Form Approved

OMB No. 0990-0390

Expiration Date 04/30/2018

**The Healthy Behavior Data Challenge**

Phase 1 Submission Template

**Introduction**

The Healthy Behavior Data Challenge responds to the call for new ways to address the challenges and limitations of self-reported health surveillance information and tap into the potential of innovative data sources and alternative methodologies for public health surveillance.

The Healthy Behavior Data Challenge will support the development and implementation of prototypes to use these novel methodologies and data sources (e.g., wearable devices, mobile applications, and/or social media) to enhance traditional healthy behaviors surveillance systems in the areas of nutrition, physical activity, sedentary behaviors, and/or sleep among the US adult population aged 18 years and older.

The collection of health data through traditional surveillance modes including telephone and in-person interviewing is becoming increasingly challenging and costly with declines in participation and changes in personal communications. In addition, the self-reported nature of responses particularly in the areas of nutrition, physical activity, sedentary behaviors, and sleep has been a major limitation in these surveillance systems, since self-reported data are subject to under/over reporting and recall bias. Meanwhile, the advent of new technologies and data sources including wearable devices (Fitbit, Garmin, Adidas, Jawbone, smart watches, activity trackers, etc.), mobile health applications on smartphones or tablets, and data from social media represents an opportunity to enhance the ability to monitor health-related information and potentially adjust for methodological limitations in traditional self-reported data.

The Healthy Behavior Data Challenge will harness this potential and identify feasible alternative options for collecting health-related behaviors in new ways. Conducted in two phases, Phase I (Prototype Development) entails Challenge participants developing a concept proposal for obtaining data collected from wearable devices, mobile applications and/or social media for public health surveillance purposes.

The Healthy Behavior Data Challenge participants will propose data sources and approaches for aggregating data from wearable devices, mobile applications and/or social media in the areas of nutrition, physical activity, sedentary behaviors, and/or sleep. In Phase II (Prototype Implementation), a subset of submissions (up to 3) with promising concepts will be invited to test their proposed approaches for ongoing public health surveillance.

**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 |  |  |
| Small Data Lab | | |
| Team Lead |  | City/Province |
| Deborah Estrin |  | New York |
| E-mail |  | Phone Number |
| [destrin@cornell.edu](mailto:destrin@cornell.edu) |  | (310) 866-0116 |
| Subject-matter/domain expertise |  |  |
| Computer Science |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Team Member #1** |  | **E-mail** |  | **Subject-matter/domain expertise** |
| **Aaron Baum** |  | **aaron.baum@mssm.edu** |  | **Economics** |
| **Team Member #2** |  | **E-mail** |  | **Subject-matter/domain expertise** |
| **JP Pollak** |  | **jpp9@cornell.edu** |  | **Information Science** |
| **Team Member #3** |  | **E-mail** |  | **Subject-matter/domain expertise** |
|  |  |  |  |  |
| **Team Member #4** |  | **E-mail** |  | **Subject-matter/domain expertise** |
|  |  |  |  |  |
| **Team Member #5** |  | **E-mail** |  | **Subject-matter/domain expertise** |
|  |  |  |  |  |
| **Team Member #6** |  | **E-mail** |  | **Subject-matter/domain expertise** |
|  |  |  |  |  |

|  |
| --- |
| **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** |
| Cornell Tech |  | tech.cornell.edu |  | University |

1. How did you find out about this challenge?

|  |
| --- |
| Mailing list |

1. Submission Overview

|  |
| --- |
| **Project Title** |
| **Pushcart: measuring nutrition using digital receipts** |
| **Project Overview** |
| Describe in 500 words or less:   * What aspects of sleep, physical activity, nutrition, and sedentary behavior do you propose to report on and why are they important for public health surveillance? * Provide a brief description of the source(s) of data that will be used to report on these aspects, how your team proposes to access them, and why they are appropriate for use in public health surveillance? * How do you see your concept improving on current public health surveillance in the areas of sleep, physical activity, nutrition, and sedentary behaviors?   Pushcart is an effortless and compatible passive data collection tool that provides objective, high frequency, household-level data on food consumption and nutrition. Pushcart can serve as a proxy for variety of nutritional metrics on the household level including: total calories consumed and division to different food groups (e.g. fruits, vegetables, sugar) continuously on a weekly basis.  Pushcart provides an objective measure of nutritional habits by analyzing online shopping email receipts from grocery stores (e.g. Amazon) and prepared-food delivery services (e.g. seamless). Pushcart is plug-and-play compatible with any service that generates email receipts. There is no single technology platform or grocery service to which Pushcart is tied and the service provider does not have to participate. The consumer unilaterally signs up for Pushcart by permitting all emailed receipts to be automatically sent to the platform for processing.  Public health surveillance of nutrition is hard. Traditional surveillance methods of individual nutrition require unrealisticly high engagement by completing daily food diaries or over relay on simple self-report measures that are subject to under-reporting and recall bias, which is particularly disadvantageous for groups with poor memory and/or low levels of literacy. Pushcart can provide an objective and effortless measure of nutrition and food consumption that could give insight about trends and improve existing public surveillance methods. |

