Social Determinants of Health: Insights from Location Big Data

Meghanath Macha
Information Systems and Management, Carnegie Mellon University, mmacha@cmu.edu

Beibei Li
Information Systems and Management, Carnegie Mellon University, beibeili@andrew.cmu.edu

Natasha Zhang Foutz
McIntire School of Commerce, University of Virginia, nfoutz@virginia.edu

Working Paper, February 11, 2022*

ABSTRACT

Rocketing hospitalization rates and costs call for deeper understanding of the connection between health outcomes and individuals' social determinants of health, such as lifestyle and socio-economics. Such knowledge holds important implications to health marketing, policy making, and policy communication. Building on the literature that has focused on either lifestyle identification or the association between health outcomes and limited behavioral features derived from small samples, we propose a novel framework leveraging the population-scale location data that capture granular individual behavior 24/7. This framework integrates unsupervised topic models and sequential deep learning models to characterize individual lifestyles and quantify their association with future hospitalizations, while integrating other social determinants of health. Applied to 45 million location records from a major metropolitan region in the U.S., the framework successfully uncovers heterogeneous lifestyles. Several key findings then emerge. An individual's lifestyle choice turns out to be a more critical predictor of future hospitalization than his/her socio-economic factors or accessibility to healthcare facilities. Lower income population can present healthy lifestyles, while high-income population can present unhealthy ones. Population with lower accessibility to healthcare or facilities can present healthy lifestyles; while population with higher accessibility can present unhealthy ones. Individuals with busy, varying work routines and limited gym visits are 2.01 times likely to be hospitalized within a year, compared to the population average. Importantly, *regularity*, rather than total time spent, toward healthy or unhealthy activities, predict future hospitalization. Overall, an individual's lifestyle choice is more critical than the socio-economic, accessibility and their community factors, consistent with a recent review on social determinants in EHR [4].

1. Introduction

Social Determinants of Health. The U.S. sees 35.7 million hospital stays per year, representing a hospitalization rate of 104.2 stays per 1,000 population. Hospitalizations incur enormous

* The authors thank the participants at the CMU and UVA workshops for valuable feedback. The authors contributed equally.

costs, $417.4 billion per year and $11,700 per stay. These costs will continue to skyrocket, as the hospitalization rate increases with age, from 17.1 (1-17 years old) to 455.7 (85+) per 1,000 population [10, 32]. Moreover, only 20% of an individual's health is attributable to access to healthcare, whereas 80% to the remaining components of the *social determinants of health*: physical environment, socio-economic factors, and lifestyle choices (American Hospital Association [18]). It is thus imperative to leverage new data and big data, beyond the conventional patient data available to healthcare professionals, and develop new tools to understand factors that predict individual health risk, reduce hospitalization rate, and promote population health via effective marketing strategies.

A variety of literature empirically investigated the importance of the factors beyond the clinical wall on health outcomes. Diet, smoking cessation, exercise, and sleep are shown to critically improve life expectancy and reduce hospitalization costs (cf. [4] for a recent review). Up to one-third of premature deaths in the U.S. arise from conditions modifiable via lifestyle choices [21]. Socioeconomic factors, such as income and education, are associated with life expectancy with the greatest disparities occurring in the mid-adulthood [15]. These studies primarily rely on identification of social determinants of health from electronic health records (EHR), medical claims, and individual surveys. Deviating from the status quo, this research examines the associations between health outcomes and individual's social determinants of health by leveraging atomic, longitudinal individual smartphone location data.

Location Data. Identifying social determinants of health from mobile location data presents several significant advantages over conventional data sources. First, mobile location data are straightforward to collect, merely an app permission away, tracked in the background in most mobile ecosystems, and readily accessible to data users. Sustained data collection also requires little to no effort from an individual or data user, compared to a hospital visit or medical claim filing. Second, mobile location data offer an extensive, spatio-temporal profile of an individual by delineating day-to-day behavior, mobility, lifestyle choices, and social relations [12]. Meanwhile, these data embed rich points-of-interests (POIs), such as restaurants, gyms, pharmacies, and hospitals, home and work locations [23]. Third, mobile location data portray a much richer context than EHR, such as the longer-term precursors (i.e., locations visited and behaviors before) and aftermath of a hospital visit. Fourth, mobile location information can help fill any data void (e.g., when no EHR or health insurance is available for a first-time patient) or verify survey responses. Fifth, mobile location data permit continuous monitoring of social determinants of health, thus facilitating adaptive interventions to mitigate future health risk [37]. In a nutshell, we aim to propose a framework to identify the social determinants of health from these behaviorally rich individual location data and empirically quantify their association with future health outcomes of immense economic and societal values.

