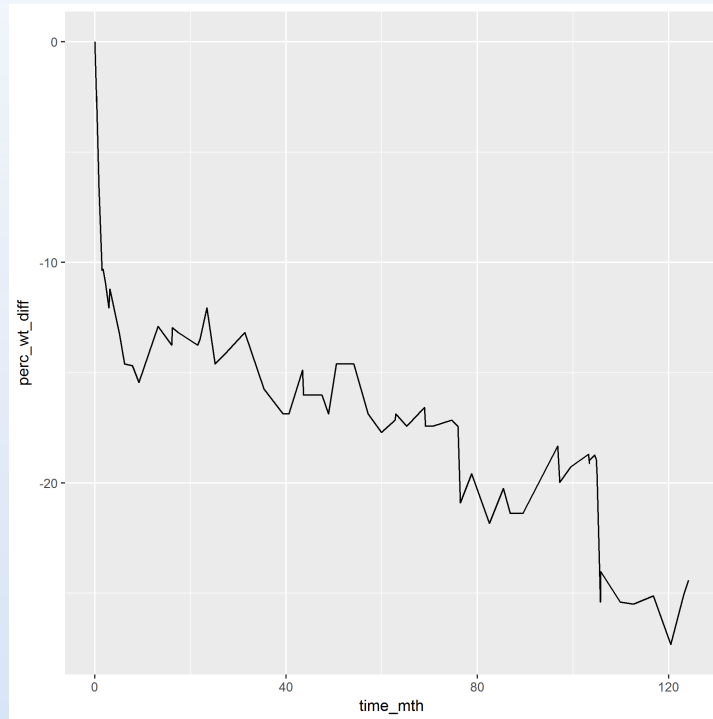
Individual-Level Data Visualization

- Pattern 2: Regain in weight change after the first major weight loss



Individual-Level Data Visualization

- Pattern 3: Lose more in weight change after the first major weight loss



Summary of Individual-Level Patterns

- First stage: major weight loss patterns
 - Almost everyone lose weight to some degree
 - within 12-18 months post-surgery
- Second stage: different weight change patterns (assumes heterogeneous subpopulations)
 - Remain stable -> Average
 - Regain -> Suboptimal
 - Further weight loss -> Optimal

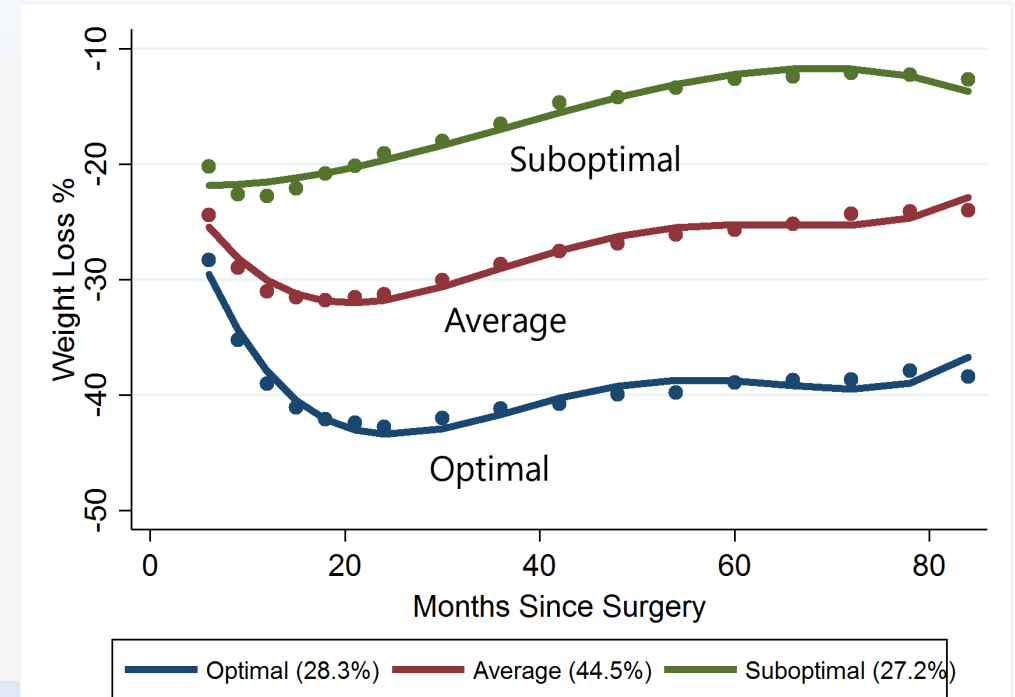
Group-Based Trajectory Modeling (GBTM)

