**The Healthy Behavior Data Challenge**

Phase 1 Submission Template

**Website**:

Additional Information:

Information on the Behavioral Risk Factor Surveillance System can be found at [www.cdc.gov/brfss](http://www.cdc.gov/brfss). Details on the HBD Challenge may be found at challenge.gov.

For Further Information Contact: Dr. Machell Town at BRFSSinnovations@cdc.gov.

**Submission Deadline**:

1. Challenge Team Information

|  |  |  |
| --- | --- | --- |
| Team Name |  |  |
| RTI PGHD | | |
| Team Lead |  | City/Province |
| Robert Furberg |  | RTP, NC |
| E-mail |  | Phone Number |
| [rfurberg@rti.org](mailto:rfurberg@rti.org) |  | 919-452-8021 |
| Subject-matter/domain expertise |  |  |
| Digital Health & Clinical Informatics |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Team Member #1** |  | **E-mail** |  | **Subject-matter/domain expertise** |
| **Kelly Evenson** |  | [kelly\_evenson@unc.edu](mailto:kelly_evenson@unc.edu) |  | **Phyical Activity Epidemiology** |
| **Team Member #2** |  | **E-mail** |  | **Subject-matter/domain expertise** |
| **Jacqueline Bagwell** |  | [jbagwell@rti.org](mailto:jbagwell@rti.org) |  | **Clinical Informatics** |
| **Team Member #3** |  | **E-mail** |  | **Subject-matter/domain expertise** |
| **Ron Thigpen** |  | [rthigpen@rti.org](mailto:rthigpen@rti.org) |  | **Technical Development** |
| **Team Member #4** |  | **E-mail** |  | **Subject-matter/domain expertise** |
| **Stephanie Eckman** |  | [seckman@rti.org](mailto:seckman@rti.org) |  | **Survey Methodology / Sampling** |
| **Team Member #5** |  | **E-mail** |  | **Subject-matter/domain expertise** |
| **Rainer Hilschner** |  | [rhilscher@rti.org](mailto:rhilscher@rti.org) |  | **Data Science** |
| **Team Member #6** |  | **E-mail** |  | **Subject-matter/domain expertise** |
| **Jay Rineer** |  | [jrin@rti.org](mailto:jrin@rti.org) |  | **Geographic Information Systems** |

|  |
| --- |
| **Are all team members residents of the United States?** |
| **Yes** |

1. Organization (if submitting on behalf or as part of an organization)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Organization Name** |  | **Website** |  | **Type of Organization** |
| RTI International |  | [**www.rti.org**](http://www.rti.org) |  | **Nonprofit,** *501(c)(3*) |

1. How did you find out about this challenge?

|  |
| --- |
| Tiwtter |

1. Submission Overview

|  |
| --- |
| **Project Title** |
| **RTI PGHD + EMA + GPS** |
| **Project Overview** |
| We propse to capture and conflate three sources of data for analysis: 1) active (aka, self-report via ecological momentary assessment measures), 2) passive (aka, sensor data on human performance, activity, and physiology via consumer wearable devices), and 3) spatial (aka, location data derived from smartphone via global positioning system). We will specifically capture data on three major indicators: 1) Physical Activity, 2) Sendentary Behavior, and 3) Sleep.  1. Active data collection  EMA measures for the proposed study will be compiled in an application that has already been developed to support data collection using both Android and iOS devices.  Affect  The Photographic Affect Meter (PAM) is a validated, one-item, mobile App-based measure of affect that correlates with the widely used PANAS scale. In PAM, subjects select from a grid of 16 images the one that best represents their current emotional state. The 4 x 4 grid is modeled with valence (negative-positive, left to right columns) as the x-axis and arousal (low-high, bottom to top rows) as the y-axis. The PAM instrument has been used in a number of studies to evaluate affect and we have successfully used the instrument to study effects of social support on daily chronic pain and affect among chronic pain patients.  Sleep Quality  Sleep quality plays an important role in a person’s health and functioning and is connected with fatal diseases, such as diabetes and obesity. Established measures are available for assessing sleep quality (e.g. Pittsburg Sleep Quality Index, but are too long to be used for EMA. EMA researchers have identified the Daytime Symptoms in Insomnia Scale (DISS), which includes the following four daytime symptom factors that are highly correlated with sleep symptoms and reports: alertness, positive and negative mood and sleepiness/fatigue. The PAM affect measurement is well-suited to capture the first three factors in DISS. Sleepiness/fatigue will be measured using the following three items that comprise the fourth sleepiness/fatigue factor of the DISS Scale: “How (fatigued, sleepy, or exhausted) do you feel?” with items rated on a visual analog scale ranging from 1 to 100.  2. Wearables data will be captured using RTI’s Participant-Generated Health Data infrastructure which leverages Validic (<https://validic.com/>) services to enable extraction of data from over 200 connected applications and devices.  3. Consistent with our prior work, we will leverage the third-party application Moves to obtain continuous spatial data from participants (<http://bjsm.bmj.com/content/early/2016/06/10/bjsports-2016-096103>) |