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

Nutrition:

* Total calories consumed per day
* Total calories from fat (saturated, not-saturated)
* Sodium consumed
* Total carbs consumed
* Total protein consumed
* Total consumption of vitamins
* How often fruit (not including juices) was consumed (day, week, or month)
* How many times per day/week/month a green leafy or lettuce salad, with or without other vegetables, was eaten
* How often vegetables (not including lettuce salads and potatoes) was eaten (day, week, or month)
* Number of sugar-sweetened beverages consumed in a week (or per day)
* Number of caffeinated drinks consumed in a week (or per day)

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* | Grocery service (e.g. Instacart, Amazon Fresh) | Email receipts | Gmail API | None. | Retrospective and ongoing with high frequency. | Nutrition | Unknown |
| *2* | Food delivery service (e.g. Seamless) | Email receipts | Gmail API | None. | Retrospective and ongoing with high frequency. | Nutrition | Unknown |

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). (750-word limit)

The Behavioral Risk Factor Surveillance System (BRFSS) is the nation's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. The BRFSS questionnaire is designed by a working group of BRFSS state coordinators and CDC staff[[1]](#footnote-0).

Currently, the questionnaire using structured interviews to estimated nutritional habits which are subjects to recall bias and under-reporting. For example, Section 12 measures fruit and vegetable consumption using a sample of the following questions: (1) Not including juices, how often did you eat fruit? (2) How often did you drink 100% fruit juice such as apple or orange juice? (3) How often did you eat a green leafy or lettuce salad, with or without other vegetables?

Pushcart is a novel way to objectively measure frequency of specific and general food consumption over a period of time to supplement other measures such as the BRFSS questionnaire. Puschart can specifically capture household consumption changes over time in response to behavioral and policy interventions and could be added to a wide range of programs (e.g., WIC) once the techniques are refined.

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. (750-word limit)

Pushcart obtains grocery receipt data in an automated manner using email receipts and an authenticated Gmail API (Application Programming Interface) user. The grocery data provides information on product identifiers and quantity purchased for each product. Nutritional data is automatically retrieved from Nutritionix when an item is created and outputted to the API. Names of food items (and other identifying data about them) are hashed before they are saved, for internal privacy. A basic visual interface for purchases and items on the dashboard. The aggregate computed data and statistics are made available to third-party applications. Any data that is collected by our software, passively or actively, is encrypted and uploaded to a secure server and/or a security-compliant cloud based service. Extracted features contain no personally indefinable information (PII) and subjects will be given a random, meaningless code to identify them over time. Only the study investigators and the technical administrative staff will have access to the code key. The server administrator will apply the latest operating system and security updates. All data transfers will occur over HTTP, secured by an SSL certificate or SSH certificate. Login accounts will be established with only a username and password. No one except the end user will know their password, to prevent against automated attacks where malicious software attempts to guess passwords by trying many login attempts. Passwords for login accounts must be a minimum of 8 characters long. The software will run on secure cloud computing services. The system administration will meet the technical specifications of 1) limited logon access and 2) secure network connections through firewalls.

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. (500-word limit)

Empirical studies suggest that objectively documented household food purchases yield an unbiased approximation of overall diet quality among individuals when compared to 24-h diet recalls.[[2]](#footnote-1)[[3]](#footnote-2) This includes estimating percentage energy from saturated and total fat, carbohydrate, and protein, and sugar. For example, researchers documented a mean difference of ∼1% in the estimated percentage of energy from saturated and total fat across the two methods.

