Missing Data in Accelerometry

Amy H. Herring
Sara & Charles Ayres Distinguished Professor
Statistical Science, Global Health
& Biostatistics & Bioinformatics
Duke University

Based on collaborations with Kelly Evenson (UNC), Fang Wen (UNC), Nicole Butera (GWU), Siying Li (UNC), Chongzhi Di (UW), David Buchner (UIUC), Michael LaMonte (Buffalo), and Andrea LaCroix (UCSD)

February 24, 2020

Self-Reported Activity

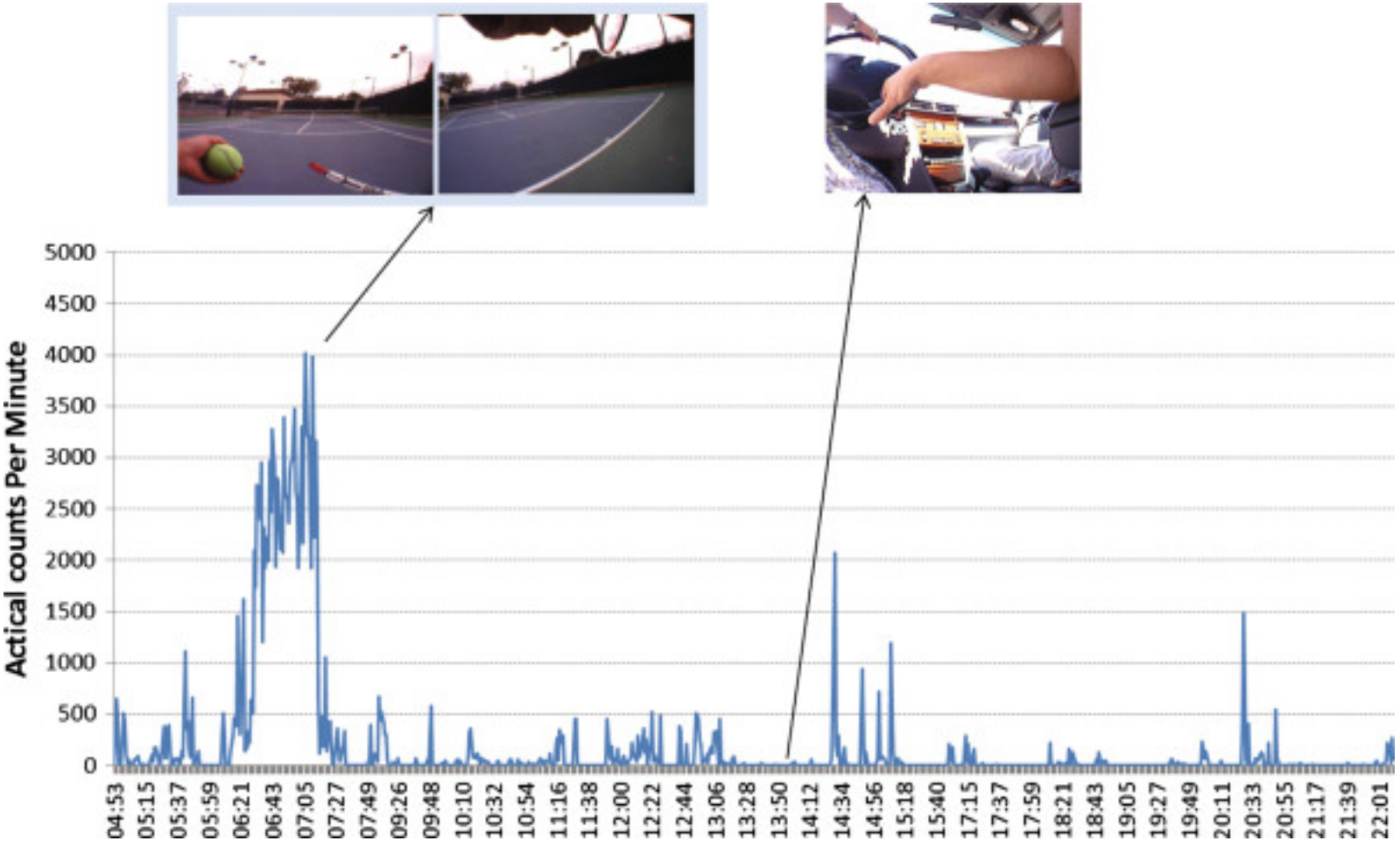
- May involve a time/activity diary or questionnaire regarding "typical" activity or activity in the past month, providing estimates of time and perceived intensity of activity; intensity can also be objectively estimated based on the named activity and lab-based values ("absolute" intensity)
- High participant burden, frequent missing/poor quality data (e.g., my mom characterizes all her walks as vigorous activity)
- Even complete data may be challenging to analyze
- Example: PIN physical activity study (Evenson et al)
  - Measured activity across multiple domains, e.g. leisure, work, outdoor/indoor household, child/adult care, and transportation
  - Self-reported time spent in activity often exceeds wake time or even 24-hour day
  - Some activities difficult to characterize
    - 8 hours of doing laundry
    - 24 hours of childcare

