Physical activity, sedentary behaviour, and sleep on Twitter: A labelled dataset for public health research

Zahra Shakeri Hossein Abad[1, 2], **Gregory P. Butler**[3], **Wendy Thompson**[2], **and Joon Lee**[1,2,4]

[1]Data Intelligence for Health Lab, Cumming School of Medicine, University of Calgary, Canada
[2]Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, Canada
[3]Centre for Surveillance and Applied Research, Public Health Agency of Canada, Canada
[4]Department of Cardiac Sciences, Cumming School of Medicine, University of Calgary, Canada
[*]corresponding author(s): Zahra Shakeri Hossein Abad (zshakeri@ucalgary.ca)

ABSTRACT

Advances in automated data processing, together with the unprecedented growth in user-generated social media (SM) content, have made public health surveillance (PHS) one of the long-lasting SM applications. However, the existing PHS systems feeding on SM data have not been widely deployed in national surveillance systems, which appears to stem from the lack of practitioners' trust in SM data. More robust datasets over which machine learning (ML) models can be trained/tested reliably is a significant step toward overcoming this hurdle. The health implications of physical activity, sedentary behaviour, and sleep (PASS) are widely studied through traditional data sources, which are often out-of-date, costly to collect, and thus limited in quantity and coverage. We present LPHEADA, a multicountry and fully Labelled digital Public HEAlth DAtaset of tweets originated in Australia/Canada/United Kingdom/United States between November 2018-June 2020. LPHEADA contains 366,405 labels for 122,135 PASS-related tweets and provides details about the place/time/demographics associated with each tweet. LPHEADA is publicly available and can be utilized to develop (un)supervised ML models for digital PASS surveillance.

Background & Summary

Almost two-thirds of the world's population now use the Internet, taking the global total to 4.57 billion (59%) by July 2020[1]. 87% of Internet users and 65% (3.96 billion) of the world's total eligible population (i.e. over the age of 13) now use social media. The combined time that these users spend on social media adds up to more than 1 million years every day[1], contributing to a large amount of user-generated data on different social media platforms. In 2020, Twitter alone reported 500 million tweets generated per day from 145 million daily active users. The low-cost data stream available on social media and other Internet-based sources serves to make research advances on Digital Public Health Surveillance (DPHS) more accessible for public health officials, clinicians, patients, and the general public. This help to disseminate insights into different aspects of public health and promote healthy lifestyles and health policies[2,3]. The open access to the public data about users and their opinions, the ease of use, and a large user base have made Twitter one of the most popular data sources for studying different aspects of public health[4,5], with Google Scholar indexing 1.32 million articles mentioning Twitter and public health. Moreover, more than 85% of Twitter users, with a wide breadth of demographic groups[4], also use Facebook, Instagram, and YouTube (this number for other platforms varies between 52% to 82%)[1], indicating that Twitter users reasonably represent active social media users in general.

Since 2011, Twitter has been the most popular form of social media used for public health communication[6,7]. A recent scoping review of 755 articles on DPHS shows that Twitter is the most studied of all platforms and most utilized platform to study communicable diseases, behavioural risk factors, mental health, drug utilization, and vaccine[7].

However, a number of limitations that mainly stem from the limitations associated with the data, are still the major obstacles towards the adoption of digital data for public health surveillance[4,7]. Given the main aims of any PHS system are to measure, monitor, and improve the overall health status of their target populations, the systematic incorporation of time, demographics (i.e. age, gender), and place data into the surveillance process is critical to the reliability and generalizability of this process[8,9]. However, nearly one-third (32%) of the digital public health surveillance studies published between 2005-2020 (with the majority of them related to behavioural risk factors surveillance) did not capture age, gender, or place information for their analyses[7]. Moreover, most studies on DPHS do not consider whether their findings are associated with the user's personal experience (self-reported or not), leading to content bias, incorrect results and potentially flawed interpretations[7].

Considering the location-dependent nature of health policies, along with the essential role of place data in assessing the representativeness of a PHS system, the impact of a PHS system can vary considerably with geographical location[10–13].

However, the number of DPHS studies that have stratified their results by more granular geographic region is small[7]. Due to a lack of annotated datasets for the development of automatic models, more than two-thirds (69%) of DPHS studies published before 2020 are limited by labour-intensive, manual, and abstract analysis methods (e.g. manual coding, qualitative analysis, and rule-based natural language processing), which makes these studies limited in terms of sample size, scope and generalizability[7].

Given that all of these challenges are data-oriented, an increase in both data quality and quantity enriched with concrete demographics and location information can help deal with all these challenges. Moreover, to facilitate the development and evaluation of robust machine learning models to address the limited scope of manual data analysis techniques, annotated datasets for various PHS aspects are required. However, only a handful of annotated datasets are publicly available for research on DPHS[14–21]. Jimeno-Yepes et al.,[15] provided an annotated dataset of 1,300 tweets related to disease symptoms and pharmacologic substances. The open dataset developed by Aphinyanaphongs et al.,[16] contains 13,146 labelled tweets resulting from hashtag filtering and covers a time span from January 2010 to January 2015. This dataset is developed for training binary classifiers to detect tweets that indicate e-cigarette use for smoking cessation. Crowdbreaks[18], an open health tracking platform, crowdsources the labelling of vaccine sentiment and COVID-19 related tweets to the public. While the dataset provided by this system, compared to other open DPHS datasets, is in a better position in terms of size, it lacks demographics and geospatial data, and each tweet is labelled by only one annotator (without any control over their labelling quality).

Given that in addition to physical inactivity, as the leading risk factor for non-communicable diseases and premature death[22], prolonged sedentary behaviour and inadequate sleep are also important risk factors for chronic disease[23], this work presents a multicountry and fully Labelled digital Public HEAlth DAtaset (LPHEADA) of tweets related to physical activity, sedentary behaviour, and sleep (PASS) that originated in Australia, Canada, the United Kingdom (UK), or the United States (US). We selected these countries as they have some of the highest proportions of social media users in the world (Australia: 71%, Canada: 66%, UK: 66%, US: 69%, and World: 51%)[1]. LPHEADA comprises 366,405 labels, labelled by 708 unique annotators on Amazon Mechanical Turk (AMT), for 122,135 unique tweets generated by 72,749 unique users between November 28th 2018 to June 19th 2020. AMT is a software service operated by Amazon that allows users (*aka* requesters) to crowdsource work–broken into micro-tasks called human intelligence tasks (HITs), to a large number of workers who are compensated for each HIT completed[24]. As LPHEADA was collected and labelled in collaboration with the Public Health Agency of Canada (PHAC) to develop PASS indicators for the Canadian population, around 81% (98,722) of the tweets included in our dataset were collected from Canada. Tweets from the US and UK make up 8% of the dataset each (US: 10,193, UK: 9,154), and Australian tweets make up the remaining 3% (4,067) of the dataset. Along with the labelled tweets, we provide detailed information about users' gender, age range, geospatial information, whether the tweet is self-reported or not, and whether it is posted by an organization. We evaluated the quality of the dataset and its labels using latent semantic analysis (LSA), linguistic analysis, machine learning models, and truth inference models. The dataset we provide in this paper can be utilized to develop (un)supervised machine learning models for digital PASS surveillance.

Methods

Collection and Preparation of the Dataset

We collected the data of this study from Twitter using the `Twitter livestream application programming interface (API)` between 28th November 2018 to 19th June 2020. The entire dataset (i.e. 1,902,980,841 tweets) was filtered for Canadian tweets potentially relevant to PASS. 103,911 tweets were selected from 22,729,110 collected Canadian tweets using keywords and regular expressions related to PASS categories. Each of these 103,911 tweets was labelled by three AMT workers, from which 98,722 tweets received three valid labels (i.e. multi/empty-labels are invalid and were rejected), with almost half of them related to physical activity. For the Canadian dataset, 610 unique workers participated in our data labelling tasks and completed 103,911 HITs, from which 5,189 HITs were removed as they did not receive three valid answers. The majority of these workers (87%, 530) completed less than 100 HITs each, among which 164 completed only one HIT each. Among the workers who completed more than 5,000 HITs, one worker completed 21,801 HITs, and three workers completed between 5,000 and 10,000 HITs.

