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

Phase 1 Submission Template

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. 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

Keeping Pace

Team Lead

City/Province

The Healthy Behavior Data Challenge

Phase 1 Submission Template

Rumi Chunara	New York
E-mail	Phone Number
rumi.chunara@nyu.edu	646-997-3312
Subject-matter/domain expertise	
Public health informatics	

Team Member #1	E-mail	Subject-matter/domain expertise
Ben Wellington	bwellington@landscapemetrics.com	Visualization of data and graphic design
Team Member #2	E-mail	Subject-matter/domain expertise
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

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

Organization Name	Website	Type of Organization
Chunara Lab	http://chunaralab.org	Research Group at NYU

3) How did you find out about this challenge?

Colleagues (word of mouth)

4) Submission Overview

Project Title

Keeping Pace: Tracking Physical Activity Together

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?

Our project focus on measurement of physical activity; one of the most important risk factors for obesity today [1]. In 2010 physical inactivity was identified as the fourth leading risk factor of non-communicable diseases and accounted for over 3 million preventable deaths worldwide [2]. Physical activity has been shown to play an integral role in the etiology and prevention of obesity as well as numerous other chronic conditions, such as cancer, coronary heart disease, diabetes, hypertension, depression and osteoporosis, which together are responsible for immense mortality, illness and expense in the United States today [1]. Thus, better understanding of physical activity is at the cornerstone of health and can have wide-ranging health, economic, and social consequences.

Simultaneously, novel Internet and mobile-connected data sources such as wearables and personal fitness trackers capture where people go and when in a very granular manner. The data are directly captured, which evades recall and information biases associated with traditional survey measures, at low cost [5]. However, the opt-in and proprietary nature of personal trackers present challenges that require appropriate thought and design into how best the data can be used [6]. Specifically, the variety and constantly changing landscape of wearable tools, which each may only gather specific types of information or only reach certain population groups, alongside a lack of linked details that are generally garnered from structured surveys that are essential for public health surveillance.

The Keeping Pace platform addresses these challenges by aggregating physical activity data from personal activity trackers alongside functionality for linking important information to the data. Keeping Pace includes an intake survey and follow-up surveys capture demographic information (essential to assess representativeness of the data) as well as allow users to label aspects of activity (e.g. enjoyment during activity) and push questions to users strategically (e.g. ask if activity was for transportation or recreation).

Data is ingested from multiple types of wearable devices or smartphone apps that users can choose to install (e.g. Moves) or is pre-installed on their phone (e.g. iPhone HealthKit). Crucially, the Keeping Pace platform also can be expanded as new trackers are developed. Current functional sources are: Fitbit trackers (all types), Run Keeper app, iPhone Healthkit, Moves app. Data are accessed securely via Application Programming Interfaces. Each tool tracks aspects of physical activity, such as number of steps and calories burned. In some cases the data is also linked to location, some devices infer the type of activity (or allow users to label it).

In summary, Keeping Pace will improve on current physical activity surveillance or simple aggregation
of data via:

1) High resolution (in time: minute, hour, and location: meter) direct traces of activity (avoiding under/over reporting and recall biases associated with self-reported survey measures)
2) Agile assessment representativeness in real-time enabling prospectively sample to include different groups, or based on upcoming important events or times (e.g. street closing)
3) An inviting interface enabling further communication with participants
4) Potentially enabling reach groups that are difficult to reach through active surveillance

[1] Warburton, D.E., C.W. Nicol, and S.S. Bredin, *Health benefits of physical activity: the evidence.* Canadian medical association journal, 2006. 174(6): p. 801-809
[2] Bull, Fiona C., and Adrian E. Bauman. *Physical inactivity: the "Cinderella" risk factor for noncommunicable disease prevention*. Journal of health communication, 2011. 16.sup2 p. 13-26.
[3] Physical Activity Guidelines Advisory Committee, *Physical activity guidelines for Americans.* Washington, DC: US Department of Health and Human Services, 2008: p. 15-34.
[4] Chunara, R. Wisk, L.E. and Weitzman, E.R. *Denominator issues for personally generated data in population health monitoring.* American journal of preventative medicine, 2017 52(4), p. 549-553.
[5] Sallis, J.F. and Saelens, B.E., 2000. Assessment of physical activity by self-report: status, limitations, and future directions. Research quarterly for exercise and sport, 2000 71(sup2), p. 1-14.