Objectively-Measured Activity

Common configuration is small accelerometer affixed to belt.

- Typical accelerometer records acceleration on 3 axes (vertical, anterior-posterior, medial-lateral) regularly (e.g., each 15 sec)
- 3-D raw acceleration data typically converted to univariate vector magnitude (VM) activity counts: $\sqrt{x_1^2 + x_2^2 + x_3^2}$
- VM counts often converted to compositional data (sedentary, light, moderate, vigorous activity) based on lab-based calibration studies
- Bouts are key
  - Common definition: 10+ min. of moderate-vigorous activity with ≤ 1 interruption of 1-2 min.
  - My workout often 6:00-6:45am spin class, fewer vigorous minutes outside of that time!

Typical Daily Count Tracing (Doherty et al., 2013)

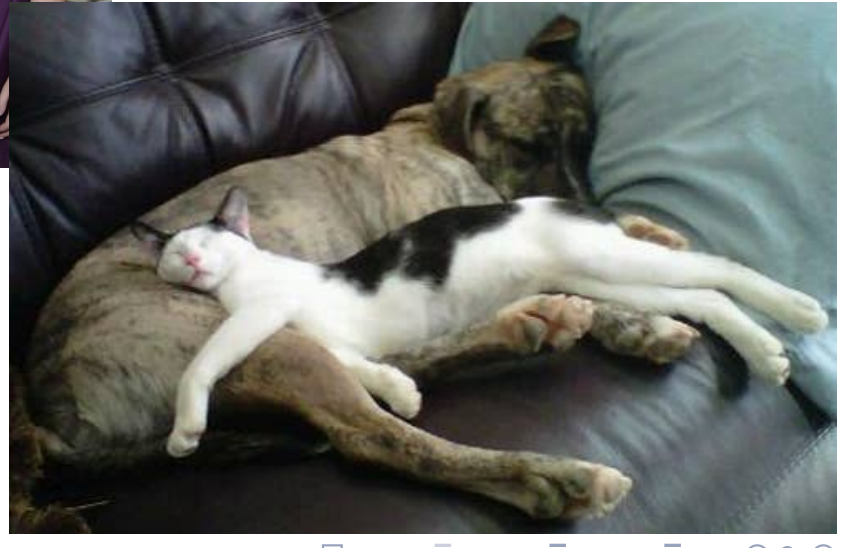


Participants in the small Doherty et al. study also wore a camera on a lanyard.

