Hallmarks of Health Big Data

Nathaniel Osgood

Nathaniel Osgood
Measuring and Modeling Health Behavior with Smartphone Mediated Data Collection
8-10-2016
Some Sources of “Big Data” Related to Health Behaviour

- Twitter (feeds)
- Facebook (status updates)
- Environmental sensors (weather, municipal, building)
- Lab test results
- Point of sale records
- Administrative data
- Questionnaire responses (mobile, web)
- Sequence data
- Voice audio
- Incoming/outgoing calls
- Communication infrastructure proximity data
- Health information browsing behavior
- Consumer electronic devices sensors (physical activity, proximity, location, etc.)
Four key “V’s” of “Big Data” (Google)

- **Volume**: Lots of evidence

- **Velocity**: High temporal resolution longitudinal data

- **Variety**: Cross-linked data sources support can “triangulation” of understanding

- **Veracity**: Physical measures are less subject to self-report, on-device self-reporting is more temporally proximate to health event (exposures, symptoms,…)
Volume

• Consider
  • N participants
  • # of records per participant (M)

• Traditional epidemiologic studies: \( N \gg M \)

• “Big data”: \( M \gg N \) common

• Common: Dozens of MB per participant/day

• This volume of data will often require different handling techniques than for traditional systems: Different
  • Storage
  • Analysis
  • Visualization
Volume & Variety: Some Statistics

Ethica - Chromium

ethicadata.com/dashboard?study=85

Overview Compliance Activity Location Battery Surveys

General

Name: New Tech for Foodborne Disease - Phase 2 Group 2
Organization: School of Public Health, University of Saskatchewan
Duration: 75 days from Jan 21, 2016 to Apr 04, 2016
Upload Server: Ethica servers

Streams:
- Gravity: 59.542.737
- Linear Acceleration: 59.476.067
- Orientation: 59.380.202
- WiFi: 8.847.544
- GPS: 7.574.966
- Survey Responses: 4.862
- Battery: 444.541
- Accelerometer: 91.560.665
- Gyroscope: 121.023.090

Registration Code: 85

Surveys

Applications 09:42  Ethica...
Velocity

• Electronic data sources often update frequently
  • Low rates: Lab data, administrative data
  • Medium Low rate: multiple times/day e.g.,
    • Facebook updates
    • Twitter
    • browsing behavior
    • app use
    • Ecological momentary assessment (EMA) responses
  • Weather
  • Point of sale
  • Medium rate: on order of seconds (e.g., GPS, building sensors)
  • Higher rates: Many times per second (e.g., accelerometers, gyroscopes)
• Such velocity provides high temporal resolution into micro-behaviours and exposures
Variety

- A given electronic data source often provide multiple lines of evidence
  - Smartwatch (e.g., Empatica E4): stress responses via electrodermal activity & Heart Rate Variability, heart rate, acceleration, skin temperature
  - Smartphone with Ethica iEpi: location, physical activity, proximity, posture, humidity, EMA responses, etc.

- For a given participant, we increasingly have multiple sources of electronic data available – both quantitative and qualitative
  - Smartphone (context and state via sensors, ecological momentary assessments)
  - Smartwatch
  - Weather
  - Point of sale
  - Facebook updates

- This evidence is cross-linked by participant and time (i.e., for a given participant & time, we can find the relevant information applying then across all data sources

- We can often triangulate state of a given participant using many lines of evidence
Example Biography: Waterborne Illness

- Bathed at Newport Beach
- Measured Risk Rating
- Beach Advisory Posted
- Beach Closure Notice Posted
- Occurrences of HCGI
- Sought Care
Example Biography: Lyme Disease

- Removed Tick
- Rash Appeared
- Joint Pain
- Stiff Neck
- Encountered Lyme Disease Advisory
- Sought Care
Activity Level over Geography
Social Interactions During and After Work

Density of people visited by two sample participants per location:
a) during the experiment, b) during working hours c) outside working hours

Sample Participant 1

Sample Participant 2
Cross-Linking of Data: Metcalf’s Law

- Opportunities for cross-linking of data => values rises as square of number of sensors

- Example cross linking (BT=Bluetooth):
  - Food Record/BT/Wifi (timing, social & geographic/spatial context of food consumption)
  - Accelerometer/GPS (with GIS)/BT (how does physical activity level change near parks? In high crime areas? How change around other people? With weather?)
  - BT/GPS/Wifi (estimates of contact location, understanding of social context/capacity of contacts)
  - BT/Wifi/GPS (indoor & outdoor positioning)
  - GPS & Accelerometer: Triggering more rapid measurement of accelerometer if moving quickly
  - Triggered surveys and any sensor: Disambiguation
Veracity

• Individual measurands can be more accurate than self-report
• Collectively, measurands can often offer strong accuracy benefits
Veracity

• Individual measurands can be more accurate than self-report
• Collectively, measurands can often offer strong accuracy benefits
A Major Barrier: Reliable evidence regarding key health behaviors & exposures is hard to secure

- Location (access to care, access to resources, barriers to activity, environmental risks)
- Physical activity (obesity, T2DM & GDM, risk of falls)
- Spatial proximity (transmission of pathogens and norms, imitative behaviour, communication)
- Social context (norms, communication, perception of safety, risk perception)
- Communication: Personal & mass media (risk perception, norms, beliefs, social cues)
- Decision-making rules & heuristics
- Exposure to environments, behaviour, etc.

Absent understanding of such behaviours & exposures, the potential to quantitatively evaluate policy trade-offs is greatly limited.
Accompanying Access & Participation Challenges

• Identifying and influencing social groupings associated with adverse health & risk behavior

• Reaching traditionally vulnerable & underrepresented groups
  • Low SES/Youth/Minority/Homeless/Mobile populations

• Need for flexible means of trialing new strategies

• Economical long-term studies and surveillance in an age of declining opt-in or participation in traditional methods such as random digit dialing
Contact Duration

CCDF of Contact Duration

- Flunet
- Cambridge/Haggle
- HED1 - Participants Only
- HED1 - All Mobile Devices
Contacts Week 2
Ethica

Where is the source of this food?

- Food purchased on campus
- Ready-to-eat food purchased off campus
- Eating at restaurant
- Eating food prepared at home

Please take a photo of the food you are eating.
Veracity

• Individual measurands can be more accurate than self-report
• Collectively, measurands can often offer strong accuracy benefits
  • “Triangulation” from multiple data sources often greatly lowers risks of ambiguity
• Examples: Classification of
  • Posture (e.g., Sitting/Standing/Lying down/Off-person)
    • Relevant measurands: Accelerometer, gyroscope, orientation, charging status
  • Indoor/Outdoor. Relevant measurands: GPS strength & availability, WiFi signal strength, charging status, light levels, temperature vs. temperature for region, weather, etc., time of day
  • In vehicle vs. not. Relevant measurands: GPS speed & strength & availability, acceleration, WiFi signal strength, gyroscope charging status, light levels
  • Smoking vs. not. Relevant measurands: Outdoor/indoor status, hand acceleration, heart rate, heart rate variability, social surrounds, GPS location, time of day
Some Attractive Characteristics of Health Big Data

• (Intensively) longitudinal
• Physical measurements
• Cross-linking w/quantitative& qualitative responses to context-sensitive EMAs
• Sufficient temporal resolution to
  – Eliminate many effects of recall bias
  – Recognize many repeated day-to-day patterns
  – Observe many exposures (e.g., via social env., physical proximity to physical activity) & outcomes
• Capacity to “triangulate” from multiple measurands
  – Deductions about behavior
  – Exposure/Response
• Automatic collection => Low Burden