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

All indicators that can be accrued in Keeping Pace based on current data sources are listed below. It should be noted that as the number and types of data sources increase, the types of indicators can also increase based on what new sources can measure.

A) Physical Activity

- Amount of MVPA[1] 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

B) Sedentary Behavior[2]

- Amount of time per day spent sedentary, excluding sleep time

[1] Moderate-to-vigorous physical activity (MVPA) is any activity with an energy expenditure >3 metabolic equivalents

[2] Sedentary behavior is any waking activity characterized by an energy expenditure ≤ 1.5 metabolic equivalents and a sitting or reclining posture

6) 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	Fitbit	Wearables (Original as well as Zip, One, Flex 2, Alta, Alta HR, Charge 2, Blaze and Surge)	Accessed via Fitbit API (already exists)	No cost (users who have purchased Fitbits can opt in to provide data)	Data is collected in real-time (current), once a user has consented, it is updated every time the user syncs their data from Fitbit device with the Fitbit app	Physical activity and sedentary behavior	Keeping Pace, Fitabase (company)
2	Moves	Smartphone app	Accessed via Moves API (already exists)	No cost to the app	Data is collected in real-time (current), once a user has consented, it is updates every time KP updates	Physical activity and sedentary behavior	Keeping Pace
3	Run Keeper	Smartphone app	Accessed via RunKeeper API (already exists)	No cost to the app	Data is collected in real-time (current), once a user has consented, it is updates every time KP updates	Physical activity	Keeping Pace
4	Apple Health Kit	Smartphone app (built-in all iPhones/Apple watches)	Accessed via HealthKit Data Download/Export (software exists and implemented in KP)	No added cost (users who have purchased iPhone devices and haven't turned off this application can opt in to provide data)	Data is collected in real-time (current), once a user has consented, it is updates every time KP updates	Physical activity and sedentary behavior	Keeping Pace, Apple
5	Planko (app designed by Chunaralab team)	Smartphone app that tracks steps connected to location	Accessed via Planko (already exists)	No cost to the app	Data is collected in real-time (current), once a user has consented, it is updates every time KP updates	Physical activity and sedentary behavior	Keeping Pace

7) 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)

There are several ways the data in KP will augment current utilized data, each described below.

Frequency and Recency of Data Currently, in order to assess physical activity there are survey-based measures such as the Pedestrian and Transport Survey (New York City) [1] the Global Physical Activity Questionnaire (WHO) [2], or specific questions on the National Health Interview Survey (NHIS) or Behavioral Risk Factor Surveillance System [3,4]. Survey data is collected infrequently and suffers from lack of recency (e.g. latest data in the WHO Global Health Observatory data repository is from 2010, the PATS survey was only administered in 2010 and 2011 and the NHIS has been deployed annually or subannually). On the other hand, data aggregated from Keeping Pace:

- high resolution (in time: minute, hour, and location: meter) compared to survey data
- enables real-time monitoring (data is collected in real-time and updated in Keeping Pace whenever the data syncs)
- agile, can assess representativeness in real-time which then can be used to prospectively sample to include different groups, or sample based upcoming important events or times (e.g. closing streets) or other interventions
- high frequency means seasonal effects can be examined (e.g. day of week, or unique times like holidays)

Direct Measures Particularly relevant for measures considered here, studies have indicated that answers from self-completion questionnaires are open to misinterpretation about how well people are aware of their activity behaviors. There is also a strong issue of perceptions from survey-based measures; individuals could incorrectly perceive they have made healthy modifications, when in fact they have not. Especially in regards to physical activity, there has been evidence of a relationship between bias and intensity of activity [5,6]. Finally, while surveys robustly capture specific information about physical activity episodes and frequency, they do not link context (location) and time to behaviors. Thus, Keeping Pace augments on existing measures by:

- direct measurement of activity instead of self-report
- high-resolution in terms of location enabling qualitative and quantitative assessment of specific built-environment relationships to activity (e.g. specific blocks where there is increased activity or data informing qualitative studies to investigate why certain parks/sidewalks are used less than others)
- Measures of activity linked to time and place to provide context and inform further qualitative studies to examine behaviors and their modifiers

Flexibility of Data Collection Formats While existing denominator-based surveillance systems provide population-representative structured data to inform policy and guide clinical and research objectives, survey timeframes and sampling frames do not flexible follow-up or feedback of health alerts and information to subjects. Specifically Keeping Pace offers flexibility in regards to:

- enabling follow-up questionnaires to ask questions in an agile manner

- through the digital mechanism can communicate with participants in a non-intrusive and fun way

Population Reach/Scale Traditional surveillance modes including telephone, paper and in-person interviewing, is becoming increasingly costly in terms of time, cost and labor. Thus, reaching specific subgroups, small geographies or even large amounts of people is challenging. Also, some surveys captured limited data via deployment of GPS-trackers, however these are often gathered for short periods in time, and from populations in specific areas. Harnessing personal tracking tools thus potentially enables:

- reaching groups that are difficult to reach through active surveillance
- scaling to reach the possible millions who use personal activity trackers (which doesn't include those who have an iPhone which also tracks activity intrinsically) [7]
- examining activity at community levels and places that do not have the infrastructure for detailed surveys

Going Beyond Simple Data Aggregation. While novel data from personal tracking tools provide opportunity for real-time and precise located data, there are still added features about the data that are important for use in public health research and surveillance. Therefore, Keeping Pace augments data aggregated via:

- linked demographic data with all of the aggregated data
- ability to query users with questions about data (e.g. for periodic, repeated routes, ask if this is a commute to work, or ask what the purpose of a route was, enjoyment during a particular activity or perception of safety)
- Ability to get in touch with users later for other targeted questions

[1] http://www1.nyc.gov/site/doh/data/data-sets/physical-activity-and-transit-survey.page
[2] Armstrong, T. and Bull, F., 2006. Development of the world health organization global physical activity questionnaire (GPAQ). *Journal of Public Health*, *14*(2), pp.66-70.
[3] Parsons, V.L., Moriarity, C.L., Jonas, K., Moore, T.F., Davis, K.E. and Tompkins, L., 2014. Design and estimation for the national health interview survey, 2006-2015.
[4] Bland, S.D., Bolen, J.C., Holtzman, D., Powell-Griner, E. and Rhodes, L., 2000. State-specific prevalence of selected health behaviors, by race and ethnicity; behavioral risk factor surveillance system, 1997.
[5] Baranowski, T., *Validity and reliability of self report measures of physical activity: an information-processing perspective.* Research Quarterly for Exercise and Sport, 1988. 59(4): p. 314-327.
[6] Sallis, J.F. and B.E. Saelens, *Assessment of physical activity by self-report: status, limitations, and future directions.* Research quarterly for exercise and sport, 2000. 71(sup2): p. 1-14.
[7] Forrester Data Report: Consumer Wearables Forecast, 2017 To 2022 (US).

8) 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)

Process: Our platform (Keeping pace) consists of three parts:

1) Input/profile page (presentation slide 3): This initial part of the platform contains a simple intake survey. Standard demographic information (age, gender, race/ethnicity, home zip code) will be collected. Next, the user is taken to a page where they can select which data sources to connect to the Keeping Pace platform. Selecting a data source is operationalized through the next part of the platform.
2) Data connection through Application Programming Interfaces: We will import data through an OAuth transaction with the API (this online technology is used to securely authenticate and share permissions between websites). Each device will connect to receive a summary of data on a daily basis. Depending on the type of device, this could consist of a list of a user's activities and activity log entries for a given day in the format requested using a specific unit system.
3) Data visualization/presentation/communication front-end (presentation slides 6-9): In an intuitive user interface, metrics such as MVPA and others that can be statistically aligned with existing survey data are communicated across multiple scales. For example, at the sub-city level (via neighborhood or census tract), state-by-state or precise-locations that can be used to precisely learn about built-environment links to physical activity. Data for this view would be anonymized (please note metrics on presentation are example/place holders and are not real data).