1. Indicators to be measured (the indicators listed below are not comprehensive and innovators are recommended to include other relevant indicators)
   1. Physical Activity

* Amount of MVPA[[1]](#footnote-0) time per day
* Amount of MVPA time per day obtained in bouts of 10 minutes or more
* Amount of MVPA time accrued while at work, at home and/or in transit
* Identification of times during the day where MVPA is high
* Daily number of steps
* Miles/km (Distance) on foot or other modes of active transportation
* Frequency of MVPA
* Calories burned
* Type of activity (aerobic, strength, etc.)
* Level of activity (low, moderate, high)
* Time spent in different domains of MVPA (home/occupational, travel and recreational)
* Location of MVPA (recreation facility, at home, at work, on sidewalk/bike lane)
* Perception of safety while active
* Enjoyment level of the MVPA
* Number/flights of stairs climbed
* Average and peak heart rate
* Hours per week adults spent in sports, fitness or recreational physical activities
* Other indicators
  1. Sedentary Behavior[[2]](#footnote-1)
* Amount of time per day spent sedentary, excluding sleep time
* Amount of time per week spent on a computer/screen including watching TV, videos, playing computer games, emailing or using the internet
* Amount of sedentary time accrued while at work, at home and/or in transit
* Sitting time at work/ number and frequency of breaks at work from sedentary time
* # of hours spent in a car or motor-vehicle
* Other indicators
  1. Sleep
* Hours of sleep per night (sleep duration)
* Amount of time awake after sleep onset
* Sleep efficiency
* Amount of time to fall asleep (i.e., sleep latency)
* Consistency of bedtime
* Consistency of wake time
* Amount of time in REM vs. non-REM sleep (duration of sleep stage)
* Type of activity directly before sleep (e.g., screen time, reading, TV)
* Sleep satisfaction in morning
* Daytime sleepiness
* Other indicators

1. Summary of proposed data source(s) (complete applicable sections)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Data Source** | | **Data Accessibility** (e.g., API, specialized software, existing data set) | **Data Cost** (i.e., fee for access, open access) | **Data Recency and Update Frequency** (i.e., how recent is the data and how often is it collected) | **Applicable Functional Area(s) and Indicator** (i.e., physical activity, nutrition, sleep, and/or sedentary behavior) | **Existing Users of the Data Source** (i.e., identify examples of organizations or other groups that have or are using the data source) |
| **Organization (e.g., company)** | Method of Collection (e.g., wearable, self-reported) |
| *1* | RTI | Self-report / EMA | Specialized software | Sample recruitment and incentives | These data will be collected daily in the proposed pilot | Serves as source of self-report to enhance passive data collection | None, this is a novel data source |
| *2* | RTI | Wearable | Extracted via API | Recruitment / incentives | These data will be captured daily | Physical activity, sedentary behavior, sleep | None, this is a novel data source |
| *3* | Moves | Smartphone | Extracted via API | Recruitment / incentives | These data will be captured daily | Concurrent location and activity data | None, this is a novel data source |
| *4* |  |  |  |  |  |  |  |
| *5* |  |  |  |  |  |  |  |
| *6* |  |  |  |  |  |  |  |
| *7* |  |  |  |  |  |  |  |
| *8* |  |  |  |  |  |  |  |
| *9* |  |  |  |  |  |  |  |
| *10* |  |  |  |  |  |  |  |

1. Describe how the data that you will use provides information and insight that is complementary to or more novel and innovative than that currently utilized for public health surveillance by CDC? (Novelty/innovation can apply at the level of the individual data source(s) selected, the specific indicators to be measured, tools/solutions that are used to capture the data, or result from newly created linked data sets).