We will obtain time-stamped household-level metrics for each grocery purchase. We will convert those to daily metrics at the household and individual-level by calibrating it information from the National Health and Nutrition Examination Survey, which is publically available and provides detailed self-reported dietary intake information.

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? (750-word limit)

The supermarket database encompassed 65% of total sales expenditure (dollars spent on products in the database compared with all available food products) and 78% of sales volume (quantities purchased of products in the database compared with all available products), and thus covered most foods and beverages typically purchased.

Pushcart data is more representative for adults aged under 50. Across this population sub-group 15% already use virtual grocery shopping and 65% of them are willing to use it in the future[[4]](#footnote-3). Among adults aged over 50 the prevalence of online grocery shopping is substantially lower. We can estimate the representativeness of a specific household be measuring the frequency of online grocery shopping compared to brick and mortar shopping.

Pushcart nutrition data is representative of a household and therefore usually difficult to amenable by age and gender. We can measure education, geography and other social economic metrics by using supplementary survey techniques.

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? (500-word limit)

Objective measures of nutrition and food consumption of US population will add significant value to the validity of current public surveillance methods in this area. It is well known that traditional public surveillance methods of nutrition are notoriously hard because of high cost, recall bias and under-reporting. Electronic grocery data have several advantages over traditional methods of dietary assessment, particularly in placing little or no burden on participants, and in being an efficient and objective measure, which is particularly advantageous for groups with poor memory and/or low levels of literacy.

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. (500-word limit)

No. The proposed project will only provide data about nutrition and food consumption.

1. Please describe a 3.5-month plan to develop a working prototype during the second phase of this challenge. This should include:
2. Details on how you will gain access to and link data from the source(s) you’ve identified.
3. Approaches/strategies that will be taken to ensure privacy/confidentiality of data before and after linkage.
4. Your approach to comparing results from your prototype to that generated from existing public health surveillance programs
5. A description of the format your prototype will take (e.g., visualization, online data tool, etc.)
6. Costs you expect to incur during this prototyping phase

(1500-word limit)

Data collection:

We will recruit up to 300 households (~1000 people) residing in the US for the study; we will select our population in an area that has high penetration of online food shopping for mid to lower income groups. Participants will authorize Pushcart to access data about online receipts from the relevant email accounts. Pushcart will analyze retrospective data and provide longitudinal summary of food consumption and nutrition in these households for a period of 12 months. In addition, Pushcart will monitor ongoing purchases for a period of 30 days.

Privacy:

Any data that is collected by our software, passively or actively, is encrypted and uploaded to a secure server and/or a security-compliant cloud based service. Extracted features contain no personally indefinable information (PII) and subjects will be given a random, meaningless code to identify them over time. Only the study investigators and the technical administrative staff will have access to the code key. The server administrator will apply the latest operating system and security updates. All data transfers will occur over HTTP, secured by an SSL certificate or SSH certificate. Login accounts will be established with only a username and password. No one except the end user will know their password, to prevent against automated attacks where malicious software attempts to guess passwords by trying many login attempts. Passwords for login accounts must be a minimum of 8 characters long. The software will run on secure cloud computing services. The system administration will meet the technical specifications of 1) limited logon access and 2) secure network connections through firewalls.

Comparing to existing programs:

We will compare summary results obtained from Pushcart to The Behavioral Risk Factor Surveillance System (BRFSS) questionnaire. Specifically, we will focus on the two nutrition relevant sections from this survey (11 – Alcohol consumption; 12 – Fruits and vegetables).

Prototype:

ThePushcart prototype application gives individuals the ability to track their nutrition effortlessly, to receive insights from the application on their phone or computer, and to share the information with their dietician. Users receive insight through intuitive, clear “visual receipts” (see figure) designed using MyPlate categories and the Traffic Light Diet. Dietitians who use Pushcart gain the ability to securely and privately know precisely what groceries their clients buy, review grocery receipts and each item’s nutritional content, rapidly tag items as high / medium / lownutritional value, monitor changes to shopping habits over time, and communicate with clients online easily to provide tips and healthy swap suggestions, while monitoring which suggestions are successful and which are not. We designed the prototype application to be modular and reusable so that it can be remixed and repurposed for new services and platform architectures, from research to commercial products. Everything down to the front-end UI is Open Source.