Feasibility:

As our platform has been built and ingested participant data already (144 participants), feasibility of the data collection process has been tested (users have signed up, consented, and data has been synced).

Privacy, security and confidentiality:

Security and consent: Although the data shared by participants, we have gone through the process of creating an IRB approved consent form which explains: the potential benefits, risks, what the process will involve, contact information with any questions. Participants security is also ensured by always having the opportunity to withdraw. Also by sharing a view of the aggregate data with participants we will increase engagement which is essential for a participatory approach. Children under the age of 18 will not be permitted to sign up (this is a commonly used exclusion criteria for Internet-based studies). Future work can specifically focus on children by explicitly including parental consent and accounting for other attributes specific to the inclusion of children.

Data Security: Security of data stored on the platform is of utmost importance, and we will take the following steps:

- data will be stored in a secure online server.
- only key personnel will have access to the data.

Keeping Pace storage is and will be expanded to incorporate data best practices such as:

- no re-identification: privacy is a moving target, it is agreed that no specific efforts to re-identify individuals will be made from the data they share in an ongoing manner
- acknowledgement and citation where possible: an effort to acknowledge the participants contributing data will be made whenever possible (e.g. can be done anonymously, in presentations, reports etc.).
- knowledge sharing: standards for data measures will be used whereever possible to ensure sustainability and knowledge translation
- data use: any use of the data must conform to all applicable laws and regulations in a given jurisdiction.

As well we will employ technical storage practices including using TLS/SSL (transport layer security and secure sockets layer), hashed passwords, limited admin access and database backups.

Privacy/confidentiality: The high resolution of data (mapped activity routes) from personal trackers should be carefully managed as location data is sensitive. Individual-level traces will not be made available except to key personnel of the surveillance tool.

9) 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)

On the individual level, there are concerns and ongoing research regarding quality and variability of measures from personal activity tracking tools [1], and due to proprietary nature, tools are not validated through a standardized approach. However, research at the aggregate scale has shown these data to directly correspond to self-report measures, and instill confidence in their use. Informed by the most current research on physical activity tracker data, here we describe how statistically sound metrics will be developed from Keeping Pace data.

- MVPA time per day: Smartphone accelerometer-based measure of average daily steps have shown to correlate with WHO guideline measures for moderate to vigorous physical activity based on self-report [2]. This is corroborated by the finding that the difference in average steps per day between females and males is strongly correlated to the difference in the fraction of each gender who report being sufficiently active according to the WHO. Also, a strong relationship between activity inequality and obesity across countries suggests that findings are robust to differences in wealth. Thus total steps per day will be used as the proxy for MVPA.
- MVPA time/day in bouts of 10 minutes or more: Using above, this can be directly computed based on timing of the continuous data and an appropriate threshold for gap times (time between first and last step / 10 minutes), where first step is the first step after an appropriate threshold gap
- MVPA time at work, at home and/or in transit: Computed using MVPA time, intake survey and linked data labels indicating place of work, home, and transit routes