In addition to the Canadian tweets, we filtered a random subset of the dataset for tweets that originated in the UK, USA, and Australia. This dataset spans the same data collection period as the Canadian dataset and contains 70,239 labels collected for 23,413 tweets (i.e. three labels per tweet). Adding the data from these countries will provide an important epidemiological diversity that can be utilized for implementing comparative studies and evaluating the generalizability of the PASS surveillance models trained on the Canadian dataset.

Labelling Process

We implemented a pipeline to create the crowdsourcing tasks, referred to as HITs by AMT, post them on AMT, collect the labels through a quality check process, approve/reject the HITs, and store the results. To minimize noisy and low-quality data, we added a qualification requirement to our tasks and granted the labelling access to workers who had demonstrated a high

degree of success in performing a wide range of HITs across AMT (i.e. master qualification). In addition to this, we added a simple qualification question to each HIT to detect spammers or irresponsible workers. Each HIT contained four questions, including the qualification question and was assigned to three workers. Through different iterations of data labelling, workers were paid from US$0.03 to US$0.05 after completing each HIT. We collected the labels for the 122,135 tweets used in this study through different iterations, from April 2019 to November 2020. We regularly checked the quality of submitted tasks to detect low-quality workers during each iteration and revoke their access to our tasks. Before the formal initiation of the process, we pilot tested the design, response time, and complexity of the HITs in two iterations and revised the workflow accordingly. To label the dataset provided in this paper, we utilized AMT as a crowdsourcing service and did not collect any personally identifiable information from the workers (participants) during the data labelling task. The experiments were carried out in accordance with relevant guidelines and the University of Calgary Conjoint Faculties Research Ethics Board's (CFREB) regulations. We implemented the entire workflow in `python` and used `Boto3 python SDK` to connect to and work with AMT.

Time Adjustment

Twitter API returns the date and time that a tweet is published in UTC. To adjust this timezone base on each tweet's location, we utilized the bounding box of coordinates, which enabled spatial mapping to tweets' respective city locations and used `timezonefinder` in `python`. Given that daytime, month, and weekday can be influential factors in twitting about each of the PASS categories, and to better utilize the datetime data (`%a %b %d %H:%M:%S %Y`), we extracted `a`: weekday, `b`: month, and `H:` hour fields and stored them as separate features. Figure 1a shows the temporal distribution of tweets for each of the PASS categories in the Canadian dataset. Moreover, the stacked area charts presented in Figures 1b-1d detail the frequency of tweets for each of PASS categories for the top ten Canadian cities.

Location Inference

The geospatial metadata provided by the Twitter API is derived from three main sources, including the (1) geotagged location, (2) profile location, and (3) mentioned location in the tweet text. Geotagged location can be defined by exact location (i.e. device location) at the time of tweeting, by assigned twitter place (i.e. at the neighbourhood level), or both. While the exact location field provides the highest level of precision, it is not a default setting, and only a small portion of users share their exact latitude and longitude (e.g. only 1-2% of tweets are geo-tagged[25]). Thus, to infer the location of each tweet in LPHEADA, we proposed and developed a five-step process that utilizes tweet-neighbourhood location (i.e. place.name and place.full_name), profile information (i.e. profile description and location), and tweet text to map twitter's geospatial metadata for each tweet t_i to physical locations in the form of $\langle c_i, p_i \mid s_i \rangle$, where c denotes the city of a tweet and $p \mid s$ represent its corresponding province or state, respectively (Figure 2). To demo the proposed process, we used the Canadian geographical names dataset (i.e. location dictionary), provided by the Geographical Names Board of Canada. Each geographical name provided by this dataset is mapped to a province and is classified to a geographic area such as city, town, village, lake, administrative sectors, or recreational centers. As illustrated in Figure 2, for each t_i, we first use a simple search function to map the 'place.name' field associated with each tweet to its corresponding c_i in the location dictionary (LD). If found, the corresponding province field p_i was defined using equation 1:

$$p_i = \begin{cases} LD[c == c_i][p] & \text{if place.full_name}_{c_i}[1] == \text{null} \mid \text{place.full_name}_{c_i}[1] == LD[c == c_i][p], \\ place.full_name_{c_i}[1] & \text{otherwise} \end{cases} \tag{1}$$

Where place.full_name$_{c_i}$[1] denotes the the second component of the field when the first component is c_i (e.g. Ontario in the illustrative example of Figure 2). This will detect geographical areas with the same names but in different provinces (e.g. Leduc is a town/city in both Alberta and Quebec).

Steps 3-5 of the process deal with unstructured text objects that can come from all three sources of geospatial information. To extract the location information from these fields, we developed a string matching function to detect the longest common substring between the unstructured text of the dataset and the area field of the location dictionary (e.g. first time boating in Lake Louise #AB is mapped to $\langle$ Lake Louise, Alberta$\rangle$ instead of to $\langle$ Louise, Quebec$\rangle$). To manage the complexity of information extraction from the unstructured text, we only used a subset of areas listed in LD, with high population density (e.g. city, municipality, town, village, and country). Thus, we excluded areas classified as lake, mountain, river, bridge, or river.

Demographic attribute inference

The demographics variable of age and gender and the information about the source of each tweet (e.g. organization vs. real users) were not available within the dataset collected from Twitter. We estimated these variables for each tweet using the `M3inference` package in `python`[26], which uses a multimodal deep neural architecture for joint classification of age (binned