Literature. Studies across disciplines have aimed to understand individual behavior from location data – characterizing mobility patterns [14], social ties [26, 8], and shopping patterns [17]. While most behavioral patterns have been leveraged for advertising [25], their relationship with health outcomes is receiving increased attention. On one hand, researchers primarily from Computer Science focus on identifying macro representations of an individual's subset of daily activities without linking to long-term health risk [22, 9, 36]. On the other, the medical community examines health outcomes, such as depression, [33], schizophrenia symptoms [2] and other standard clinical measurements [30], by analyzing micro activities, such as sleep patterns, gait, and activity rhythms. Both literature rely on sensor data from fewer than 200 individuals.

In comparison, our research is distinctive on multiple fronts. We extract a *comprehensive* range of behavioral patterns, including work, leisure, commute, and fitness, to capture "lifestyle", defined in sociology and marketing as "an activity that exhibits a pattern of behavior, consumption or leisure" [6]. We further integrate these macro representations of lifestyle with micro-level features inferred from the location data, such as accessibility to healthcare facilities and socioeconomic status, to construct an extensive profile of an individual's social determinants of health. We then quantify the link between these determinants and a key health outcome - hospitalization. Our examination of the population-scale data also permits empirical generalization and policy guidance.

Key Findings. We analyze the year-long location data from the Baltimore and DC metropolitan area. For Baltimore residents, the lifestyle identification reveals that while as expected the weekday (weekend) lifestyle is primarily characterized by work (home) routines, heterogeneous lifestyles, such as workaholics and fitness regulars, do emerge. We also find that lower income population can present healthy lifestyles, while high-income population can present unhealthy ones. Further, population with lower accessibility to healthcare or facilities can present healthy lifestyles; while population with higher accessibility can present unhealthy ones. Individuals with constant work, limited fitness, or stay at home on weekdays are 2.01 and 1.47 times more likely to have a future hospitalization within the next year compared to average (2.45%). In contrast, those who conduct fitness on weekends or weekdays are much less likely (0.52 and 0.65 times, respectively) to have a hospitalization. Interestingly and importantly, *regularity*, rather than total time spent, toward healthy activities or unhealthy activities, significantly predicts future hospitalization. Overall, an individual's lifestyle choice is more critical than the socio-economic, accessibility and their community factors. These findings strongly align with a recent review article of social determinants of health in EHR and their impact on analysis and risk prediction by [4].

Finally, to quantify the health risk, we jointly represent the multiple facets of an individual's social determinants and develop a sequential deep learner to predict future hospitalization. The proposed learner, dealing with a huge class imbalance (2.45 % on average are hospitalized) achieves

a PR AUC and ROC AUC of 0.28 and 0.85 respectively. From an ablation study of the proposed learner and several baselines, we confirm that individual behavioral features, such as lifestyles and day-to-day activities, significantly contribute to the predictive performance for both Baltimore (16.6% increase in PR AUC) and D.C. residents (30% increase). These findings remain consistent across the proposed learner and considered baselines.

2. Related Work

Behavioral Routine and Activities:

We break down this stream based on the type of data.

Smartphone Data Researchers, primarily from the CS community developed several machine learning techniques to recognize low-level individual activities (e.g., sitting, standing, or walking) and high-level activities, often referred to as *lifestyles* or *routines*, (e.g., eating at a restaurant, taking a subway) from various types of sensor data collected from smartphones. While some of these methods are supervised [22], due to the practical limitation of acquiring labeled data for activities, a majority of the recent focus has shifted towards unsupervised methods [9, 39]. [9] apply LDA and ATM on labeled cell tower data to automatically discover routines, including "being at work" or "going home from work". [39] propose a probabilistic generative model for learning individuals' latent behavior patterns based on unlabeled cell tower data. This sub-stream of literature limit their focus to a subset of an individual's daily activities (such as work or shopping patterns) and do not analyze potential long-term health signals from the identified representations.