- Times during the day when MVPA is high, MVPA frequency: Over all participants (stratified by demographics), we will identify times of highest MVPA and compute MVPA time per day
- Daily steps, distance, calories burned, number/flights of stairs climbed and average and peak heart rate: Accessed directly from each source with step counts, distance by activity type, calories burned, flights climbed (HealthKit) or heart rate available (Fitbit)
- Activity type (aerobic, strength, etc.) or level (low, moderate, high): Accessed directly from each source with activity type available (type), from heart rate, activity measures and also inferred based on MVPA time (level)
- Time spent in different MVPA domains (home/occupational, travel and recreational) or locations: Computed using MVPA time, intake survey and linked data labels from users indicating place of work, home, and transit and GIS data of transit routes and sidewalk/bike lanes
- Perception of safety or enjoyment while active: Computed using MVPA time and linked labels from users indicating place of work, home, and transit routes or enjoyment level during activity
- Hours/week in sports, fitness or recreational physical activities (or time spent sedentary, excluding sleep time): Computed via activity data pre-categorized into activity type, and using post-hoc user tags for type of activity (e.g. transport, recreation, etc.) or for sedentary, via time with no activity (excluding sleep identified post-hoc, or inferred based on largest block of time with no activity in 24 hours).

[1] Dannecker, Kathryn L., et al. "Accuracy of Fitbit activity monitor to predict energy expenditure with and without classification of activities." *Medicine & Science in Sports & Exercise* 43.5 (2011): 62.
[2] Althoff, Tim et al. "Large-scale physical activity data reveal worldwide activity inequality". *Nature* (2017) doi:10.1038

10) 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)

Representativeness is an important issue which must be considered for all data sets. Our group has written extensively on this topic [1]. For example, the New York City Physical Activity and Transit Survey, which is geographically well-weighted, only includes adults, so it does not represent anyone under the age of 18. Collection of the personally-generated data via our platform in concert with an intake survey links the data with demographic information (gender, age, race/ethnicity, home location) in order to assess who the data is generated by and meaningfully stratify or extrapolate the data to larger groups. In sum, all of the data will have linked demographics thus the data set will be possible to disaggregate by age, gender, education, and race/ethnicity. Further demographic characteristics can be solicited via post-data collection surveys.

Understanding the precise demographics of the data will enable prospective focused sampling for underrepresented groups. Based on initial data collection, as well as statistics about use of physical activity wearables, adolescents and young adults are among the groups most heavily engaged. Our experience in recruiting for other Internet-based public health initiatives (www.goviralstudy.com) has shown that targeted enrollment can also increase representation from other groups.

Yet, there may be fundamental limits due to use by some groups (e.g. the very young or old) who simply do not use these types of personal tools. For example, the very young or very old (similar to the PATS survey which does not include those under 18 years old, the WHO GPAQ via the Global School-based student health survey which doesn't include anyone under 13), or those in rural populations. Current research has shown that use of these tools is very feasible and acceptable in these populations [2], thus our approach which incorporates multiple tools and is designed to capture allows for underrepresented groups.

[1] Chunara, R. Wisk, L.E. and Weitzman, E.R. Denominator issues for personally generated data in population health monitoring. *American journal of preventative medicine*, 2017 52(4), p. 549-553.
[2] Batsis, John A., et al. "Use of a wearable activity device in rural older obese adults: A pilot study." *Gerontology and geriatric medicine* 2 (2016): 2333721416678076.

11) 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)

International, national and local health surveillance programs such as the WHO Global Physical Activity Questionnaire, CDC National Health Interview Survey or New York City Physical Activity and Transport survey measure physical activity levels and related metrics. These denominator-based surveillance systems provide population-representative structured data to inform policy and guide clinical and research objectives. Acquiring data through these means is becoming costlier, in terms of time, labor and monetary costs. Further, survey timeframes and sampling frames do not support real-time monitoring, robust subgroup investigation, flexible follow-up or feedback of health questions or alerts to participants. Simultaneously, sizeable populations worldwide generate vast quantities of digital data from Internet and mobile connected tools. Mining such data delivers low-cost, high-resolution views into public health phenomena to complement traditional systems.

While the data offers important and relevant information, due to the observational, proprietary nature of the data there are challenges to aggregating the data in a clear and consented manner by participants. To-date as far as we are aware, no public health organizations use the data in an aggregate manner. Our team's experience in engaging participants (through multiple participatory public health platforms such as www.goviralstudy.com, www.flunearyou.org) as well as communication and visualization of scientific data (http://www.landscapemetrics.com/portfolio/) are essential towards acquiring and using the information effectively.