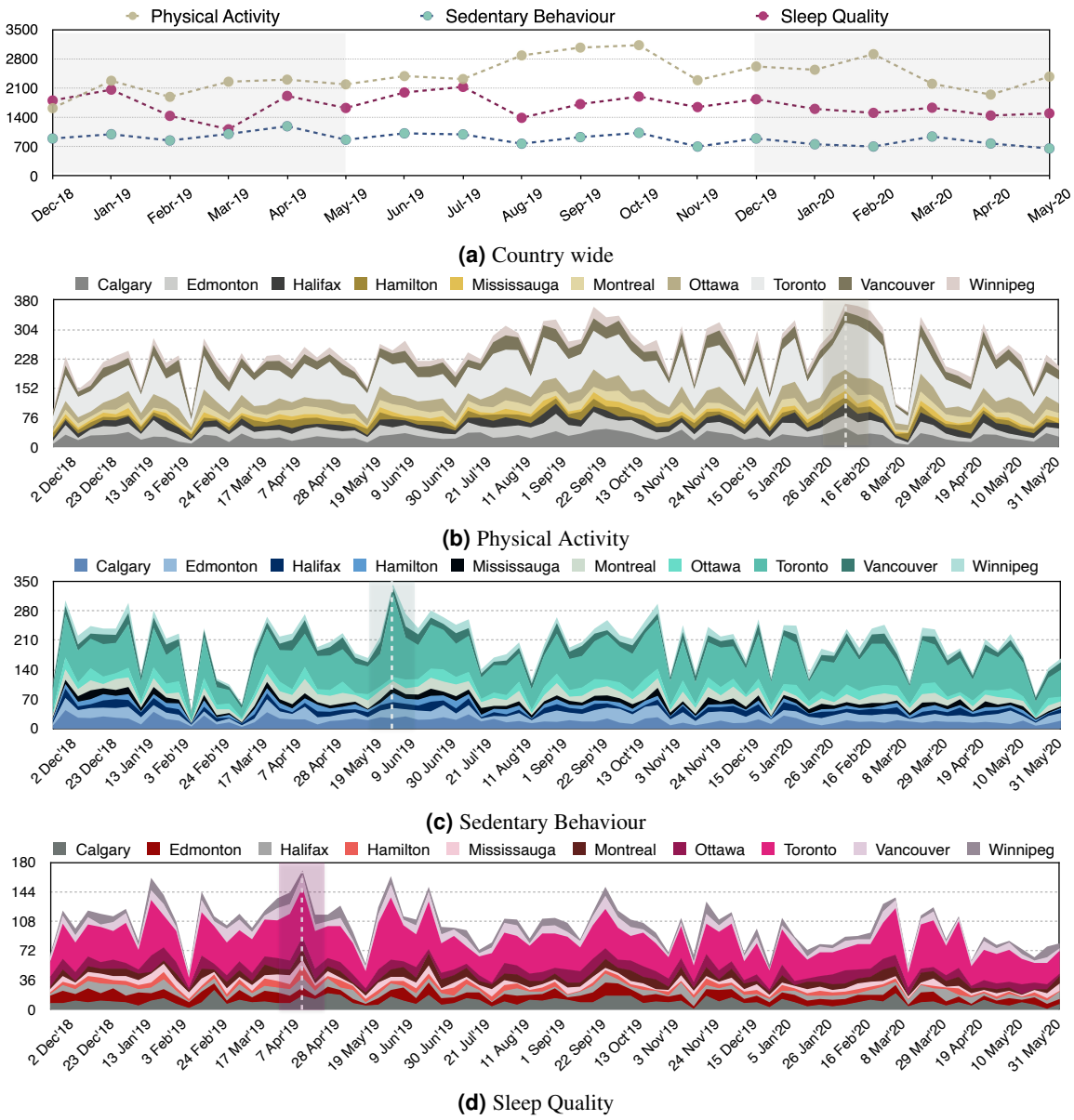


Figure 1. The temporal distribution of the tweets for the Canadian dataset. To make fair comparisons, we used the data from December 1st 2018 to May 31st 2020 for these visualizations and removed the data collected during the last two days of November 2018 and the first two weeks of June 2020.

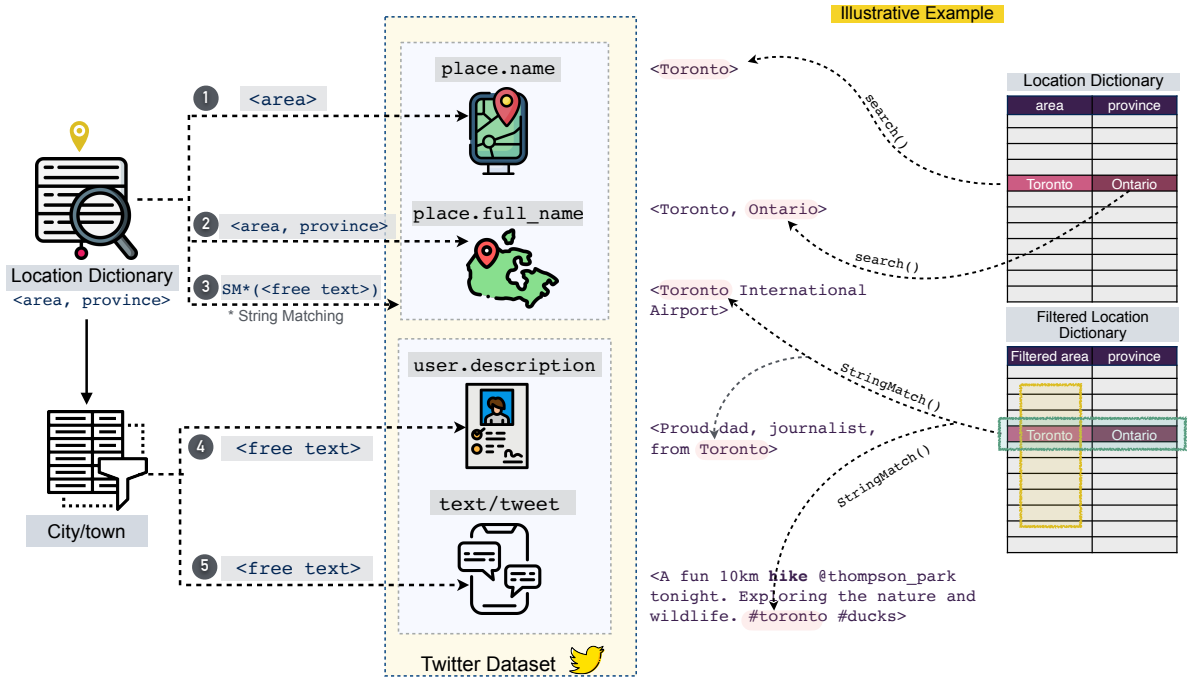


Figure 2. Five-step location inference process. The location dictionary is an external regional geographical metadata used to extract the exact locations of the tweets. The `area` field in this process refers to regions at different scales such as city, regional, municipality, town, township municipality, municipal district, dispersed rural community, village, or country.

into four groups: $\leq$ 18, 19-29, 30-39, $\geq$40), gender, and information-source of social media data. This approach uses four sources of information, including username, screen name, biography, and profile image of public profiles to develop two separate pipelines for processing a profile image and each of the three text sources of information. The models provided in this package are trained on 14.53M, 2.61M, and 23.86M profiles for each of the gender, age, and is-organisation categories, respectively.

Data Records

LPHEADA comprises 366,405 labels for 122,135 unique tweets generated by 72,749 unique users between November 28th 2018 to June 19th 2020. This dataset is organized into 12 subsets (three PASS categories for each of the four countries). Table 1 provides the demographics of the dataset, including the number of tweets per PASS category for each country, labels' characteristics, and demographics characteristics of the users. Each unique tweet is assigned a unique integer, known as `TweetID`. Each ID is mapped to the core Twitter metadata and to three crowd-generated labels for both binary and multi-class classification tasks. Figure 3 visualizes this hierarchy. As illustrated in this Figure, for each labelled tweet, LPHEADA provides the following data categories:

Labels. Let L denote the set of j unique class label, t present the tweet text, and w_k present the kth worker who labeled the tweet, where $k \in \{1,2,3\}$. Each $l_j \in L$ is defined based on two conditions: whether or not the tweet is self-reported ($c_1 \in \{0,1\}$); and whether or not the tweet reports a recent PASS experience ($c_2 \in \{0,1\}$). For each PASS category across the four countries, the dataset contains the following two subsets of labels for each tweet:

Multi-class Labels– In this subset, each tweet t is mapped to quadruple $\mathcal{L} = \langle tweetID, (w_1, l_{j_1}), (w_2, l_{j_2}), (w_3, l_{j_3}) \rangle$, where $j = 5$ and each l_j is defined based on the value of both c_1 and c_2 conditions and can be formulated as $\{11, 10, 01, 00\}$. We also let workers choose a fifth option, called `Unclear`, to ensure they do not give random labels to tasks that they are not confident of performing successfully.

Binary Labels– Each label l_j in this subset is defined based on logical `AND` relationship between conditions c_1 and c_2. Thus, $j = 3$ with $l_1 = 1$ if the tweet presents a self-reported PASS surveillance and $l_1 = 0$, otherwise. Like the multi-class category, each tweet is mapped to a quadruple, and there is a class called 'unclear' (l_3). The binary labels did not directly come from the AMT workers and were generated by dichotomizing the collected labels.

Table 1. The characteristics of the dataset. The three labels collected for each tweet are consolidated into a single label using majority voting. The discrepancy between the number of binary and multi-class labels is due to the way that majority voting calculates the truth label for each of these categories.