- A statistical method that is designed to identify a finite number of groups of individuals following similar trajectories over time of a single outcome
 - Identify distinctive trajectories: groups of individuals with similar trajectories
 - Estimate the proportion of the population following each such trajectory group
 - Relate group membership probability to individual characteristics
- Two key outputs of the GBTM
 - the shape of the trajectory, typically defined by a polynomial function of time,
 - the probability of trajectory group membership

Summary results of GBTM

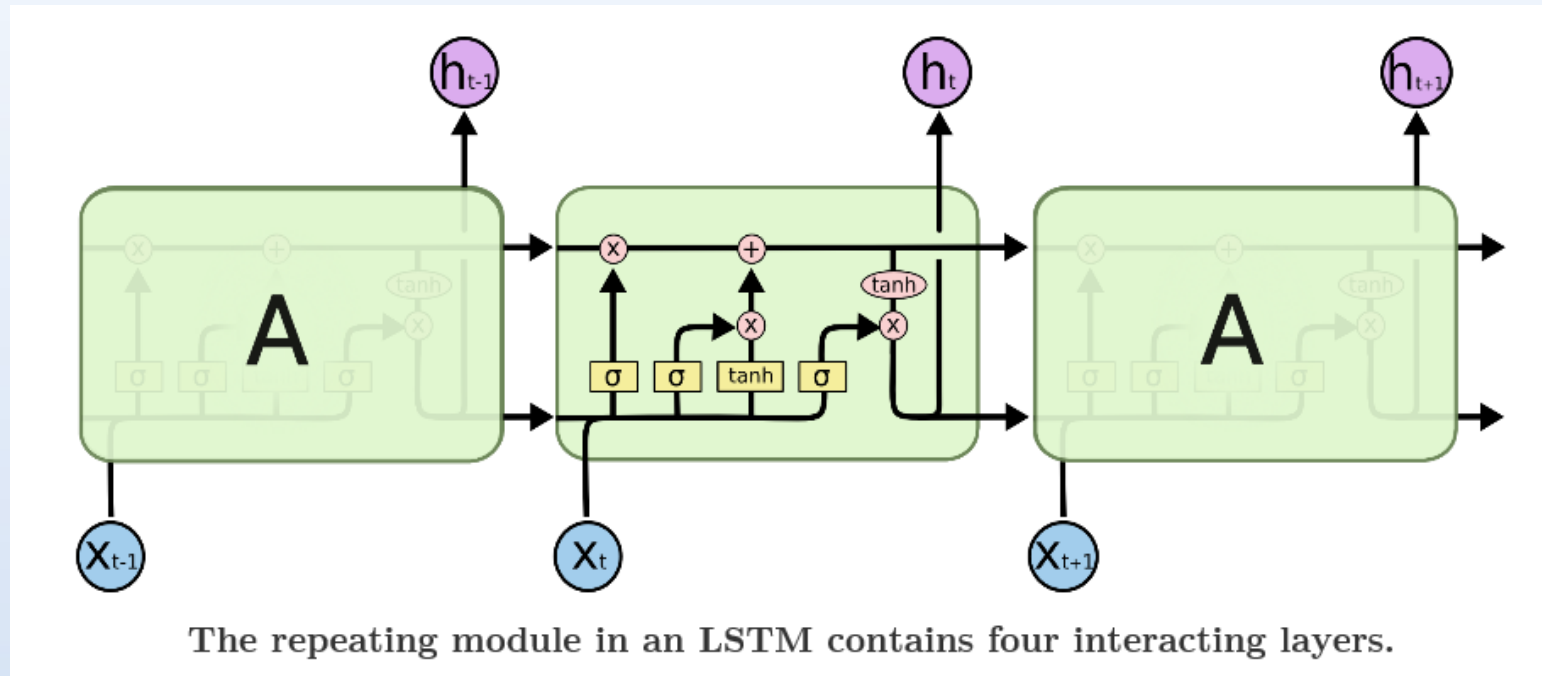
Highlights

- Weight loss trajectories up to seven-years following Roux-en-Y Gastric Bypass (RYGB) were not uniform, with 27% of patients maintaining <20% body weight loss.
- Patients that experienced poorer long-term weight loss were more likely to be male and have diabetes, but less likely to have a smoking history or take sleep medications.
- Lower BMI at the time of surgery and less early postoperative weight loss were associated with poorer weight outcomes.