While the data acquired from personal health trackers may not be representative of the general population, the ability to stratify by demographic information based on the intake survey, as well as any information collected retrospectively through surveys on Keeping Pace should provide the ability to know *whom* the data represents, and with enough data make balanced comparisons to existing survey measures.

Finally, the data gathered through Keeping Pace can go beyond existing public health survey-based measures. Survey measures only provide a static and select view of any relationships. For example the 2017 BRFSS physical activity questions ask about number of physical activity episodes during the past week/month, and average time. Thus the survey doesn't capture measures such as heart rate, calories burned, or uncover nuances such as how measures vary by time. Temporal relationships are important in order to understand if there are specific periods at which environmental parameters may be better utilized (for example if activity in a certain area is highest weekends versus weekdays). Thus the direct and multiple measures captured through Keeping Pace can offer added quantitative information for public health surveillance.

12) 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)

The data generated by the Keeping Pace project will be relevant to multiple other parts of public health surveillance.

Infectious Diseases: To accurately assess the daily movements of people, to-date infectious disease modelers have used high-resolution data sources such as Call Data Records (CDR) to approximate human mobility (via time and location measures). Thus data from Keeping Pace can provide mobility measures (e.g. for parameterizing transmission models, especially at high resolution) [1]. Downstream processing and analysis of the data would be required to:

- assess overall movement between locations over different time periods
- impute location frequency where data is sparse (we have developed algorithms to do this with high accuracy [2])
- stratify movement by demographic group

Chronic diseases: Physical activity has been shown to play an integral role in the etiology and prevention of obesity as well as numerous other chronic conditions, such as cancer, coronary heart disease, diabetes, hypertension, depression and osteoporosis [3]. Comprehensively measuring physical activity behavior in precise spatio-temporal context offers a mechanistic view of how we interact with our environment that can directly inform prevention in complement to important work on elucidating obesogenic environment features (e.g., price and availability of food).

Further, models of disease that specify behaviors and enable their close monitoring are likely to better inform intervention and prevention efforts by improving effectiveness and optimizing investment. A core such behavior is physical inactivity; it is a modifiable behavior embedded in and influenced by a network of social-environmental factors (e.g., social exercise norms, transportation systems, etc.). However until now, it has been limited how well we can specify detailed information about behaviors in models of disease risk. The measures generated through the Keeping Pace platform, being linked to precise times, places and groups can be linked to these external factors improving modeling efforts, such as those assessing the efficacy of proposed interventions or interactions between the social environment and behavior change. The measures can be used as computed in their various forms, there is not any specific down-stream processing needed, unless there are niche applications. Thus the data gathered through the Keeping Pace system can be used in detailed models of how policies can shape behaviors and the simulate the effect of different public health programs.

[1] Wesolowski, Amy, Nathan Eagle, Andrew J. Tatem, David L. Smith, Abdisalan M. Noor, Robert W. Snow, and Caroline O. Buckee. "Quantifying the impact of human mobility on malaria." *Science* 338, no. 6104 (2012): 267-270.

[2] Rehman, N.A., Chunara, R. Filling in the Timeline: Predicting Individual Level Mobility Patterns from Sparse Digital Traces. *International conference on Ubiquitous computing* (under review).

[3] Warburton, D.E., Nicol, C.W., and Bredin, S.S. 2006. "Health Benefits of Physical Activity: The Evidence." *Canadian medical association journal* 174 (6):801-809.

