Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group

BIOS 7720: Applied Functional Data Analysis

Leture 1: Course Introduction

Andrew Leroux

March 2, 2021

Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group

Roadmap

1. Introduction to applied functional data analysis
2. Motivating data examples
3. Course outline
4. Software
5. Final projects
6. New FDA working group

A Bit About Me

- Academic Background
  - Sc.M. Biostatistics 2014-2015 (JHU)
  - Research Associate 2015-2016 (JHU)
  - Ph.D. Biostatistics 2016-2020 (JHU)
  - Assistant Professor 2020-
- Mentored students and junior faculty on
  - Functional data analysis (FDA)
  - Analysis of wearable device data

Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group

A Bit About Me

- Applied work
  - Wearable/implantable device data
    - Accelerometry (physical activity/sleep)
    - Heart rate
    - Continuous blood glucose monitoring
  - Ecological momentary assessment (EMA)
    - Patient reported measures collected via smartphone app
    - examples: mood, pain, hunger, energy, etc.
- Methods research at the intersection of FDA and
  - Survey statistics
  - Longitudinal data analysis
  - Survival analysis
  - Shared parameter models (e.g. joint modelling)

Student Introductions

Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group

A bit about yourselves: program, year, area of research

Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group

Course Logistics

- Course syllabus posted on Canvas
- Class meets Tuesday/Thursday 2:30-3:50PM
  - Mostly didactic lectures with in-class exercises
  - 2 lectures devoted to in-class work/feedback on group projects
- Student evaluation
  - 4 homework assignments (60% of students' grade)
  - Final project and presentation (40% of students' grade)
  - Assignments and projects can be worked on in groups of up to 3
    - All students must independently write up their own homeworks
    - A portion of students' grades (20%) on their final projects is determined by peer evaluations

Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group

Course Logistics

- **No** background in FDA assumed
- Generalized linear mixed models are great
- Support only offered for *R*, but use whatever works for you
  - Some software packages for FDA (later)
  - Coding up your own implementations
- All assignments fully reproducible
- (optional) Email me to set up a 1-on-1 meeting

Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group

What are “Functional Data"?

- Conceptualized as arising from a smooth underlying process
  - Theoretically continuous (infinite dimensional)
  - In practice observed discretely
- Functional data are characterized by
  - Correlation along some domain
    - Temporal (time of day, day of year, etc.)
    - Spatial (geographic, anatomical)
  - Density of measurements
    - typically densly measured
    - regular versus irregular

What are “Functional Data"?

Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group

- Types of functional data
  - Continuous
  - Binary
  - Count
  - Categorical (nominal/ordinal)
- Practical challenges
  - Dimensionality
  - Non-linear data and/or association structures
  - Software development

Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group

What are "Functional Data"?

- Functional data can be
  - Outcome (response variable)
  - Predictor (explanatory variable)
  - Both outcome and predictor
- Statistical methods for scalar data extend to functional
  - Multilevel (hierarchical)
  - Longitudinal
  - Multivariate

Goals of Functional Data Analysis

Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group

- Characterize principle sources of variation in the data
- Study how patterns/variation in the data are associated with outcomes
- Study how individual characteristics are associated with patterns/variation in the data

Examples of Functional Data that I've Worked with

Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group

- Child Growth data
- Neuroimaging data (Multiple Sclerosis Lesion Progression)
- Wearable device data
  - Accelerometry (physical activity)
  - Heart rate

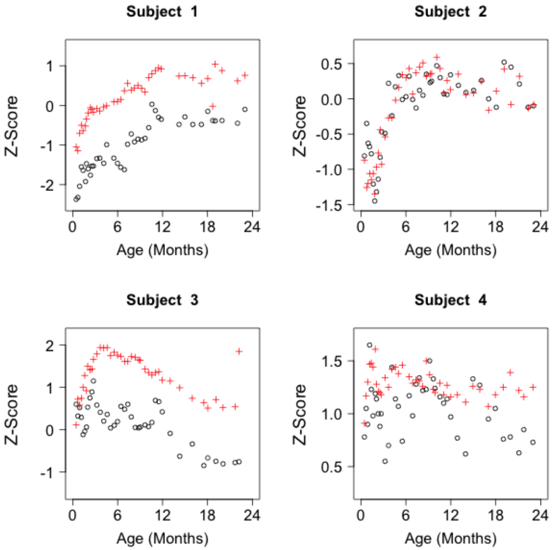
Child Growth: Length/Weight-for-Age Z-scores

Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group

- Context
  - Child growth data
  - Age- and sex-standardized measurements
    - LAZ: length-for-age z-score
    - WAZ: weight-for-age z-score
- Scientific Question/Goals
  - Which children are at risk of stunting (length)?
  - Dynamic prediction of future length
  - Quantify uncertainty
- Statistical Challenges
  - Highly non-linear growth curves
  - How to incorporate weight?
  - How to incorporate historical effect of length?

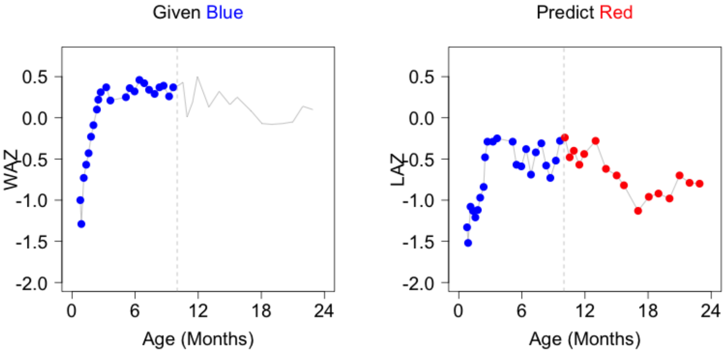
Child Growth: Length/Weight-for-Age Z-scores

Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group



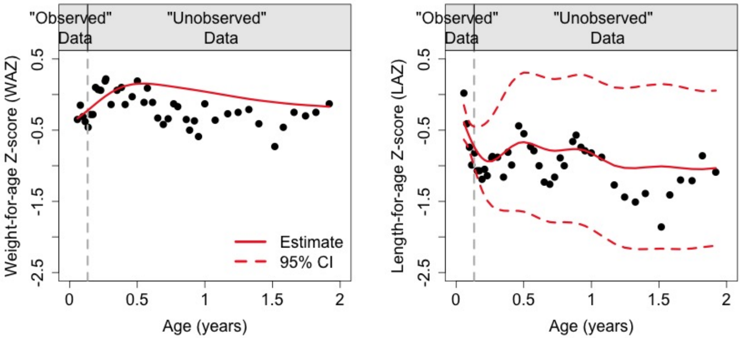
Child Growth: Length/Weight-for-Age Z-scores

Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group



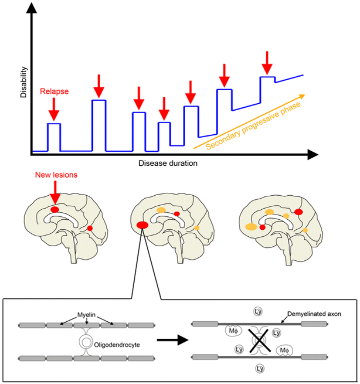
Child Growth: Length/Weight-for-Age Z-scores

Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group



Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group

Multiple Sclerosis: Background and Motivation



[Nakahara et al., 2012]

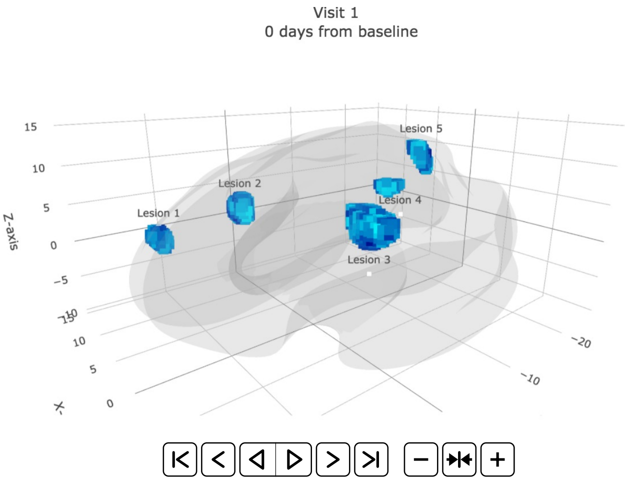
Multiple Sclerosis: Background and Motivation

Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group

- Measurement of lesions
  - Brain measured in voxels ($1mm^3$)
  - Images registered longitudinally
  - Segmentation [Sweeney et al., 2013a, Sweeney et al., 2013b]
- NINDS longitudinal study 2000-2008 [Sweeney et al., 2016]
  - 48 participants
  - 751 lesions
  - 182,006 voxels
  - 24 observations per voxel (4.4 million data points)
  - 30 days between scans
- Remyelination and MRI [Franklin and ffrench-Constant, 2008]

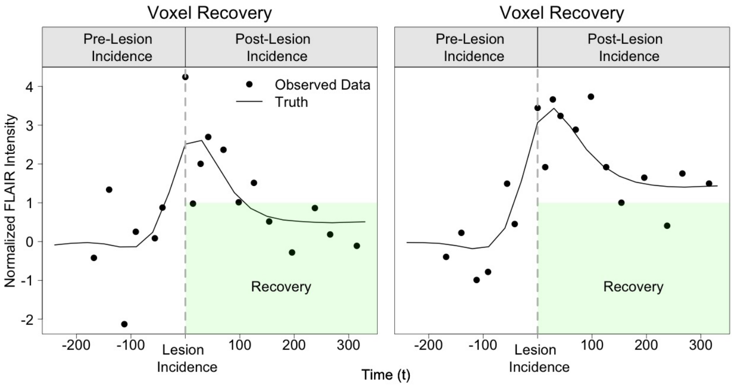
Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group

Neuoimaging: Voxels in MS Patients



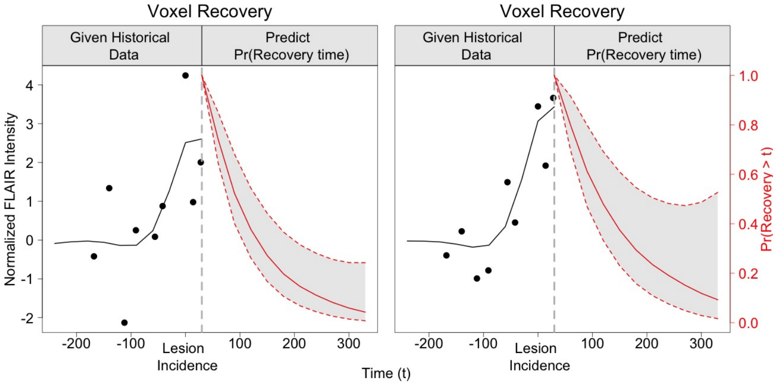
MS Lesion Recovery: Conceptual Framework

Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group



MS Lesion Recovery: Dynamic Prediction

Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group



Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group

Wearable and Implantable Technology

- Wearable and implantable devices are smart electronic devices that can be worn on the body as implants or accessories
- Emerging technology (increasing variety of sensors and signals measured)
- Growing popularity in health research and consumer tech

Why Wearables?

Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group

- Major "modifiable" determinants of health
  - What you put into your body (e.g. food, alcohol, medications)
  - What your body outputs (e.g. movement, physical activity)
  - Environment (e.g. pollution)
- Wearable accelerometers measure (a proxy of) "output"
- Simple summaries of PA are "best" predictors of mortality

Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group

Why Wearables?

- Major “modifiable" determinants of health
  - What you put into your body (e.g. food, alcohol, medications)
  - What your body outputs (e.g. movement, physical activity)
  - Environment (e.g. pollution)
- Generally, wearables can provide information on each of these

Why Wearables?