Long Short Term Memory (LSTM)

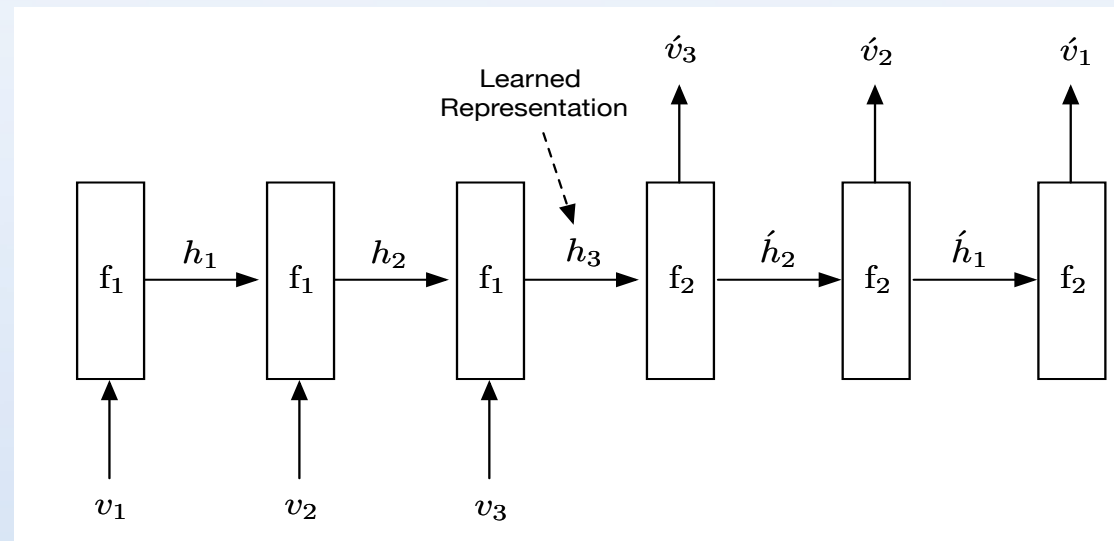
- Recurrent Neural Network (RNN) and its special kind, Long Short Term Memory (LSTM) network have successfully tackled many time-relevant problems.
 - LSTM (introduced by Hochreiter & Schemidhuber 1997) is capable of learning long-term dependencies.