Fig. Prototype Visual receipt

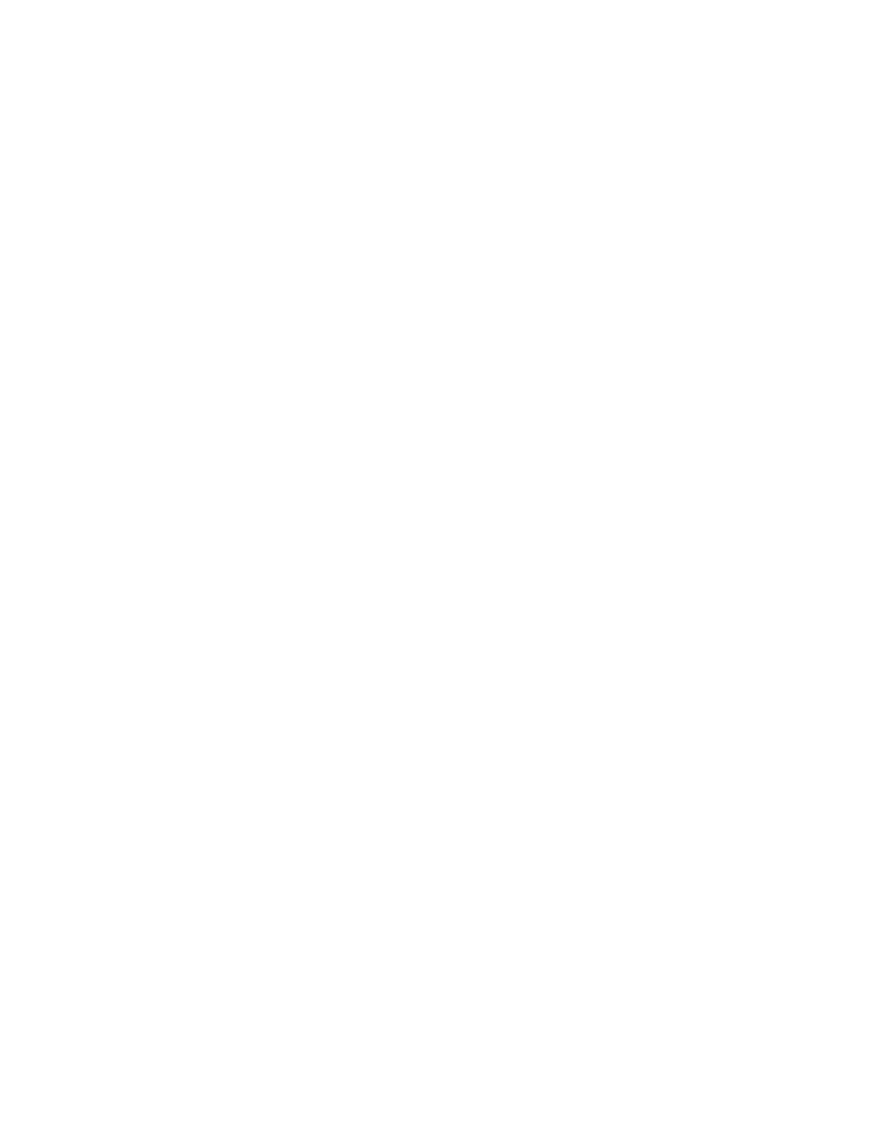
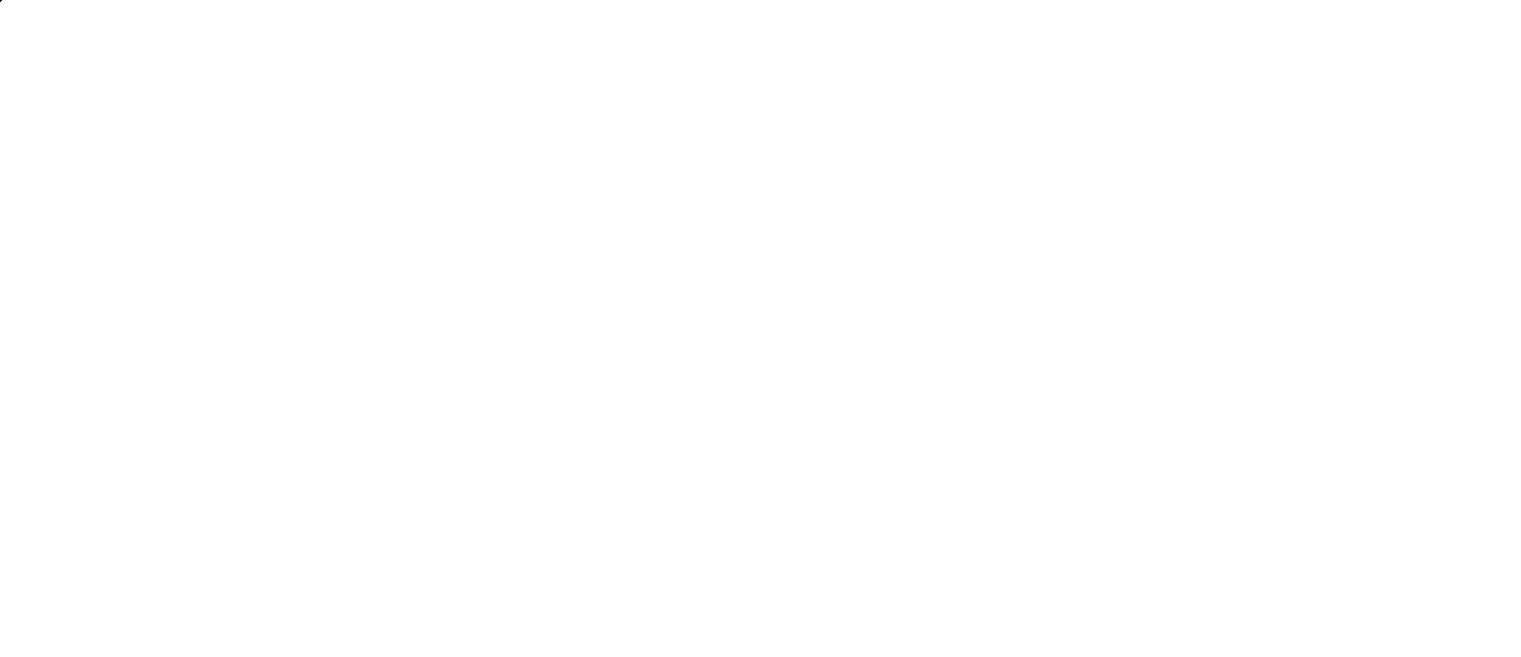