Surveys and Health Records Several other measures of routines have been developed in the medical literature via surveys or individual health records [5, 19]. These measures are based on smoking cessations, physical activity, diet quality, alcohol consumption and body weight. Healthy Eating Index-2015 (HEI-2015), computed based on individual surveys is a measure for assessing whether a set of foods aligns with the Dietary Guidelines for Americans (DGA). Alternatives to HEI with stronger correlations to chronic diseases was proposed by [5]. Acquiring longitudinal measures of such nature - for instance, via surveys, would require frequent interaction with individuals making them less practical than smartphone data based inference. Next, we discuss works that study associations and impact of behavioral routines, activities on future health events.

Behavior as Health Determinants: Individual behavior, measured as dietary, alcohol and tobacco consumption have been studied to determine health status of a population [19]. Impact of of lifestyle factors, determined by physical activity, high dietary score AHEI-2010 [5] on premature mortality was studied by [20]. Other factors such as health care resources [24], socio-economic factors [27] have been studied to impact health outcomes of a population. This stream of study primarily rely on data from surveys, EHR, medical claims differing from our work.

Prior studies have associated sensor measurements of sleep patterns, gait, activity rhythms, indoor activities and outings, and mobility with standard clinical measurements and survey data. Mobility metrics derived from location data have been used to describe the patterns of behavior and subjective experience associated with depressive symptoms [33], and mood patterns associated with schizophrenia symptoms [2]. [30] study relationship between location and transition patterns of an individual's indoor mobility behavior, namely the frequency, duration and times being carried out, with the driving and motor skill. Wearable sensor data was used to infer physical activity in patients with knee osteoarthritis [1]. [7] introduces the notion of an activity curve, which represents a visual abstraction of an individual's routines and develops a technique to detect changes in routines and perform health assessment. Our work complements this line of literature by identifying, associating and leveraging several individual social determinants to predict future health outcomes, from location data. To the best of our knowledge, we are not aware of other works that involve prediction of future health events from location data.

3. Framework

The primary objectives of our framework are two-fold. First is to identify an individual's social determinants of health from the location data: such as lifestyles, socioeconomic status, and accessibility to various resources. Second is to quantify the relationship between these determinants and individual health risk. We will introduce the relevant notations next.

Definition 1 (Trajectory) *A trajectory T_i of an individual i is defined as a temporally ordered set of tuples $T_i = \{(l_1^i, t_1^i), ..., (l_{n_i}^i, t_{n_i}^i)\}$, where $l_j^i = (x_j^i, y_j^i)$ is a location where x_j^i and y_j^i are the coordinates of the geographic location, and t_j^i is the corresponding time stamp.*

Definition 2 (Activity-Trajectory) *An activity trajectory D_i of an individual i is defined as mapping T_i to activities that exhibit a pattern of behavior. D_i is a temporally ordered set of tuples $D_i = \{d_1^i, , ..., d_{n_i}^i\}$, $d_j^i = (a_j^i, c_j^i)$, where $a_j^i = act(l_j^i), l_j^i \in T_i$ is an activity by the individual inferred from a location closest to x_j^i and y_j^i, and c_j^i is a coarser timestamp of t_j^i. Also, denote W as the universe of all temporal activities d_j^i across D_i.*

Definition 3 (Lifestyle) *A lifestyle L_i of an individual i is defined as a set of activities and their corresponding timestamps $L_i = \{d_1^i, d_2^i, ..., d_Y^i\}$, $d_j^i = (a_j^i, c_j^i) \in W$, $|L_i| = Y$ that globally represent an individual's day-to-day temporal activities across T_i.*

Next, we illustrate the transformation of individual trajectories (T_i) to activity trajectories (D_i) (Section 3.1) and detail the identification of lifestyles (L_i, Section 3.2). In Section 3.3, we discuss the remaining social determinants and our learner to quantify health risk. We present model-free analysis to understand if lifestyles signal future hospitalizations and discuss the performance of the proposed learner in predicting them in Section 5.