- Wearable accelerometers measure (a proxy of) “output"
- Simple summaries of PA are “best" predictors of mortality [Smirnova et al., 2019, Leroux et al., 2020]

Table 1. Demographic and Clinical Characteristics Separated by Alive and Deceased Status 5 Years After Participation in the Accelerometry Study, National Health and Nutritional Examination Survey Pooled Cohorts Study, United States, 2003–2006

		Alive	Deceased	
Rank	Characteristics	Mean(*SD*)/*N*(%)[a]		AUC[b]
1	TAC	217,926 (111,868.8)	136,307.9 (94,316.3)	0.771
2	Age	65 (9.3)	73.4 (8.9)	0.758
3	MVPA	14.7 (17.3)	6.5 (12.1)	0.745
4	ASTPsl/nw	0.29 (0.08)	0.37 (0.11)	0.733
5	Sedentary time	1,102.4 (105.1)	1,184 (110.5)	0.728
6	TLAC	2,811.7 (705.6)	2,278.7 (744.9)	0.721
7	TLAC 12–2 pm	410.2 (113.9)	333.4 (125.3)	0.697
8	TLAC 4–6 pm	381 (112.2)	309.3 (116.4)	0.697
9	TLAC 2–4 pm	398.1 (116.7)	323 (121.5)	0.693
10	TLAC 6–8 pm	320.5 (118.4)	250.7 (108.7)	0.692
11	TLAC 10 am–12 pm	410.9 (127.4)	335.3 (132.3)	0.682
12	Mobility problem	768 (28.6%)	172 (57.9%)	0.672

Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group

- High dimensional multilevel and/or longitudinal time series data
  - Signal processing
  - Functional data analysis
  - Feature extraction
- Methodological challenges specific to wearables
  - Study protocol (device location, battery life, convenience, comfort)
  - Wear vs non-wear (complex missing data patterns)
- Computational challenges
  - Data storage
  - Data analysis

NHANES accelerometry

- The National Health and Nutrition Survey is a cross-sectional study of the US population performed in 2-year waves
- Complex survey structure
- Accelerometry data available for the 2003-2004 and 2005-2006 waves
  - Acceleration summarized into minute-level "activity counts"
  - Up to 7 days of data for each participant
  - Study protocol: remove the device at bedtime

NHANES accelerometry: data structure

- Accelerometry data downloadble from NHANES is in long format
- Very large file sizes ($\approx$ 2.5 GB)
- Unintuitive data structure

SEQN	PAXSTAT	PAXCAL	PAXDAY	PAXN	PAXHOUR	PAXMINUT	PAXINTEN	PAXSTEP
31128	1	1	1	1	0	0	166	4
31128	1	1	1	2	0	1	27	0
31128	1	1	1	3	0	2	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

NHANES accelerometry: proposed data strucutre

- Wide format instead of long format[1] (≈ 60 MB)
- 7 rows per participant, descending cronologoical order

	Unique Identifier		Quality Flags		NHANES wave	Activity Counts				
	SEQN	PAXDAY	PAXCAL	PAXSTAT	SDDSRVYR	MIN1	MIN2	MIN3	⋯	MIN1440
(a)	31128	1	1	1	4	166	27	0	⋯	0
	31128	2	1	1	4	0	0	0	⋯	0
	⋮	⋮	⋮	⋮	4	⋮	⋮	⋮	⋮	⋮
	31128	7	1	1	4	0	0	0	⋯	0
	⋮	⋮	⋮	⋮	4	⋮	⋮	⋮	⋮	⋮
(b)	31193	2	2	1	4	0	0	0	⋯	1921
	31193	3	2	1	4	335	2598	2185	⋯	46
	31193	4	2	1	4	0	0	0	⋯	0
	⋮	⋮	⋮	⋮	4	⋮	⋮	⋮	⋮	⋮
(c)	31880	2	2	2	4	32767	32767	32767	⋯	32767
	31880	3	2	2	4	32767	32767	32767	⋯	32767
	⋮	⋮	⋮	⋮	4	⋮	⋮	⋮	⋮	⋮
(d)	32008	5	1	2	4	0	0	0	⋯	0
	32008	6	1	2	4	NA	NA	NA	⋯	NA