13) Please describe a 3.5-month plan to develop a working prototype during the second phase of this challenge. This should include:

1) Details on how you will gain access to and link data from the source(s) you've identified.
2) Approaches/strategies that will be taken to ensure privacy/confidentiality of data before and after linkage.
3) Your approach to comparing results from your prototype to that generated from existing public health surveillance programs
4) A description of the format your prototype will take (e.g., visualization, online data tool, etc.)
5) Costs you expect to incur during this prototyping phase

(1500-word limit)

1) Details on gaining access: We will use Application Programming Interfaces (APIs) to access data from the included tools. We will import data through an OAuth transaction flow with the API (this online technology allows people to give your application/website permission to access data on their behalf). In order to execute this, the user has to select which tool they would like to give access to Keeping Pace. The platform then automatically requests "permission" from the selected tool's API on behalf of the user. The API then provides a token (request code) and secret to Keeping Pace. Finally the user is directed to the tool's own website with the token. There, they enter in their own password (so Keeping Pace never sees the password), along with the secret. Then, they are redirected back to Keeping Pace with a permission token. Now, Keeping Pace has permission to access the user's data from the select tool. All of these steps are performed seamlessly/automatically, so the users only has to simply click on a button to select the tool, and then enter their password in on the tool website.

As we already have implemented this procedure, we will simply enable access and ensure the front end of the intake area matches the experience. This approach will also enable us to add access to new personal tools as they arise in the future.

2) Approaches/strategies that will be taken to ensure privacy/confidentiality of data before and after linkage:

- OAuth authorization described above is used to securely authenticate and share permissions between websites).
- When users create a Keeping Pace account, their password is hashed and stored in the database. At no point is the plain-text (unencrypted) password ever written to the hard drive. When the user attempts to login, the hash of the password they entered is checked against the hash of their real password (retrieved from the database). If the hashes match, the user is granted access. If not, the user is told they entered invalid login credentials.

- Transport Layer Security/Secure Sockets Layer (TLS/SSL) protocols are used. These are state of the art protocols which improves security by providing a digital certificate that authenticates storage systems and allows encrypted data to pass between the system and a browser. These protocols are built into all major browsers. Therefore, installing a digital certificate on the storage system enables TLS/SSL capabilities between system and browser.
- Data will be backed up automatically every night in order to prevent and avoid data loss or tampering.
- Users will have the ongoing ability to share parts of their data or intake survey, or downstream responses, to ensure privacy/confidentiality (i.e. they can choose questions not to answer or parts of their data not to share).

3) Comparing results: Comparison of results between our data and existing survey measures will be performed at available scales of survey data. First, we will compare data within New York City, as the NYC Pedestrian and Transport (PAT) survey provides detailed data at an appropriate resolution [1]. As well, we anticipate that it will be feasible to have enough data in one location (New York City) for the comparison initially. The New York City Department of Mental Health and Hygiene (NYC DOMH) has generated multiple detailed physical activity surveys. We anticipate that early data will be focused in this one urban area (due to population density, higher use of personal trackers and existing preliminary data gathered by our study team which was primarily in New York City).

The survey data allows stratification via subgroups (age, gender and neighborhood) enabling linking social media data with data from these surveillance systems. We will compare based on measures including MVPA or “any physical activity in the recreation domain” by week and home location which are available from PAT and can be derived from KP data as described in section 9. Measures will be compared with data from the most recent secular survey using via balanced correlation analysis. Chunara has worked with these specific municipal data in previous work and is familiar with potential limitations inherent in the survey-responses such as self-report and recall bias [2, 3]. Demographic data will be used to select balanced populations from both the KP and survey data for fair comparison of outcomes as we have done in a recent study [4]. These correlations are not trivial and will provide insight into where data is over or under represented or other qualitative changes that result in differences (for example since the latest NYC data is from several years ago, there could be environmental or other shifts that result in population physical activity changes). Shifts in reports from other national survey data approximate 10% over the last 10 years [5], but shifts in activity data are not available. Thus, we will conduct a sensitivity analysis to assess how annual changes across DOMH panels could impact study outcomes.

Subsequent downstream comparisons can be performed state by state using CDC data from the NHIS or BRFSS surveys which also provides measures of the number of active minutes per week and MVPA and can be stratified by group.

Data regarding specific physical activity forms (e.g. running, biking) are more nuanced and can also be compared with enough data.