Variable	Canada (98,722)			United States (10,193)			United Kingdom (9,154)			Australia (4,067)		
	PA	SQ	SB	PA	SQ	SB	PA	SQ	SB	PA	SQ	SB
N	48,576	32,779	17,367	5,053	3,065	2,074	4,076	3,001	2,077	2,216	1,312	539
Labels												
Binary (%)												
Yes	15337 (32%)	11814 (36%)	6514 (38%)	1,196 (24%)	1,032 (34%)	731 (35%)	1092 (27%)	766 (25%)	356 (17%)	729 (33%)	499 (38%)	75 (14%)
No	33239 (68%)	20965 (64%)	10853 (62%)	3,857 (76%)	2,033 (66%)	1,343 (65%)	2984 (73%)	2235 (75%)	1721 (83%)	1487 (67%)	813 (62%)	469 (86%)
Multi-class												
YY	17,298 (36%)	12,818 (39%)	7,174 (41%)	1,431 (28%)	1,150 (38%)	831 (40%)	1,227 (30%)	934 (31%)	468 (23%)	846 (38%)	566 (43%)	97 (18%)
YN	6,583 (14%)	7,720 (24%)	1,895 (11%)	1,445 (29%)	1,104 (36%)	623 (30%)	1,037 (25%)	1,017 (34%)	916 (44%)	905 (41%)	404 (31%)	147 (27%)
NY	4,407 (9%)	2,242 (7%)	1,471 (9%)	634 (12%)	215 (7%)	147 (7%)	226 (6%)	332 (11%)	118 (6%)	208 (9%)	72 (5%)	33 (6%)
NN	19,622 (40%)	9,339 (28%)	6,502 (37%)	1,512 (30%)	564 (18%)	469 (22%)	1,559 (38%)	694 (25%)	572 (27%)	233 (11%)	258 (20%)	258 (48%)
NC	666 (1%)	660 (2%)	325 (2%)	31 (1%)	32 (1%)	10 (1%)	27 (1%)	24 (1%)	3 (≈ 0%)	24 (1%)	12 (1%)	4 (1%)
Users												
Unique (n)	22,601	16,984	11,490	4,911	2,994	2,048	3,810	2,653	2,002	1,735	1,004	517
Valid (n)	21,772	14,919	10,912	3,660	2,614	1,759	2,614	2,157	1,840	1,531	858	517
Gender												
Female (%)	8,471 (39%)	7,270 (49%)	4,486 (41%)	1,448 (40%)	1,537 (59%)	860 (51%)	911 (35%)	1,068 (50%)	625 (34%)	520 (34%)	401 (47%)	205 (40%)
Male (%)	13,301 (61%)	7,649 (51%)	6,426 (59%)	2,212 (60%)	1,077 (41%)	899 (49%)	1,703 (65%)	1,089 (50%)	1,215 (66%)	1,011 (66%)	457 (53%)	312 (60%)
Age range (%)												
≤ 18	1,772 (8%)	2,372 (16%)	1,361 (12%)	490 (13%)	518 (20%)	318 (18%)	136 (5%)	248 (12%)	136 (5%)	114 (7%)	132 (15%)	68 (13%)
(19-29)	5,804 (27%)	5,575 (37%)	3,358 (31%)	1,469 (40%)	1,450 (55%)	841 (48%)	639 (24%)	784 (36%)	539 (29%)	421 (28%)	301 (35%)	152 (29%)
(30-39)	5,609 (26%)	3,265 (22%)	2,605 (24%)	761 (21%)	363 (14%)	302 (17%)	705 (27%)	540 (25%)	486 (26%)	421 (28%)	209 (25%)	139 (27%)
(≥ 40)	8,527 (39%)	3,707 (25%)	3,588 (33%)	940 (26%)	283 (11%)	298 (17%)	1,137 (44%)	585 (27%)	659 (36%)	575 (37%)	216 (25%)	158 (31%)
Cities (n) [Top five]	Toronto	Toronto	Toronto	Los Angeles	Houston	Los Angeles	Cardiff	Leeds	Glasgow	Sydney	Sydney	Sydney
	Ottawa	Calgary	Calgary	Houston	Los Angeles	Houston	East Midlands	Sheffield	Manchester	Melbourne	Melbourne	Melbourne
	Calgary	Edmonton	Ottawa	Manhattan	San Antonio	Brooklyn	Bristol	London	Sheffield	Brisbane	Brisbane	Brisbane
	Vancouver	Ottawa	Edmonton	Chicago	Chicago	Chicago	Lambeth	Liverpool	Leeds	Perth	Adelaide	Perth
	Edmonton	Montréal	Vancouver	Florida	Brooklyn	Florida	Liverpool	Scotland	Edinburgh	Adelaide	Perth	Adelaide
Source												
Organization	3,109 (14%)	727 (5%)	1,105 (10%)	63 (2%)	97 (3%)	76 (4%)	368 (14%)	94 (4%)	143 (8%)	135 (9%)	22 (3%)	26 (5%)
Users	18,663 (86%)	14,192 (95%)	9,807 (90%)	3,391 (93%)	2,551 (98%)	1,1,638 (96%)	2,246 (86%)	2,063 (96%)	1,697 (92%)	1,396 (91%)	836 (97%)	491 (95%)

User's demographics data. In the demographic dataset, each tweet t is mapped to quadruple $\mathcal{D} = \langle tweetID, a, g, o \rangle$, where $g \in \{\text{male, female}\}$ presents the gender of the user who posted the tweet, $a \in \{\leq 18, 19-29, 30-39, 40 \leq\}$ represents their age range, and $o \in \{0,1\}$ shows the source of the tweet (i.e. $o = 1$ if the tweet posted by an organization, $o = 0$ otherwise). Figure 4 shows the demographic distribution of the Canadian dataset based on gender and age-range of the unique users associated with tweet IDs. We can see that at least 75% of female users in this dataset are inferred to be younger than 40, while this number for the male users is 60%. Also, the most populated age category for female users across all PASS categories is 19-29, while this range for male users is ≥ 40. Excluding the (female, sleep quality) category, the age range ≤ 18, regardless of the user's sex, is the least populated category across all PASS categories.

Inferred location data. Each row of the location dataset is presented in the form of $\mathcal{A} = \langle tweetID, c, p \rangle$, where c and p denote the city and province/state associated with each tweet, respectively. Using the TweetID parameter, the location data can be mapped to other datasets, including labels, user's demographics, time, and twitter metadata. Each of the c and p variables are inferred based on the raw variables in Twitter metadata, including text, place objects, and user's profile description (Figures 2 and 3). For example, Figure 5 shows the distribution of labelled tweets for each PASS category across Canadian provinces (i.e. p). For the top five provinces, the overall size of the dataset is directly proportional to the population size of each province. However, as only English tweets from Twitter users are included in the dataset, LPHEADA represents only English-speaking Quebecor's and Francophone Quebecor's tweets in English, placing the province in fourth place. Moreover, with a lower population than BC, Alberta contains more SB and SQ tweets and places in second place (Figure 5).

Temporal data. The temporal dataset inferred from the 'created_at' field of the Twitter metadata presents the adjusted time of each tweet based on the tweet's location. Each row of this dataset is presented in the form of $\mathcal{T} = \langle tweetID, h, d, m \rangle$, where h, d, and m present the hour, weekday, and month associated with each tweet, respectively. The 'year' value does not need any adjustment and can be extracted directly from the original 'created_at' field. For example, Figure 1 represents the frequency of tweets in each PASS category across Canada at both national (Figure 1a) and city levels (Figures 1b,1c, and 1d). The highlighted area in Figure 1a demonstrates the dataset's temporal windows that can be used to compare different aspects of PASS surveillance between 2019 and 2020.