<https://colah.github.io/posts/2015-08-Understanding-LSTMs/>

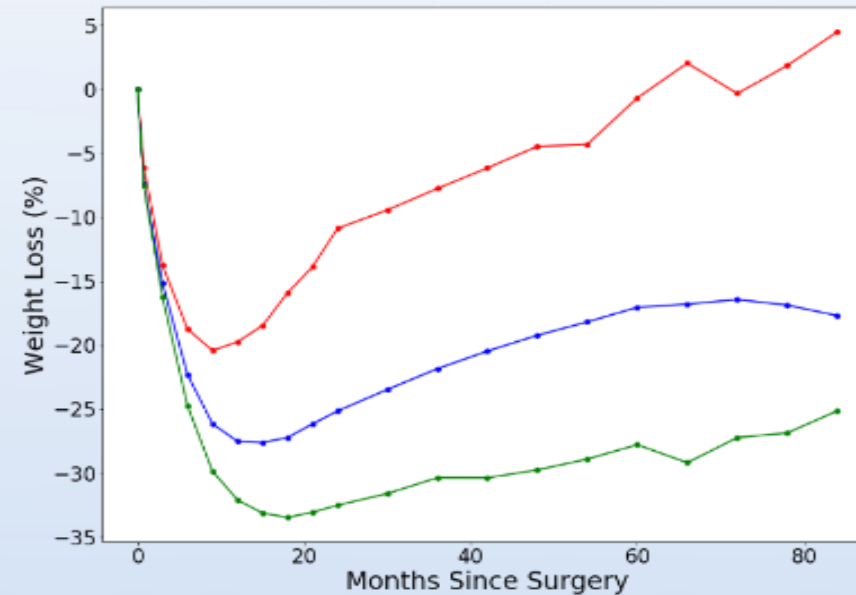
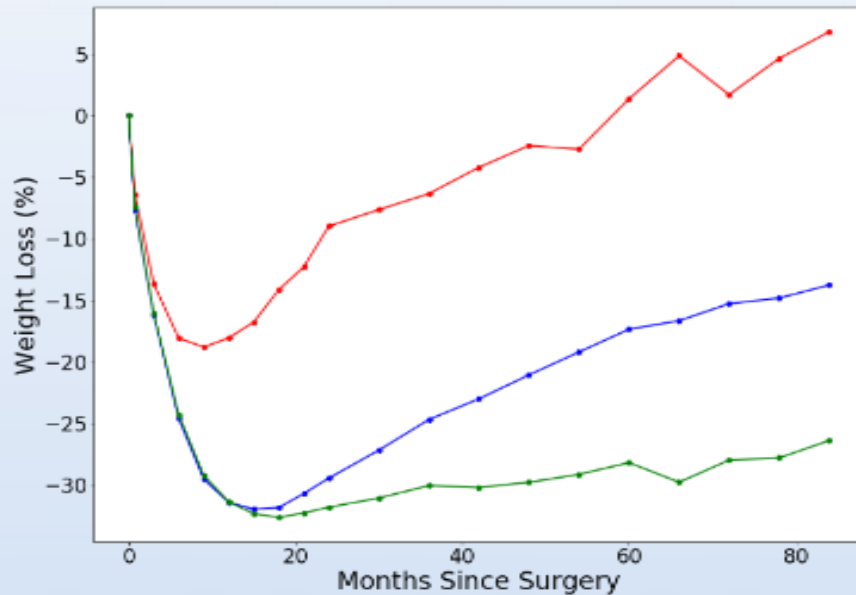
LSTM Autoencoder

- Autoencoders are a type of self-supervised learning model that can learn a compressed representation of input data.
 - For a given dataset of sequences, an encoder-decoder LSTM is configured to read the input sequence, encode it, decode it, and recreate it. The performance of the model is evaluated based on the model's ability to recreate the input sequence.



Clustering (K-Means and GMM)

Subgroup	Red	Blue	Green
K-means Count (%)	82 (2.9%)	1128 (39.6%)	1641 (57.6%)
GMM Count (%)	63 (2.2%)	438 (15.4%)	2350 (82.4%)



Concluding remarks

- Cluster analysis of longitudinal EHR data
 - Group-based trajectory modeling
 - Long Short Term Memory networks
- Potential applications
 - Identify heterogenous subgroups in longitudinal data
 - Link covariates to identified trajectory membership
 - Relate trajectories to subsequent outcomes in predictive modeling

Our Predictive Modeling Pipeline

Data Preparation

- Import data from EHR
 - Input: features and label
- Impute missing values
 - MICE
- Detect outliers

Model Development

- Feature selection and engineering
- Run algorithms
 - Logistic regression
 - Random forest
 - XGBoost
- Cross-Validation
 - Tune hyper-parameters

Model Evaluation

- Performance report
 - AUC, etc.
 - Feature importance
- Risk prediction
 - Probability of developing disease
 - High risk vs. low risk

Acknowledgements



- Collaborators
 - Kunpeng Liu, Dr. Michelle R Lent, Dr. Peter N Benotti, Dr. Anthony T Petrick, G Craig Wood, Dr. Christopher D Still, Dr. H Lester Kirchner
- Funding
 - Pennsylvania Department of Health (#SAP 4100070267)
 - AMS Student & Early Career Award (SDSS 2020)
- Contact
 - yhu1@geisinger.edu

References

- Lent MR, Hu Y, Benotti PN, Petrick AT, Wood GC, Still CD, Kirchner HL. Demographic, clinical, and behavioral determinants of 7-year weight change trajectories in Roux-en-Y gastric bypass patients. *Surgery for Obesity and Related Diseases*. 2018 Nov 1;14(11):1680-5
- <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>
- <https://machinelearningmastery.com/lstm-autoencoders/>

Missing Data Imputation

- In the imputation step, we learned from the series of past observations to predict the next value in the sequence.
- Mean Absolute Error (MAE)
 - MAE of Median imputation: 0.28;
 - MAE of LSTM imputation: 0.003.