Activity group	Place type of location
hospital	hospital, doctor
health	physiotherapist, pharmacy, dentist, drugstore
ncessityshopping	store, supermarket, convenience_store, home_goods_store, grocery_or_supermarket, hardware_store
fitness	gym
publictransport	transit_station, train_station, bus_station, light_rail_station, subway_station
owntransport	car_wash, car_repair, parking, gas_station, taxi_stand
religious	church, mosque, hindu_temple, synagogue
recreation	amusement_park, tourist_attraction, zoo, park, theatre, sports_stadium, concert, bowling_alley, art_gallery, aquarium, museum, movie_rental, book_store, library, movie_theater, campground
travel	hotel, lodging, rv
personalcare	beauty_salon, spa, hair_care
leisureshopping	clothing_store, department_store, shopping_mall, shoe_store, electronics_store, furniture_store
unhealthyactivities	casino, liquor_store, bar, night_club, cigarette
restaurant	restaurant, food, meal, bakery, cafe, meal_delivery, meal_takeaway
home	highest dwell time location from 1 - 6 AM of an individual
work	per day, highest dwell non-home location.

Table 1 Activity groups

3.1. Locations to Activity Trajectories

Prior studies have used sensor data to study association of micro activities, such as daily sleep patterns, gait, and activity rhythms, with health [33, 30]. Our mapping of individuals' locations to POI categories, such as restaurants and groceries, opens up a new realm of possibilities to study both macro and micro patterns of an individual. For instance, macro movement and temporal patterns across competing brands inferred from such mapping were used to decide the placement of a new franchise. Further, micro, day-to-day individual-specific patterns such as number of visits, time spent at various business types can predict the individual's next likely location [25].

To identify individual lifestyles, we map the locations to POI categories by using Google Places API[1](second column of Table 1) and use the SafeGraph definitions of work to define *home*, *work*, *full-time*, and *part-time* work [2]. Next, we group POI categories with similar semantics (first column in Table 1) to form 15 activity groups that form the universe of all activities a^i_j. Further, to abstract away variations of the exact time in day-to-day activities, a coarser timestamp of t^i_j (timestamp associated with an individual's location), c^i_j is associated with each activity : 12 - 2 AM, 3 - 5 AM, 5 - 7 AM, 7 - 9 AM, 9 - 11 AM, 11 - 2 PM, 2 - 5 PM, 5 - 7 PM, 7 - 9 PM, 9 - 12 PM. The resulting tuples of $d^i_j = (a^i_j, c^i_j)$ across individual trajectories form the universe (W as defined in Def. 2) of temporal activities d^i_j.

3.2. Lifestyle Identification

Automatic discovery of individual lifestyles from location data is a non-trivial problem given the massive scale and high dimensionality. Besides, the differences in an individual's activities across days, and the differences from other individuals' activities add further complexity. We take an

[1] POIs can be readily identified by matching the longitudes/latitudes using Google Places API `https://developers.google.com/places/web-service/supported_types`

[2] Social Distancing Metrics Schema by SafeGraph `https://docs.safegraph.com/docs/social-distancing-metrics`

unsupervised topic modeling approach that has shown potential for uncovering complex temporal and behavioral patterns to identify *work*, *home*, and *consumption* routines [36, 9] on smaller location data sets. Specifically, we leverage the concept of probabilistic Author Topic Model (ATM), designed for text documents [31] to model an individual's day-to-day activities. Leveraging the granular location data, we extend this line of literature by incorporating an extensive set of 15 POI or activity types to represent an individual's lifestyle.

3.2.1. Author Topic Model: LDA is a probabilistic, unsupervised learning model of a bag of words and of hidden discrete variables called topics. For text modelling, we may view each document as a mixture of various topics, where each topic is characterized as a distribution over words. ATM [31] subsumes LDA and assumes authors of documents represent a multinomial distribution over topics where each topic is a probability distribution over words. A document with multiple authors has a distribution over topics that is a mixture of the distributions associated with the authors. When generating a document, an author is chosen at random for an individual word in the document. This author picks a topic from their multinomial distribution over topics and then samples a word from the multinomial distribution over words associated with that topic. This process is repeated for all words in the document. Formally, the probability of a word w_t assuming K topics, A authors, D documents and W unique words is: $P(w_t) = \sum_{k=1}^{K} P(w_t|z_t = k)P(z_t = k)$ where z_t is a latent variable showing the topic from which the t^{th} word is drawn. The aim of ATM inference is to determine the word distribution $P(w|z = k) = \phi_w^{(k)}$ for each topic k and the distribution of topics for authors $P(a = k) = \theta_k^{(a)}$ for each author a. $P(\theta)$ is a Dirichlet(α) and $P(\phi)$ is a Dirichlet(β), where α and β are hyper-parameters. Gibbs approximation proposed in [31] can be used to estimate these as