NHANES accelerometry: *rnhanesdata* package

(1) Processed data	
processed physical activity data	"PAXINTEN_C.rda" and "PAXINTEN_D.rda"
wear/non-wear flags data	"Flags_C.rda" and "Flags_D.rda"
covariates data	"Covariate_C.rda" and "Covariate_D.rda"
mortality data	"Mortality_2011_C.rda" and "Mortality_2011_D.rda"
(2) Data processing functions	
NHANES activity processing code	"process_accel()"
NHANES wear/non-wear flag code	"process_flags()"
NHANES mortality	"process_mort()"
NHANES data merging	"process_covar()"
(3) Helper functions	
Calculate survey weights on subsets	"reweight_accel()"
Identify "good" days of accelerometry data	"exclude_accel()"
(4) Raw data	
NHANES covariate data	"ALQ_C.XPT","ALQ_D.XPT", "BMX_C.XPT","BMX_D.XPT", . . .
NHANES linked mortality data	"NHANES_2005_2006_MORT_2011_PUBLIC.dat" "NHANES_2003_2004_MORT_2011_PUBLIC.dat"

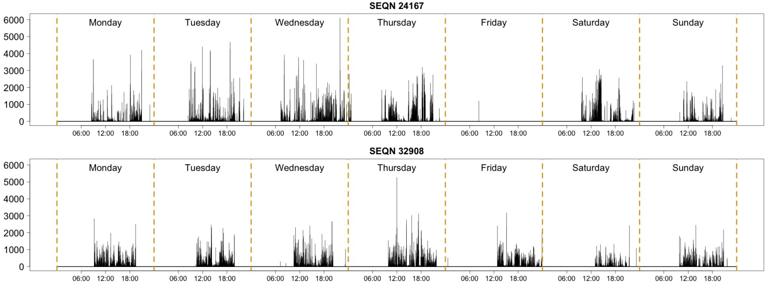
rnahnesdata: Package Installation

- Package installation may take a few minutes due to the size of the processed data.
- Requires the devtools package
- See ?"rnhanesdata-package" for details
- Published a paper on this [Leroux et al., 2019]
- A vignette: https://andrew-leroux.github.io/rnhanesdata/articles/NHANES_accelerometry_introduction.html

```
if(!require("rnhanesdata")){
    devtools::install_github("andrew-leroux/rnhanesdata")
    require("rnhanesdata")
}
```

NHANES accelerometry: EDA

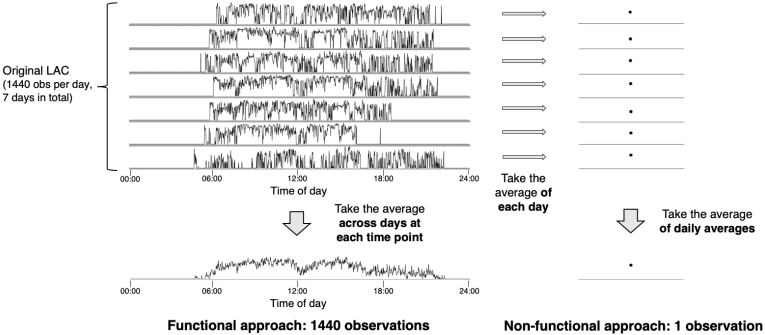
- 7 days of data for two participants at the minute level
- Estimated non-wear time has been imputed as 0
- Dominated by a few very large values



Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group

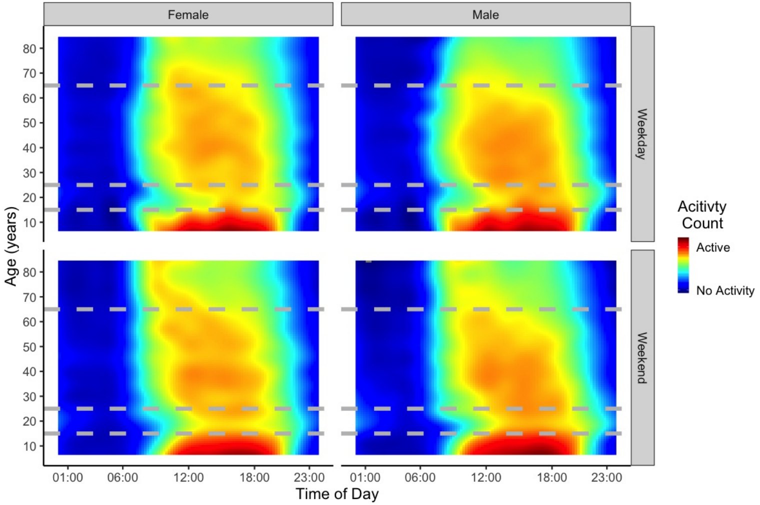
Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group