Objectively-Measured Activity

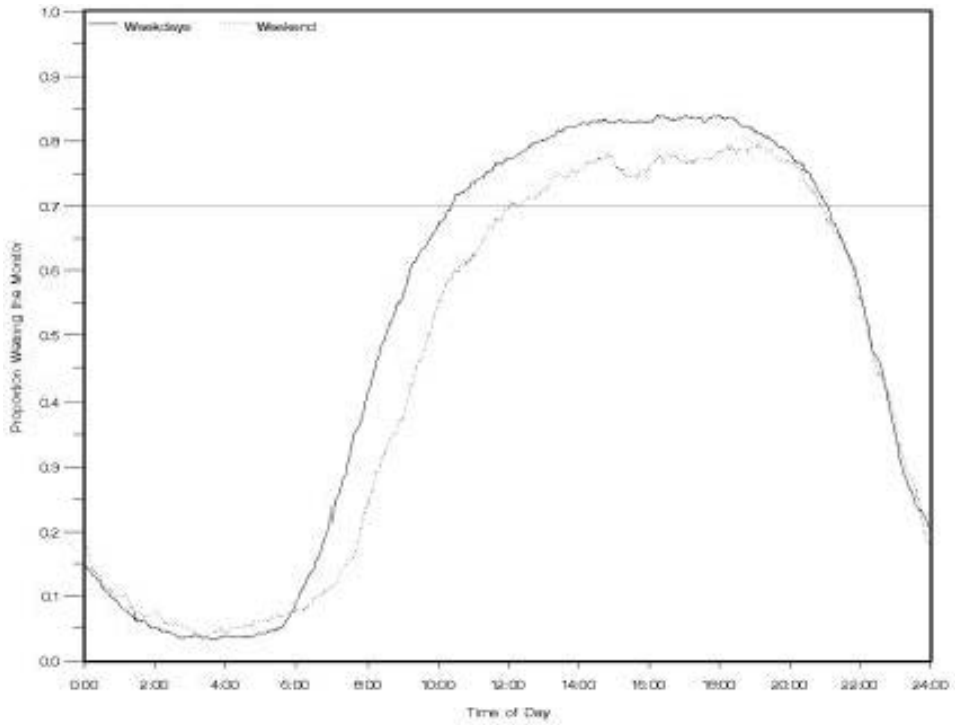
- Participants are often asked to wear an accelerometer all day for an entire week, or during waking hours
- Compliance, however, is an issue (wear time determined algorithmically, e.g. no acceleration for at least 90 minutes, with up to 2 minutes of non-zero data if upstream and downstream 30 minutes show no acceleration)

More Non-Compliers



Wear Time in Pregnancy, Infection, and Nutrition Study

- Asked to wear accelerometer except during sleep, bath, or swim for 7 days
- Average wear time <13 hours/day
- 70% of women "wore" 9am-9pm weekdays
- 70% of women 10:45am-9pm weekends



Wear Time in WHI-OPACH

Women in the OPACH study were asked to wear the accelerometer for seven consecutive days during waking and sleeping hours, except during bathing or swimming.

- 95.3% of women had some non-wear time
- Median non-wear time 12.6% of the 7-day period
- Only 2.4% of days completely missing

Typical Approaches to Handling Non-Wear Time

- Consider data only from adherent days of adherent participants: "complete case (CC)"
  - Adherent: e.g., wear accelerometer $\geq$ 10 hours/day on 4-7 days
  - Often scale everyone to 10 hours/day (daytime) for comparability
  - Assumes adherent and non-adherent participants and days comparable
- "Available case (AC)" analysis
  - Use all observed data
  - Often scale everyone to 10/24 hours/day for comparability
  - Again, strong assumptions about activity during non-wear times

Imputation and Other Approaches to Handling Missing Data

Challenges include

- Selection of level on which to impute data
  - Often data summarized and imputed at daily or even weekly level (e.g., average min. of vigorous, moderate, light, sedentary activity), ignoring compositional nature of summary data (e.g., $>$24 hrs/day activity) or associations among measures (e.g., vigorous bouts and minutes of vigorous activity)
  - Data on original scale (e.g., 3 accelerations every 15 sec.) are high-dimensional correlated time series
- At either level, data are highly skewed and zero-inflated, without an obvious mapping to a parametric distribution
- Data correlated within individuals
- Continuous episodes (bouts) often of interest