Twitter metadata. In addition to the inferred data records mentioned above, TweetIDs presented in LPHEADA can be used to retrieve Twitter metadata. This metadata, in addition to the tweet text, place object, time of the tweet, and user IDs, provides more details on the tweet and user objects, including:

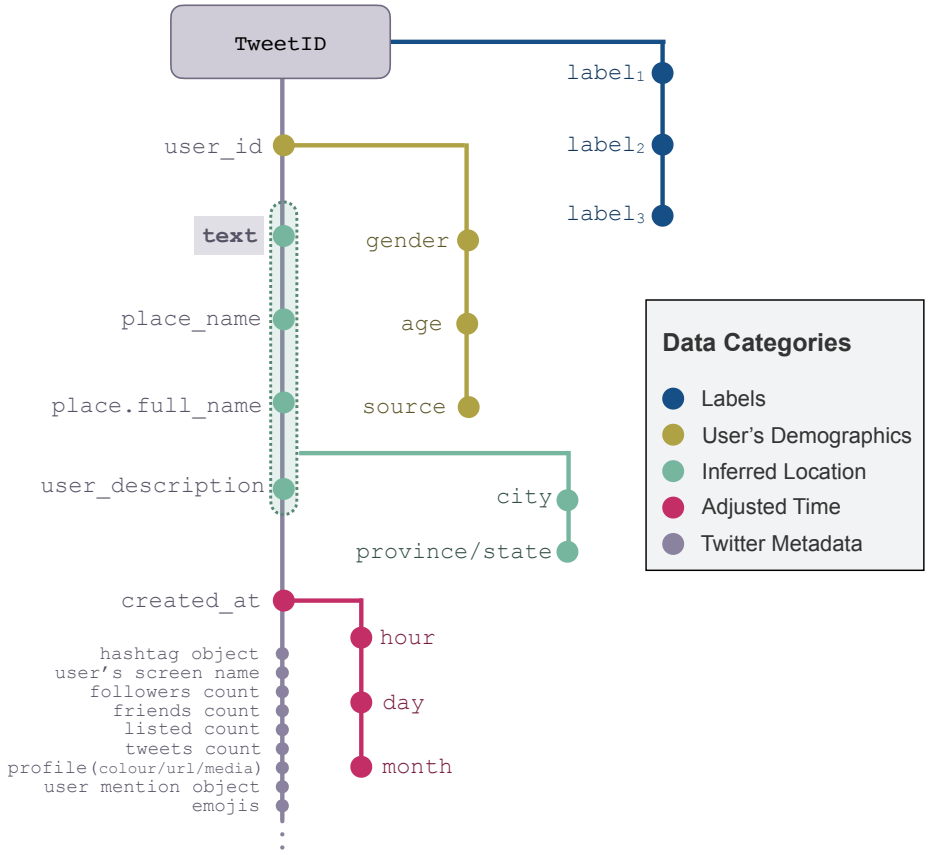


Figure 3. Overview of information tracking using TweetID. Each `tweet/text` is identified by a unique `TweetID` (provided by LPHEADA). This ID is mapped to metadata that include `user_id, place_name`, `place_full_name, user_description`, and `created_at`. Three labels are provided for each `TweetID` that can be used for developing machine learning models. `user_id` was used to infer the demographics of each tweet including `genders, age range`, and `source`. Adjusted time (`month/day/hour`) was extracted using `created_at`, and `text, place_name, place_full_name`, and `user_description` was used to identify the `city` and `state/province` mapped to each `TweetID`.

User object: This object comprises of user's screen name, description, followers count, friends count, listed count (i.e. the number of public lists that the user is a member of), tweets count (i.e. the number of (re)tweets issued by the user), and profile characteristics (e.g. image, colour, and URLs).

Tweet object: This object comprises of hashtags mentioned in each tweet, emojis, user mentions, urls, and media (e.g. images and videos). For example, Figure 6 illustrates the distribution of the top 10 hashtags per label for each of the PASS categories in the entire dataset. Hashtags are basically keywords or word strings prefixed with the symbol '#' that are used for categorizing and communicating tweets pertaining to the same topics. While the high level of intersection between the hashtags of positive and negative classes in our dataset makes this feature a less discriminating feature for the development of machine learning models (e.g. annotated hashtags in Figure 6), this field can still be utilized by PASS-related advocacy campaigns on Twitter to brand their movement and open up their campaigns to users who need more information about the context[27]. As tagged tweets are easily archived and accessible, the hashtag field can be effectively leveraged to improve the public's engagement in digital PHS discussions.

Technical Validation

To verify the quality of crowd-generated labels and set a baseline for the dataset, we conducted four studies. First, we used a series of statistical inference models to verify the quality of the labels provided in this dataset. Second, we evaluated the semantic consistency between the datasets collected from the countries included in our repository. Third, we trained and tested 12 binary classifiers using labels provided in the dataset. Finally, to investigate the structural differences between all subsets of LPHEADA, we conducted linguistic and lexical analysis and visualized the results for further comparisons. Moreover, to

Physical Activity (21,772 users)

Gender	<=18	19-29	30-39	>=40	Total
Female	9% (779)	37% (3,160)	28% (2,408)	25% (2,124)	39% (8,471)
Male	8% (993)	20% (2,704)	24% (3,201)	48% (6,403)	61% (13,301)

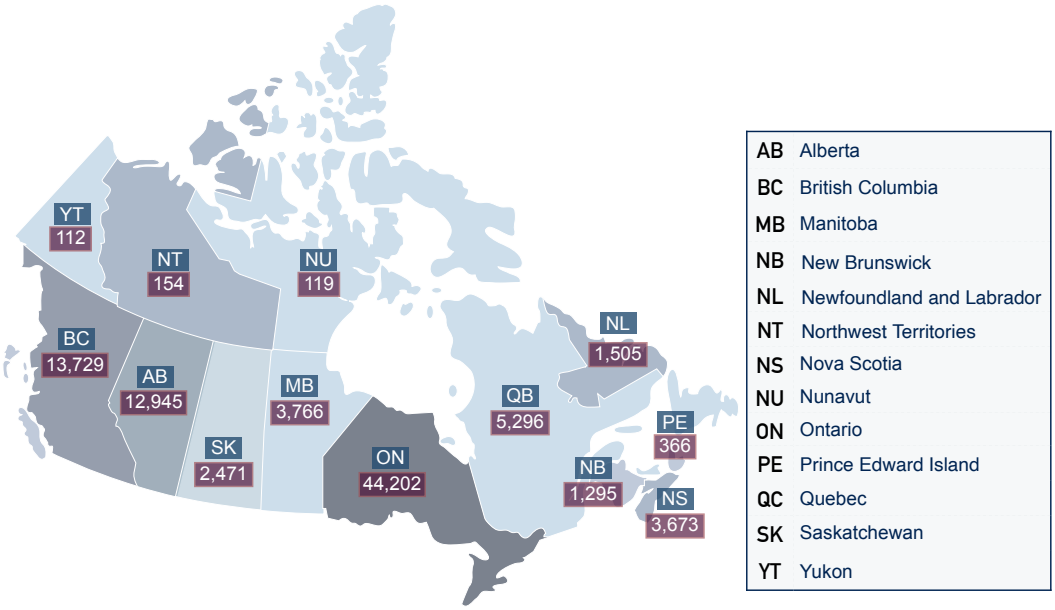
Sedentary Behaviour (10,912 users)

Gender	<=18	19-29	30-39	>=40	Total
Female	14% (607)	41% (1,822)	25% (1,108)	21% (949)	41% (4,486)
Male	12% (754)	24% (1,536)	23% (1,497)	41% (2,639)	59% (6,426)

Sleep Quality (14,919 users)

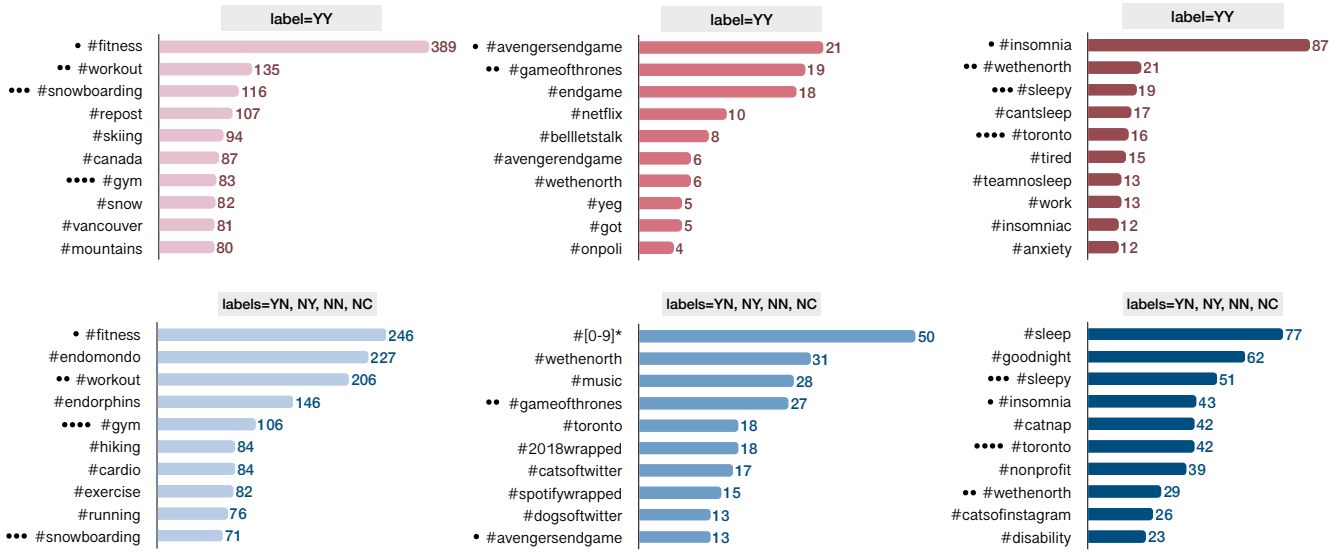
Gender	<=18	19-29	30-39	>=40	Total
Female	17% (1,267)	46% (3,307)	28% (1,569)	15% (1,127)	49% (7,270)
Male	15% (1,105)	30% (2,268)	22% (1,696)	34% (2,580)	51% (7,649)