$$\phi_k^{(w)} = \frac{n_k^{(w)} + \beta}{n_k^{(.)} + W\beta}; \theta_k^{(a)} = \frac{n_k^{(a)} + \alpha}{n_.^{(a)} + K\alpha} \tag{1}$$

where $n_k^{(w)}$ and $n_k^{(a)}$ are the number of times word w and author a have been assigned to topic k, respectively. Similarly, $n_k^{(.)} = \sum_{1:W} n_k^{(w)}$, $n_.^{(a)} = \sum_{1:K} n_k^{(a)}$ are the word-topic and author-topic sum, respectively. Next, we detail our ATM-based lifestyle identification from the individual activity trajectories (D_i).

3.2.2. Activity trajectories to Lifestyles: To identify lifestyles, we make an analogy between text documents and day-to-day activities, authors, and individuals. We view each activity d_i^j in D_i, the mapped activity trajectory as a word w. We represent each day's activities of an individual (author) as a bag of words – document d. We view activities across multiple days of an individual i as unique documents of an author a. Based on these, we estimate the two ATM model parameters $\phi_k^{(d_i^j)}$, $\theta_k^{(i)}$ using Eq. 6 which represents the probability of activity for each topic k, and

the probability of topics k for an individual i, respectively. Given these probability distributions, we can rank activities for each topic (i.e., lifestyle) discovered. We can also rank topics for individuals which we view as their primary lifestyles.

We represent each lifestyle as the top Y activities ranked by their relevance [34] – a convex combination of topic-specific probability of each activity (first term in Eq. 2) and lift (second term in Eq. 2, p_{d_j} is the empirical distribution of activity d_j).

$$r(d_j) = \lambda log(\phi_k^{d_j}) + (1-\lambda) log \frac{\phi_k^{d_j}}{p_{d_j}} \tag{2}$$

Next, we assign the most probable topic from the estimated author-topic distribution θ_k^i as the primary lifestyle of an individual. Combining this with the top Y activities ranked by relevance, we can represent an individual i's lifestyle as $L_i = \{(d_1^i, d_2^i, ..., d_Y^i)\}$, $d_j \in W$. This completes the identification of the individual's lifestyle L_i from T_i. We augment these macro representations with other facets of social determinants extracted from location data that capture the micro day-to-day activities, accessibility to various resources, and socio-economics of an individual's neighborhood.

3.3. Other Social Determinants

In Table 2, we describe different facets of individual social determinants extracted from the location data and the proxies used to indicate an individual's health outcome - hospitalization. To construct these, we glean through the literature on the prediction of health outcomes from medical claims [11] or EHR data [16] across disciplines and make necessary adaptations to compute them from the individual location data. These features also form the input and output of our prediction model detailed later.

1) **Lifestyles**: We identify individual weekday and weekend lifestyles from their respective activity trajectories using the above ATM.

2) **Activity**: While lifestyles capture an individual's global routines, the behaviorally rich location data also enable us to capture the day-to-day micro activities. We leverage the transformed activity trajectories D_i (as defined in Def. 2) to compute an individual's daily visit frequencies and dwell time for each of the 15 activity groups a_j^i as additional dynamic, numerical individual features.

3) **Mobility**: Mobility metrics have been shown associated with health outcomes [33]. They capture an individual's daily mobility patterns based on the locations visited in T_i, such as the individual's frequency to, time spent at [28], and distance traveled to a location [38]. We also compute other richer mobility metrics, such as entropy and radius of gyration [14]. All these are daily, dynamic, numerical, individual level features.