Fig. Prototyped dashboard to review shopping habits over time



Cost:

In the prototyping phase, most of the cost will be attributed to recruiting households to participate in the user study of Pushcart. In addition, minimal cost for storing the data and using Nutritionix API.

1. Significance and Relevance Summary

|  |
| --- |
| In 200 words or less, provide a brief summary of your project using language that is easily understood by the general public. Note: this description will be shared with a broad audience and should not include any information you would not want shared widely.  Pushcart is an effortless and compatible passive data collection tool that provides objective, high frequency, household-level data on food consumption and nutrition. Pushcart can serve as a proxy for variety of nutritional metrics on the household level including: total calories consumed and division to different food groups (e.g. fruits, vegetables, sugar) continuously. Pushcart provides a metric of nutritional habits by analyzing online shopping email receipts from grocery stores (e.g. instacart) and delivery services (e.g. seamless). Pushcart can provide an objective and effortless measure of nutrition and food consumption that could give insight about trends and improve existing public surveillance methods. |

Public reporting burden of this collection of information is estimated to average 60 minutes per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. An agency may not conduct or sponsor, and a person is not required to respond to a collection of information unless it displays a currently valid OMB control number. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to CDC/ATSDR Reports Clearance Officer; 1600 Clifton Road NE, MS D-74, Atlanta, Georgia 30333; ATTN: PRA (0990-0390).

1. . https://www.cdc.gov/brfss/questionnaires/index.htm [↑](#footnote-ref-0)
2. . Eyles, Helen, Yannan Jiang, and Cliona Ni Mhurchu. "Use of household supermarket sales data to estimate nutrient intakes: a comparison with repeat 24-hour dietary recalls." Journal of the American Dietetic Association 110.1 (2010): 106-110. [↑](#footnote-ref-1)
3. . Appelhans, Bradley M., et al. "To what extent do food purchases reflect shoppers’ diet quality and nutrient intake?." International Journal of Behavioral Nutrition and Physical Activity 14.1 (2017): 46. [↑](#footnote-ref-2)
4. . Nielsen Report on “The Future of Grocery”, 2015 [↑](#footnote-ref-3)