NHANES accelerometry: Data Reduction



NHANES Accelerometry: Epidemiology of PA and Aging

Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group

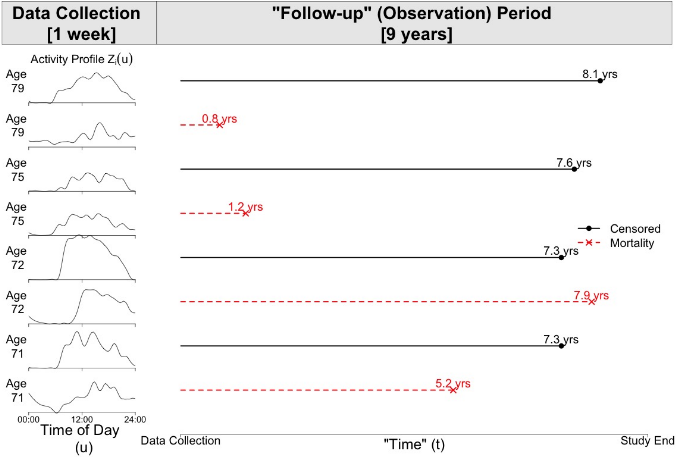


NHANES Accelerometry: Predicting Mortality

Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group

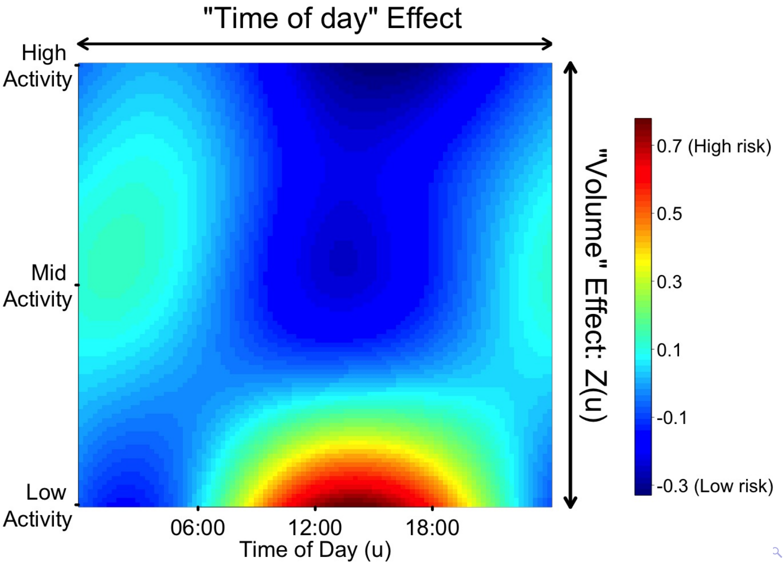
- **Context**
  - NHANES accelerometry data (2003-2006 waves)
  - 3,208 participants with at least 3 days of "good" data
  - Ages 50 to 85.
  - Administratively censor at 9 years
  - 599 events (mortality) with average time to event 4.3 years
- **Scientific Question**
  - Objectively measured PA associated with mortality
  - Does the association depend on diurnal patterns of activity? [Cui et al., 2021]

Accelerometry: Predicting Mortality



Accelerometry: Predicting Mortality

Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group



Lectures: Key Topics

1. Introduction to **smoothing** using (penalized) splines
   - Basis expansions
   - Generalized additive models (GAMs)
2. Functional principal component analysis
3. Functional regression
   - scalar outcome, functional predictor
   - functional outcome, scalar predictor
   - functional outcome, functional predictor

FDA Software

Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group

- If you haven't already, install
  - *R* free, available at `https://cran.r-project.org`
  - *RStudio* (optional) free, available at `https://rstudio.com`
- Software packages in *R*
  - *mgcv*
  - *refund*
  - Others (will introduce along with specific topics)