Ubiquitous mobile phones offer personalized data-driven insights into their user’s lifestyles, habits, and daily routines through data generated as users move about and engage with social technologies. The use of passive data in place of survey data can reduce respondent burden and improve measurement quality. However, variables such as beliefs, attitudes, emotions, and intentions, are still best collected via surveys. For this reason, Mick Couper and others have argued that the skillful combination of survey and passive data is the future of our industry. We propose a pilot to demonstrate the infrastructure to administer short surveys via smartphones and tablets while also gathering passive data from the device and other wearable devices. Rather than asking respondents a large set of questions, clients could ask a few questions and infer additional variables from the passively-collected data.

1. Describe the process you will use to link the data from the different sources you’ve identified. Include a description of feasibility and any considerations that will be made to ensure the privacy, security and confidentiality of the data and data subjects throughout this process.

We will not capture any PII/PHI and refer reviewers to our prior work described in the publication entitled, “Establishing Linkages Between Distributed Survey Responses and Consumer Wearable Device Datasets:A Pilot Protocol” to describe our approach.

<https://www.researchprotocols.org/2017/4/e66/>

1. Describe how the linked data set(s) or individual data source(s) will be used to develop values for your proposed set of metrics in sleep, sedentary behaviors, nutrition, and/or physical activity.

We refer reviewers to the uploaded data fields document that inventories variables captured via passive data collection, specifically those related to sleep, sendendary behavior, and physical activity.

1. Describe the representativeness of your data set for public health surveillance (e.g., to what population groups or sub-groups can you meaningfully extrapolate the results of your data set?). How amenable will this data set be to disaggregation by age, gender, education, geography, or other demographic characteristics?

Consistent with the methods described in the paper linked to in item 8, we will initiate recruitment via Mechanical Turk, screen participatns by age, gender, education, geography, or other demographic characteristics and invite a sufficient number of respondents to result in a final active sample of ~300 individuals to participate in one month of data collection.

1. How useful will your data set be for public health surveillance, how significant/relevant and generalizable are the results that you expect to obtain? (

We refer reivewers to our recent publication entitled, “Assessing Validity of the Fitbit Indicators for U.S. Public Health Surveillance”

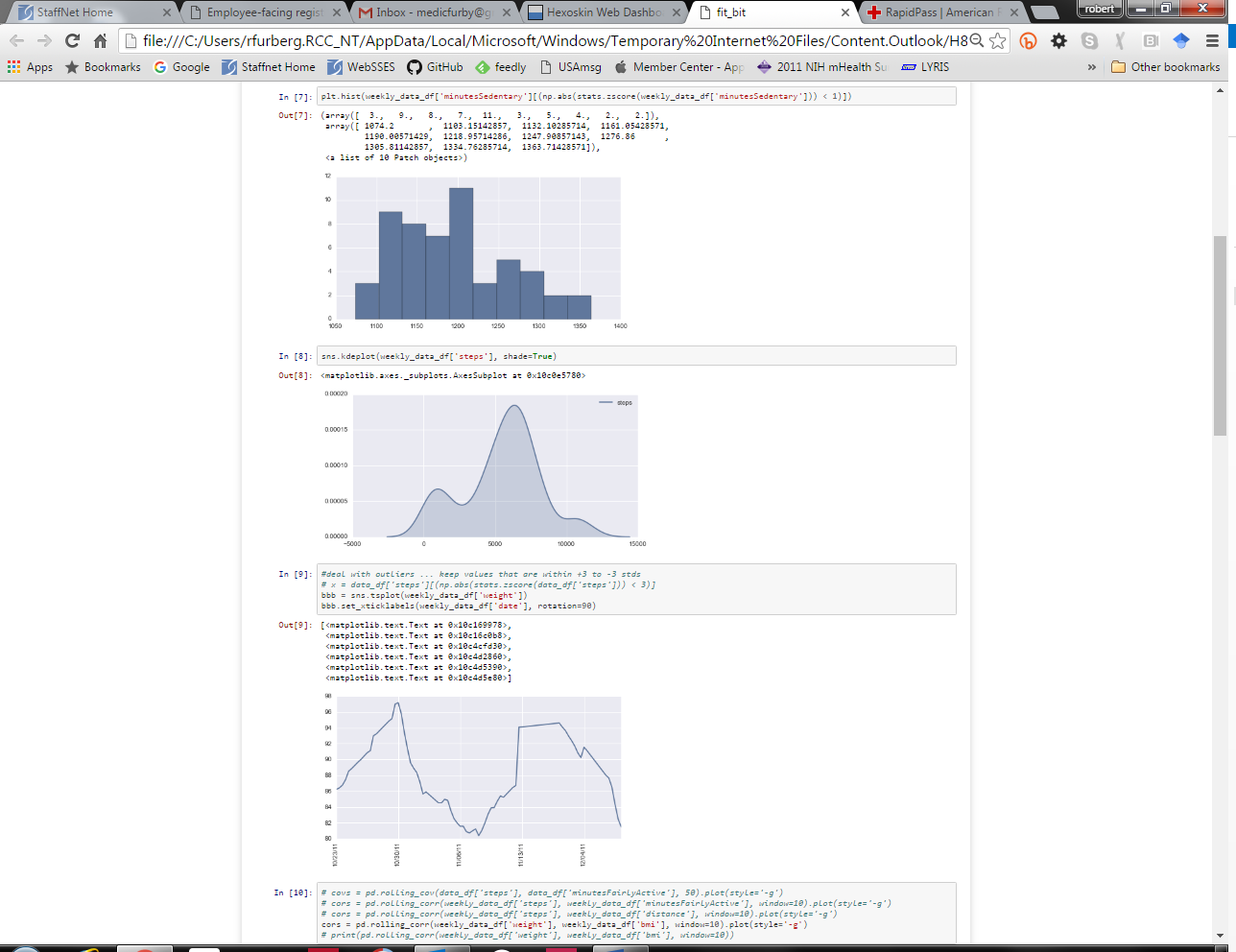
<http://www.ajpmonline.org/article/S0749-3797(17)30304-5/fulltext>