Figure 4. Demographic information of 42,603 unique users associated with the tweets originated in Canada.



	AB	BC	MB	NB	NL	NT	NS	NU	ON	PE	QC	SK	YT
Physical Activity	6,692	7,512	1,726	598	797	80	1,780	56	20,092	192	2,177	1,398	67
Sleep	4,048	4,032	1,321	439	430	56	1,274	42	15,788	100	2,226	661	25
Sedentary Behaviour	2,205	2,185	719	258	278	18	619	21	8,322	74	893	412	20

Figure 5. Geo-spatial details of the Canadian dataset



(a) Physical Activity **(b)** Sedentary Behaviour **(c)** Sleep Quality

Figure 6. The distribution of top 10 hashtags per label for each of the physical activity, sedentary behaviour, and sleep quality categories. The number at the end of each bar presents the frequency of its corresponding hashtag. The intersections between two class of labels for each PASS category are annotated using filled circles (•). This Figure is based on all data collected from Canada, US, UK, and Australia.

address unseen technical issues of the dataset, we provide a public issue tracker for handling bug reports, describing solutions to technical issues, data updates, and other issues and contributions.

Methods of label agreement

To measure the consistency of labels generated by AMT workers, we calculated label consistency ($\mathcal{LC}$) as the average entropy of the collected labels for each PASS category[28]. For each tweet $t_i \in \mathcal{T}_s$, where $\mathcal{T}_s$ denotes the set of all tweets related to surveillance category $s \in \{\text{physical activity, sleep quality, sedentary behaviour}\}$, n_{ij} defines the number of answers given to the jth choice ($j \in \{1,2,3,4,5\}$), as we have five choices for each tweet). We calculate $\mathcal{LC}_s$ as:

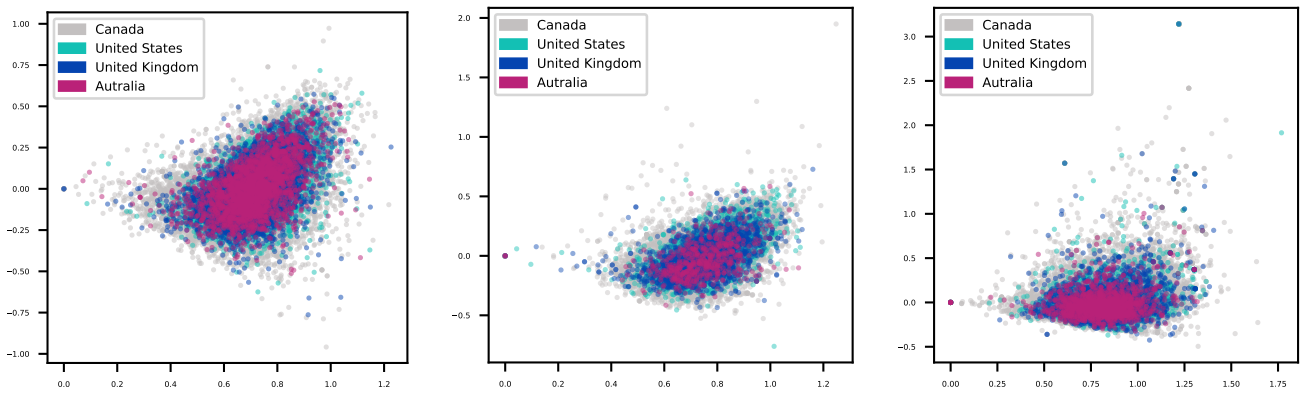
$$\mathcal{LC}_s = 1 - \left(-\frac{1}{|s|} \cdot \sum_{i=1}^{i=|s|} \sum_{j=1}^{j=5} \frac{n_{ij}}{3} \times log_3 \frac{n_{ij}}{3} \right) \tag{2}$$

$|s|$ denotes the size of the surveillance category s and as we collect three labels for each tweet, the denominators in the entropy formula receive the constant value of three. $\mathcal{LC}$ ranges from 0 to 1, and the values close to zero show less consistency between the workers' input. After calculating the label consistency ($\mathcal{LC}$) for each PASS category, we had $\mathcal{LC} > 0.52$ for the multi-class labelling, and $\mathcal{LC} > 0.73$ for the binary labelling tasks.

To consolidate the collected labels for each tweet, we used the majority voting (MV) technique (Table 1). Defining the estimated label as $\hat{l}_i$, and the submitted label by worker w as l_w, the MV approach, for a binary labeling task, assigns 1 to $\hat{l}_i$ if $\frac{1}{3} \cdot \sum_{i=1}^{3} l_w^i > 0.5$, and assigns 0, otherwise. The discrepancy between the number of binary and multi-class labels presented in Table 1 caused by the way that MV approach calculates the truth label for each of these categories. In addition to MV, there are models, such as David and Skene (DS)[29], Raykar (RY)[30], and generative model of labels, abilities, and difficulties (GLAD)[31], that incorporate the error rate of annotators (workers), task complexity, and context-sensitive features into the inference process and can be utilized to predict truth labels from crowd-labelled data.

Semantic consistency

To validate the semantic consistency of the datasets collected from different countries, we transformed the dataset of each PASS category into a semantic space of low dimensionality using latent semantic analysis (LSA). For the vector presentation of each dataset, to capture high-level semantics of the text, we ran the PASS category of each country through a pre-trained word2vec embedding model. This model contains 300-dimensional vectors of three million words and phrases trained on 100 billion words from a Google News dataset. Then, the resulting 300-dimensional vectors were averaged for each tweet. For each tweet



(a) Physical Activity **(b)** Sedentary Behaviour **(c)** Sleep Quality

Figure 7. Scatter plots of the first two dimensions of latent semantic analysis (LSA) performed on PASS categories and classified based on the geographic source of the data.

T composed of words $\langle w_1, w_2, \ldots, w_n \rangle$, with $\vec{\boldsymbol{u}}_{w_i}$ defining the embedding of w_i, the embedding of tweet T can be calculated as: $\vec{\boldsymbol{u}}_T = \frac{1}{n}\sum_{i=1}^{i=n} \vec{\boldsymbol{u}}_{w_i}$. We then applied truncated SVD on the new vectorized dataset and only kept the top two dimensions of the dataset containing the most variance (e.g. those directions in vector space of the dataset that contain more information). The scatter plots presented in Figure 7 illustrate our datasets in a two-dimensional latent semantic space. The high level of overlap between the datasets of each PASS category indicates that the data from different countries cover similar semantic space, but the space is scaled differently based on the size of datasets.