Goals of the Projects

Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group

- (Applied statistics) Apply FDA methods to answer a question relating to public health
- (Communication, written) Write up results in a manuscript format
  - Introduction
  - Methods
  - Results
  - Discussion
- (Communication, written / Reproducible research) Create a fully reproducible vignette presenting the methods and results
- (Communication, verbal) Present your project to the class

Project Details

Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group

- Project centered around a “new" analysis
  - Can be related to thesis/disseration work
  - Must incorporate a fully fleshed out data analysis
- Project will be done stepwise
  1. Proposal (due 04/08)
     - Identify and aquire a dataset
     - Propose a scientific question to be addressed using FDA methods
  2. Progress Report
     - In-class, works-in-progress presentations (04/20, 04/22)
     - Present your data and problem to the class
     - Get feedback from
  3. Final Presentations/Report
     - In-class presentations (5/11, 5/13)
     - Submit write-up (due at 11:59PM on 5/21)

Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group

FDA Working Group

- New working group on Functional Data Analysis
- Works-in-progress group
  - Discuss ongoing projects
  - Solicit imput on methods/approaches
  - Identify opportunities for
    - Collaboration
    - Funding
- Organizers/founders
  - Andrew Leroux
  - Julia Wrobel
- Meets bi-weekly, Wednesdays 2:30 PM
- First meeting, 03/31/2021

References I

Cui, E., Crainiceanu, C. M., and Leroux, A. (2021).
Additive functional cox model.
Journal of Computational and Graphical Statistics, 0(0):1–14.

Franklin, R. and ffrench-Constant, C. (2008).
Remyelination in the cns: from biology to therapy.
Nature Reviews Neuroscience, 9:839 – 855.

Leroux, A., Di, J., Smirnova, E., Mcguffey, E. J., Cao, Q., Bayatmokhtari, E., Tabacu, L., Zipunnikov, V., Urbanek, J. K., and Crainiceanu, C. (2019).
Organizing and analyzing the activity data in NHANES.
Statistics in Biosciences, pages 1–26.

Leroux, A., Xu, S., Kundu, P., Muschelli, J., Smirnova, E., Chatterjee, N., and Crainiceanu, C. (2020).
Quantifying the Predictive Performance of Objectively Measured Physical Activity on Mortality in the UK Biobank.
The Journals of Gerontology: Series A.
glaa250.

References II

Nakahara, J., Maeda, M., Aiso, S., and Suzuki, N. (2012).
Current concepts in multiple sclerosis: Autoimmunity versus oligodendrogliopathy.
Clinical Reviews in Allergy & Immunology, 42(1):26–34.

Smirnova, E., Leroux, A., Cao, Q., Tabacu, L., Zipunnikov, V., Crainiceanu, C., and Urbanek, J. K. (2019).
The Predictive Performance of Objective Measures of Physical Activity Derived From Accelerometry Data for 5-Year All-Cause Mortality in Older Adults: National Health and Nutritional Examination Survey 2003 - 2006.
The Journals of Gerontology: Series A.

Sweeney, E. M., Shinohara, R. T., Dewey, B. E., Schindler, M. K., Muschelli, J., Reich, D. S., Crainiceanu, C. M., and Eloyan, A. (2016).
Relating multi-sequence longitudinal intensity profiles and clinical covariates in incident multiple sclerosis lesions.
NeuroImage: Clinical, 10:1 – 17.

Logistics
Functional Data
Child Growth
MS Lesion
Wearable Devices
Course Outline
Software
Final Projects
FDA Working Group

References III

Sweeney, E. M., Shinohara, R. T., Shea, C. D., Reich, D. S., and Crainiceanu, C. M. (2013a).
Automatic lesion incidence estimation and detection in multiple sclerosis using multisequence longitudinal mri.
American Journal of Neuroradiology, 32:68 – 73.

Sweeney, E. M., Shinohara, R. T., Shiee, N., Mateen, F. J., Chudgar, A. A., Cuzzocreo, J. L., Calabresi, P. A., Pham, D. L., Reich, D. S., and Crainiceanu, C. M. (2013b).
Oasis is automated statistical inference for segmentation, with applications to multiple sclerosis lesion segmentation in mri.
NeuroImage: Clinical, 2:402 – 413.