Imputation and Other Approaches to Handling Missing Data

Challenges also include incorporation of relevant variables, e.g.,

- Time of day
- General activity level of the individual
- Characteristics associated with activity (e.g., age, BMI, physical function)
- Activity levels immediately before and after the missing data
- Potential availability of self-reported data (of varying quality), which range from reports of average activity levels in a week (e.g., swimming 3 days/week), to detailed sleep (and occasionally, activity) diaries
  - We found Spearman correlations of -0.09-0.27 for moderate-to-vigorous hours/week between self-reported perceived and accelerometry-measured activity, and slightly higher correlations with self-reported MET- (intensity-adjusted) hours/week and accelerometry (0.13-0.38)

Hot Deck Imputation

Working in a resource-limited setting, we explored using an *ad hoc* hot deck imputation to handle imputation of the raw data,

- on any desired scale (e.g., 15 sec)
- prioritizing imputation within- over across- similar individuals
- allowing imputation of entire time windows of missingness at once to preserve activity "bouts"
- without needing to model complex distribution of the data

Hot Deck Imputation Procedure

For each instance (window of missing values), we did the following.

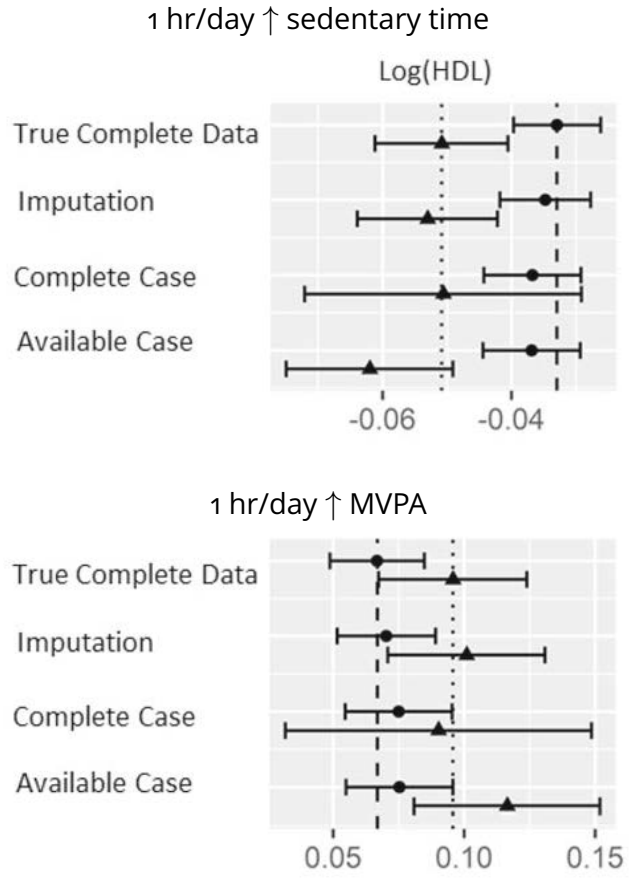
- Construct donor pool (e.g., matched on time of day, BMI, age, physical function + self-observations)
- In each imputation, randomly sample from constructed donor pool
  - May be possible to use one hot deck “match” to impute each missing instance
  - Long missing “instances” may require patchwork of donors due to prevalence of missing data in practice

Donor pool for Herring gull

Simulation Study: PA and HDL (Good) Cholesterol

- Simulation study used bootstrap samples of a "complete" dataset derived from compliers in OPACH
- MAR data imposed mimicking missing data patterns in the study
- Evaluated bias and coverage of association of 1 hour/day increase in activity domain on HDL level

→ Hot Deck Imputation >> CC, AC approaches



Truth represented by dashed (24-h day) and dotted (daytime) lines; ● estimates over 24-h day and ▲ daytime only

Interesting (to Me!) Issues

- Optimal weighting of sparse observations from self-donors versus observations from other individuals
- Characterizing uncertainty
- Proper imputation or modeling approaches that preserve bouts
- Better incorporation of self-reported activity data
- Non-ignorable missing (and observed!) data
- Ease of use/implementation for epidemiologists

Another primary interest: robustness/generalizability of dimension reduction; equivalence of classes with respect to outcomes – e.g., weekend warriors vs. daily routine