This resource provides a comparison between BRFSS data and those derived from a single vendor of wearable devices, highlighting the need for

1. Will the proposed project’s data and data sets contain information of relevance to other areas of public health surveillance (e.g., chronic or infectious disease)? If yes, please specify and describe any additional work that would be required in order to expand applicability.
2. Please describe a 3.5-month plan to develop a working prototype during the second phase of this challenge. This should include:

Please refer to the publications described above for additional details.

We will initiate screening, consenting, and enrolling individuals resulting in a representative sample of 300 participants via Mechanical Turk. The respondents will be enrolled using case identifiers to index and establish linkages between active, passive, and spatial data.

Fitbit users generally have only access to daily data snapshots aggregated from the underlying minute data Fitbit is actually recording. For an internal RTI data science project conducted in 2015, we gained access to 4 years of minute-level Fitbit data with the intent of identifying trends that are not visible in the corresponding 24h level data set. To do so we developed a flexible open source pipeline written entirely in Python that handles the entire process from ingesting the data to visualizing it. For the underlying data structure we chose Pandas dataframes (a Python based re-implementation of the R dataframe structure) to achieve maximum flexibility for analyzing, manipulating and augmenting data. To be able to easily explore the data set we chose the Jupyter (formerly IPython) interactive computational environment. Jupyter is a web application that allows users to create “living” documents with executable code, equations and text encapsulated in individual cells. Aside from enabling developers to quickly compose complex documents, Jupyter notebooks can easily be shared with other users. Documents can also be downloaded as html or PDF files. To visualize output we used the default Python matplotlib library as well as Seaborn (other visualization libraries such as plotly are possible) as show in Figure 1. One of the reasons we chose Pandas was the ability to easily augment Fitbit data with external data sources. We were interested, for example, in correlating weather conditions with activity patterns so we downloaded weather data from Weather Underground and added it to the existing Fitbit dataframe. We anticipate using similar methods to correlate wearable data from the cohort with data from other sources using either the Bokeh or the plotly library.

Initial analysis will focus on providing general, descriptive statistics on technology adoption and the characteristics of various subgroups, including high or low adopters, and the characteristics of those participants. We will conduct exploratory data analysis and creative data fusion with other external datasets of interest (e.g., environmental, geographic data). 

|  |
| --- |
| This project will explore how the combination of survey responses, wearable device data, and location tracking can improve how we monitor changes in health in our population. The pilot will demonstrate how the use of apps and wearables can provide detailed information on the health of our population. |

1. Significance and Relevance Summary

1. Moderate-to-vigorous physical activity (MVPA) is any activity with an energy expenditure >3 metabolic equivalents [↑](#footnote-ref-0)
2. Sedentary behavior is any waking activity characterized by an energy expenditure ≤ 1.5 metabolic equivalents and a sitting or reclining posture [↑](#footnote-ref-1)