4) **Accessibility**: Recent studies leveraging medical data also reveal the importance of neighborhood social demographics in predicting patient re-admission and length of stay [16]. We hence

leverage the transformed activity trajectory (D_i) and compute individual *accessibility* - the closest distance to various resources, such as hospitals, parks, fitness centers, pharmacies, public transport, and work from individual's *home* location. All these are static (time-invariant), numerical, individual level features.

5) **Socio-economics**: Based on the individual's *home* location from the transformed activity trajectories and publicly available Census data[3], we also compute several census block level socio-economic factors as in [16]. These are static and comprise of both categorical (*employment_type* - *part-time/full-time/nowork*) and numerical features (*population* of individual's census block).

6) **Hospitalization**: To identify if an individual has a future hospitalization, we overlay the day-to-day location trajectories on the publicly available location repositories of medical facilities. Specifically, we use the public data sets of hospitals, emergency medical services, and urgent care facilities from Homeland Infrastructure Foundation Level Data (HIFLD)[4]. Based on the overlaid data of medical facilities, we construct proxies to indicate the occurrence of individual's hospitalization event. Specifically,we assign an individual's $hospitalization = 1$ in an observation period, if the individual, whose *work* location is not at a medical facility, has at least 4 hours of activity at a medical facility – two of which occur during late night (12 AM - 5 AM) and the other two during 5 AM - 12 AM. We further assign $hospitalization\_night = 1$ if an individual has spent at least two late night hours at a medical facility (12 AM - 5 AM).

3.4. Health Risk Quantification

Our quantification of an individual's healthcare risk hinges on learning a model from the location data to accurately predict the future health outcome, hospitalization in our empirical study. We perform both model-free and Logit Regression analyses; and find consistent, qualitative (Figures 6b, 6d) and quantitative (Table 8) evidence that different lifestyles leads to heterogeneous rates of hospitalization.

3.4.1. Modelling Hospitalization Multiple types of (dynamic, static, categorical, numerical) individual features (Table 2) can capture multi-faceted social determinants, but also entail modeling challenges, such as the need to jointly represent all feature types, account for feature interactions, and concatenate features strategically to circumvent a sub-optimal predictive model. We address these challenges by separately learning the representations of the dynamic and static features that account for the interactions among different types of features. Next, we combine these, allowing for the interactions among the two representations, to learn a final joint representation of

[3] We obtain the Census Block Group (CBG) level data from SafeGraph: `https://docs.safegraph.com/docs/open-census-data#section-censusdemographic-data`.

[4] The latitudes and longitudes of hospitals, emergency medical services, and urgent care reported by state and federal resources are available at `https://hifld-geoplatform.opendata.arcgis.com/datasets/`

all the features. To achieve this, we represent the dynamic features by a Context-LSTM (CLSTM) cell proposed by [13], a modification of the traditional LSTM cell, widely used for word translation and time series modelling. CLSTM incorporates both dynamic and static contextual features to a time series. In [13], the dynamic contextual features are the latent topics that are jointly represented with words; and each word of the time series is concatenated with an embedding of the topic to predict the next likely word in a sentence. Extending this to our setting, lifestyles (**Lifestyle** features in Table 2) are latent topics learned from different activities and serve as a context to the activity-related dynamic features (**Activity** in Table 2). Viewing lifestyles as a context to the other dynamic features (**Mobility** in Table 2) also leads to better predictions[5]. Next, we concatenate these representations for a given time period with the embeddings of the static categorical and numerical features (**Social Demographics** and **Accessibility**) to jointly learn the representation of all features to predict an individual's future hospitalization. Such concatenations of multiple views of an individual's features to form a unified representations are widely studied in multi-modal learning (cf. [29] for a review).