Moreover, to further investigate the internal consistency of the datasets presented in this paper, we trained three convolutional neural network (CNN) multiclass classifiers (i.e. one for each PASS category) to classify the tweets into one of Canada, USA, Australia, or UK classes. Given the highly imbalanced distribution of the classes in our dataset due to the highly unequal number of samples from each country, we used average precision metric (AP) to measure the discrimination ability of our predictive models. The poor performance of these classifiers (AP_{PA}: 37%, AP_{SB}: 32%, AP_{SQ}: 31%) in detecting each tweet's country implies a high level of semantic and syntactic cohesion among the 4 countries in our dataset.

Classification of PASS categories

For the PASS classification experiment, we utilized a standard CNN classifier with one layer of convolution with global max-pooling on top of a word2vec embedding trained on 100 billion words of Google News. The vectors have a dimensionality of 300 and were trained using the continuous bag-of-words architecture[32]. We used the binary labels of the dataset to train and evaluate the model on each of 12 datasets provided in LPHEADA. Due to the imbalanced distribution of binary labels across all of these datasets (Table 1), in addition to precision, recall, F1, and AUC scores, we used average precision (AP) to measure the weighted mean of precision at different thresholds to make the score robust to heterogeneous and imbalanced class distributions. Like AUC score, AP is a model-wide and threshold-free evaluation metric. However, for imbalance class distributions with the negatives outnumbering the positives, AP is more informative than AUC, as it mainly evaluates the fraction of true positive samples among positive predictions and is more robust to the relationship between false-positive (FP) and false-negative (FN) rates[33]. As shown in Table 2, for each of the Canada, USA, and UK datasets, we find a steady increase in the overall performance of the classifier as the size of the dataset increases (i.e. $\mid PA \mid > \mid SQ \mid > \mid SB \mid$). Interestingly, the UK dataset achieves the highest performance for the PA category than all the other countries.

Linguistic Properties

To understand and validate the linguistic properties of each dataset, we measured and visually compared the following metrics for each PASS category grouped by countries: (i) sentence count, (2) grammar score (i.e. number of grammar errors), (3) the average number of syllables per word and the average sentence length (i.e. Flesch-Kincaid Grade Level index[34]), (4) the average number of words per sentence and the percentage of words with three or more syllables (i.e. Gunning Fog index[35]), (5) a combination of average sentence length and percentage of difficult words (i.e. Dale–Chall readability[36]), (6) Sentence length and number of polysyllables (i.e. Linsear Write readability[36]), (7) number of characters (i.e. Coleman-Liau Index[36,37]), (8) the average number of characters per word and number of words per sentence (i.e. automated readability index[38]), and (9) the text standard score based on number of sentences, words, syllables, and characters in each tweet (i.e. text readability consensus). Figure 8 illustrates these comparisons based on the minimum (red), average (pink), and maximum (light pink) value of each

Physical Activity				
Metrics	Canada	United States	Unided Kingdom	Australia
Precision	78%	75%	80%	65%
Recall	78%	76%	81%	68%
F1	78%	76%	80%	64%
AUC_{ROC}	83%	76%	84%	64%
AP	81%	75%	81%	66%
Sedentary Behaviour				
Metrics	Canada	United States	Unided Kingdom	Australia
Precision	73%	62%	79%	74%
Recall	73%	64%	83%	86%
F1	73%	62%	79%	80%
AUC_{ROC}	76%	62%	63%	57%
AP	78%	65%	66%	66%
Sleep Quality				
Metrics	Canada	United States	Unided Kingdom	Australia
Precision	76%	70%	73%	61%
Recall	76%	70%	76%	60%
F1	75%	70%	73%	60%
AUC_{ROC}	81%	70%	66%	63%
AP	83%	72%	71%	64%

Table 2. Binary classification of tweet using bidirectional LSTM. We used the same classifier to classify the dataset based on the countries: AP (PA: 37%, SB: 32%, and SQ: 31%).

feature. While all datasets have similar behaviour in terms of each feature's minimum value, the Canadian dataset has a lower score for the average number of syllables per word and the average sentence length for all PASS categories. Interestingly, the sleep quality dataset, compared to other PASS categories, has a higher value for the maximum number of 'grammar errors' and 'sentence count' metrics, while all datasets show the same behaviour for the minimum and the average values of these metrics. These location-specific linguistic characteristics should be considered when utilizing these datasets to train and evaluate PASS surveillance ML models. For example, a model trained on the Canadian dataset may not present some linguistic features of a dataset that originated in Australia and vice versa.

Usage Notes

Data access

LPHEADA is publicly available online. Retrieving the data using TweetIDs provided in the dataset requires Twitter's approval. Data access also requires a data use agreement between the data user and Twitter to govern the access and use of the licensed material returned by Twitter API. Once approved, Twitter metadata can be downloaded as a JSON file to be mapped to other subsets of data (e.g. location, time, and demographics) using the TweetID field.

Example Usage

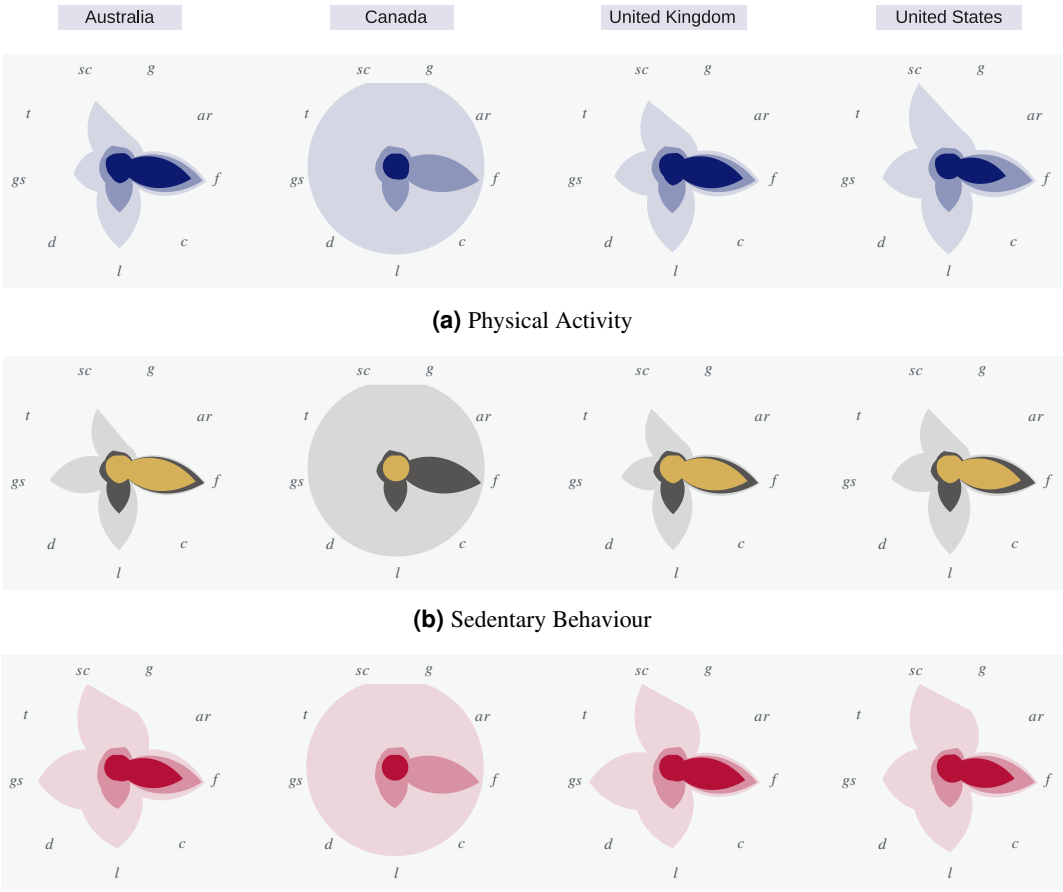
We have provided publicly accessible instructions and Jupyter Notebooks[39,40] to illustrate the application of the data (available at https://github.com/data-intelligence-for-health-lab/Lpheada-Labelled-Public-HEAlth-DAtaset).