4) Prototype format: Keeping Pace consists of an intake data form, backend storage of data, and visualization display of data. Landscape metrics are experts in designing and building interactive data visualizations and data stories. They have worked closely with a variety of scientific partners

including Dr. Chunara's research group previously, to create engaging visualizations and tools that clearly and effectively communicate data, bolstering outreach and facilitating decision-making. They will ensure that the intake form and data visualization parts of Keeping Pace are designed an optimal, intuitive and aesthetically pleasing manner.

The user interacts with Keeping Pace by first signing up through an intake survey (presentation slide 3) which requires sex/gender, race/ethnicity, age, home location and any other basic demographics needed linked to all physical activity data. The user then proceeds to allow access to data from any tools of their choice (as detailed above). Once connected, their data will then prospectively sync in Keeping Pace at a given sync rate (currently every 24 hours). The user is also given opportunity to post-hoc tag any contributed physical activity data based on questions such as activity type, enjoyment level, perception of safety.

The output visualization will enable viewing of activity data measures at multiple spatial scales (e.g. sub-city or state-by-state, slides 6-9 in the presentation). Further, the data can be viewed via *data filters* to only see data for certain demographics (age groups, genders, race/ethnicity) or activity type. As well, there is opportunity for a precise view, to examine specific paths of activity where available. This can be done in conjunction with other environmental layers such as location of parks, sidewalks, commercial or recreation areas. These layered views are useful in public health surveillance to understand how people interact with the built environment and known features of the environment that should promote or discourage activity .

5) Major expected costs during the prototyping phase:

- Front-end application development costs: Landscape will work in tight conjunction with Dr. Chunara's group to ensure that the front-end is composed of all necessary components and will be intuitive for participants.
- Back-end application development costs (data analysis and aggregation)
    - o Storage, website hosting costs
    - o Staff time for data handling between new interface and existing system

Overall timeline:

- Platform design parameters, reconciling with function and existing system (3 weeks)
- Front-end design (3 weeks)
- Back-end security and application development (3 weeks)
- Overall implementation and testing (4 weeks)
- Write-up of user guide (2 weeks)

[1] http://www1.nyc.gov/site/doh/data/data-sets/physical-activity-and-transit-survey.page

[2] Chunara, R., Bouton, L., Ayers, J.W. and Brownstein, J.S., 2013. Assessing the online social environment for surveillance of obesity prevalence. *PloS one*, *8*(4), p.e61373.

[3] Bai, H., Chunara, R. and Varshney, L.R., 2015. Social capital deserts: obesity surveillance using a location-based social network. *Proceedings of the Data for Good Exchange (D4GX), New York.*

[4] Rehman, N., Liu, J. and Chunara, R., 2016. Using propensity score matching to understand the relationship between online health information sources and vaccination sentiment. In *Proceedings of Association for the Advancement of Artificial Intelligence Spring Symposia* (pp. 23-25).

[5] Chen, C.M., Hsiao-ye, Y., and Faden, V.B. 2013. Trends in Underage Drinking in the United States, 1991-2011. In *Surveillance Report #96*. Bethesda, MD: National Institute on Alcohol Abuse and Alcoholism

14) 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.

Keeping Pace is composed of three components that together enable the aggregation and use of personal health activity tracking tools used by millions of people. The platform addresses challenges of: privacy, representativeness of tool use, lack of context around data and the changing landscape and use of different types of proprietary tracking tools. First, an ethically-sanctioned consent form and ingested pilot data from participants ensures we can feasibly preserve privacy and security of participants in a manner acceptable to them. Second, going beyond simple data aggregation, the Keeping Pace platform couples automatic data ingestion with an intake survey and ability to push questions to participants at specific times or conditions, importantly linking activity measures to demographic information (to assess representativeness and shape prospective recruitment) and provide context about the movement data (e.g. assessing if it is for recreation or transportation). Realized activity measures and linked locations and times can be used to compute measures that are reliable and augment existing public health practice. Third, Keeping Pace has a front-end by which anonymous, aggregate data at multiple resolutions will be shared with public health practitioners, illustrating metrics that can be aligned robustly with existing public health measures.