Code availability

All code used in the experiments described in the manuscript was written in Python 3 and is available through our GitHub repository (https://github.com/data-intelligence-for-health-lab/Lpheada-Labelled-Public-HEAlth-DAtaset). We provide all necessary instructions, required libraries, and sample Jupyter Notebooks allowing replicating our experiments and utilizing the dataset.

References

1. Kemp, S. Digital 2020: July global statshot. *DATAREPORTAL. Available online: https://datareportal. com/reports/digital-2020-july-global-statshot (accessed on 8 January 2021)* (2020).
2. Fung, I. C.-H., Tse, Z. T. H. & Fu, K.-W. The use of social media in public health surveillance. *West. Pac. surveillance response journal: WPSAR* **6**, 3 (2015).



(a) Physical Activity

(b) Sedentary Behaviour

(c) Sleep Quality

Figure 8. *sc*: sentence count, *g*: Gunning Fog index, *ar*: automated readability index, *f*:Flesch-Kincaid grade level, *c*: Coleman-Liau Index, *l*: Linsear Write readability, *d*: Dale–Chall readability, *gs*: grammar score, and *t*: text readability consensus.

3. Brownstein, J. S., Freifeld, C. C. & Madoff, L. C. Digital disease detection—harnessing the web for public health surveillance. *The New Engl. journal medicine* **360**, 2153 (2009).

4. Sinnenberg, L. *et al.* Twitter as a tool for health research: a systematic review. *Am. journal public health* **107**, e1–e8 (2017).

5. Pershad, Y., Hangge, P. T., Albadawi, H. & Oklu, R. Social medicine: Twitter in healthcare. *J. clinical medicine* **7**, 121 (2018).

6. Raghupathi, W. & Raghupathi, V. Big data analytics in healthcare: promise and potential. *Heal. information science systems* **2**, 1–10 (2014).

7. Shakeri H.A, Z. & et al. Digital public health surveillance: a systematic scoping review. *NPJ digital medicine.* **In press**.

8. Teutsch, S. M. Considerations in planning a surveillance system. *Princibles practice public health surveillance* 18–28 (2010).

9. Declich, S. & Carter, A. O. Public health surveillance: historical origins, methods and evaluation. *Bull. World Heal. Organ.* **72**, 285 (1994).

10. Croner, C. M. Public health, gis, and the internet. *Annu. Rev. Public Heal.* **24**, 57–82 (2003).

11. Graham, M., Hale, S. A. & Gaffney, D. Where in the world are you? geolocation and language identification in twitter. *The Prof. Geogr.* **66**, 568–578 (2014).

12. Doan, S., Ritchart, A., Perry, N., Chaparro, J. D. & Conway, M. How do you #relax when you're #stressed? a content analysis and infodemiology study of stress-related tweets. *JMIR public health surveillance* **3**, e35 (2017).

13. Woo, H. *et al.* Identification of keywords from twitter and web blog posts to detect influenza epidemics in korea. *Disaster medicine public health preparedness* **12**, 352–359 (2018).

14. Chorianopoulos, K. & Talvis, K. Flutrack. org: Open-source and linked data for epidemiology. *Heal. informatics journal* **22**, 962–974 (2016).

15. Jimeno-Yepes, A., MacKinlay, A., Han, B. & Chen, Q. Identifying diseases, drugs, and symptoms in twitter. (2015).

16. Aphinyanaphongs, Y., Lulejian, A., Brown, D. P., Bonneau, R. & Krebs, P. Text classification for automatic detection of e-cigarette use and use for smoking cessation from twitter: a feasibility pilot. In *Biocomputing 2016: Proceedings of the Pacific Symposium*, 480–491 (World Scientific, 2016).

17. Mowery, D. *et al.* Understanding depressive symptoms and psychosocial stressors on twitter: a corpus-based study. *J. medical Internet research* **19**, e48 (2017).

18. Müller, M. M. & Salathé, M. Crowdbreaks: Tracking health trends using public social media data and crowdsourcing. *Front. public health* **7**, 81 (2019).

19. O'Connor, K. *et al.* Pharmacovigilance on twitter? mining tweets for adverse drug reactions. In *AMIA annual symposium proceedings*, vol. 2014, 924 (American Medical Informatics Association, 2014).

20. Watkins, N. G., Lovell, N. H. & Larsen, M. E. Smct-an innovative tool for mental health analysis of twitter data. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 4114–4117 (IEEE, 2018).

21. Zhao, L. *et al.* Simnest: Social media nested epidemic simulation via online semi-supervised deep learning. In *2015 IEEE International Conference on Data Mining*, 639–648 (IEEE, 2015).

22. Lee, I.-M. *et al.* Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. *The lancet* **380**, 219–229 (2012).

23. Roberts, K. *et al.* At-a-glance: the physical activity, sedentary behaviour and sleep (pass) indicator framework. *Heal. Promot. Chronic Dis. Prev. Canada: Res. Policy Pract.* **37**, 252–256 (2017).

24. Paolacci, G., Chandler, J. & Ipeirotis, P. G. Running experiments on amazon mechanical turk. *Judgm. Decis. making* **5**, 411–419 (2010).

25. Tweet geospatial metadata. *Available online: https://developer.twitter.com/en/docs/tutorials/tweet-geo-metadata (accessed on 12 January 2021)* .

26. Wang, Z. *et al.* Demographic inference and representative population estimates from multilingual social media data. In *The World Wide Web Conference*, 2056–2067 (2019).

27. Ojo, A., Guntuku, S. C., Zheng, M., Beidas, R. S. & Ranney, M. L. How health care workers wield influence through twitter hashtags: Retrospective cross-sectional study of the gun violence and covid-19 public health crises. *JMIR Public Heal. Surveillance* **7**, e24562 (2021).

28. Zheng, Y., Li, G., Li, Y., Shan, C. & Cheng, R. Truth inference in crowdsourcing: Is the problem solved? *Proc. VLDB Endow.* **10**, 541–552 (2017).

29. Dawid, A. P. & Skene, A. M. Maximum likelihood estimation of observer error-rates using the em algorithm. *J. Royal Stat. Soc. Ser. C (Applied Stat.* **28**, 20–28 (1979).

30. Raykar, V. C. *et al.* Learning from crowds. *J. Mach. Learn. Res.* **11** (2010).

31. Whitehill, J., Wu, T.-f., Bergsma, J., Movellan, J. R. & Ruvolo, P. L. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in neural information processing systems*, 2035–2043 (2009).

32. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546* (2013).

33. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one* **10**, e0118432 (2015).

34. Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L. & Chissom, B. S. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Tech. Rep., Naval Technical Training Command Millington TN Research Branch (1975).

35. Gunning, R. *et al.* Technique of clear writing. (1952).

36. Mohammadhassanzadeh, H. *et al.* Using natural language processing to examine the uptake, content, and readability of media coverage of a pan-canadian drug safety research project: Cross-sectional observational study. *JMIR formative research* **4**, e13296 (2020).

37. Coleman, M. & Liau, T. L. A computer readability formula designed for machine scoring. *J. Appl. Psychol.* **60**, 283 (1975).

38. Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L. & Chissom, B. S. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Tech. Rep., Naval Technical Training Command Millington TN Research Branch (1975).

39. Pérez, F. & Granger, B. E. Ipython: a system for interactive scientific computing. *Comput. science & engineering* **9**, 21–29 (2007).

40. Kluyver, T. *et al.* Jupyter notebooks-a publishing format for reproducible computational workflows. In *ELPUB*, 87–90 (2016).

Acknowledgements

This work was supported by a postdoctoral scholarship from the Libin Cardiovascular Institute and the Cumming School of Medicine, University of Calgary. Also, this work was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (RGPIN-2014-04743). The Public Health Agency of Canada funded the Amazon Mechanical Turk costs.

Author contributions statement

Z.SH. was responsible for data collection and curation, model development, documentation and maintenance of the dataset, data analysis and visualization and wrote the paper. G.B. and W.T. reviewed the paper and provided comments. J.L. contributed to the conception and design of the study and revised the manuscript.

Competing interests

The authors declare no competing interests.