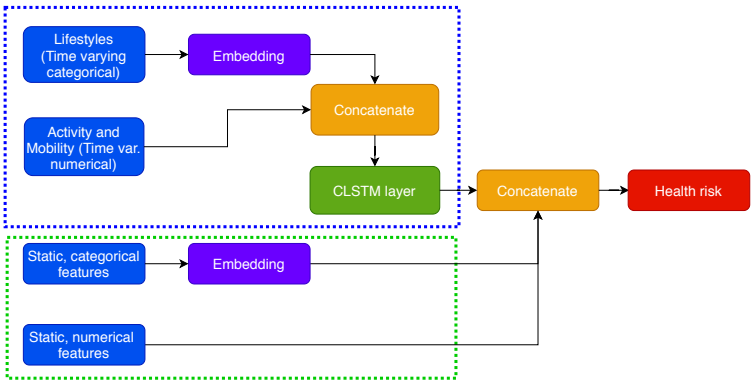


Figure 1 Architecture diagram of the proposed learner.

An overview of the architecture diagram of the proposed sequential deep learning model is presented in Figure 1. The blue box in the figure illustrates the modelling of an individual's temporal features at a day-level with a CLSTM cell (multiple days as a CLSTM layer), where the lifestyle serves as a context for the activity and mobility features. The green box shows the representations of the individual's static features, which are later concatenated with the temporal representations to predict the individual's hospitalization. Next, we formally detail the transformations performed by various layers (non-blue boxes in Figure 1) in the proposed learner.

[5] This is not surprising since an individual's lifestyle is likely correlated with his/her daily mobility behavior and hence a better predictor of his/her health outcome when we explicitly factor in both the lifestyle and mobility behavior.

Feature grouping	Name	Definition	Time Varying	Baltimore				D.C.			
				Mean	Std. Dev.	Min	Max	Mean	Std. Dev.	Min	Max
Lifestyle	lifestyle	Weekend and weekday lifestyle	✓	Refer Figure 4, 5				Refer Figure 8, 9			
Activity	home_freq	Daily frequency & dwell time at individual home	✓	5.99	17.1	1	249	5.93	18.0	1	237
	home_dwell		✓	1.68	2.80	0	24	1.55	2.82	0	24
	health_freq	Daily frequency & dwell time at health activity	✓	1.42	7.14	0	183	1.07	6.06	0	232
	health_dwell		✓	0.43	1.42	0	13.49	0.33	1.13	0	12.17
	necessityshopping_freq	Daily frequency & dwell time at necessity shopping	✓	1.72	8.07	0	216	1.46	7.29	0	215
	necessityshopping_dwell		✓	0.59	1.58	0	7.79	0.43	1.30	0	6.83
	publictransport_freq	Daily frequency & dwell time at public transport	✓	1.26	6.31	0	238	2.35	10.2	0	229
	publictransport_dwell		✓	0.48	1.34	0	6.85	0.71	1.80	0	6.89
	religious_freq	Daily frequency & dwell time at religious places	✓	0.73	4.82	0	162	0.591	4.39	0	198
	religious_dwell		✓	0.26	1.07	0	5.53	0.19	0.875	0	4.39
	work_freq	Daily frequency & dwell time at work	✓	3.21	8.67	0	233	4.20	10.3	0	192
	work_dwell		✓	1.32	2.36	0	24	1.26	2.64	0	24
	hospital_freq	Daily frequency & dwell time at hospitals	✓	0.13	2.20	0	164	0.08	2.13	0	156
	hospital_dwell		✓	0.04	0.43	0	24	0.02	0.24	0	24
	personalcare_freq	Daily frequency & dwell time at personal care	✓	0.30	3.47	0	208	0.281	2.97	0	161
	personalcare_dwell		✓	0.09	0.59	0	2.55	0.08	0.55	0	1.96
	restaurant_freq	Daily frequency & dwell time at restaurants	✓	0.78	4.22	0	145	1.32	5.90	0	177
	restaurant_dwell		✓	0.27	0.83	0	3.92	0.41	1.07	0	4.21
	unhealthyactivites_freq	Daily frequency & dwell time at unhealthy activities	✓	0.16	2.21	0	140	0.15	0.87	0	162
	unhealthyactivities_dwell		✓	0.04	0.39	0	4.32	0.02	0.22	0	12.6
	leisureshopping_freq	Daily frequency & dwell time at leisure shopping	✓	0.26	2.45	0	149	0.28	2.40	0	173
	leisureshopping_dwell		✓	0.10	0.59	0	4.91	0.09	0.54	0	5.88
	hotel_freq	Daily frequency & dwell time at hotels	✓	0.32	2.68	0	122	0.59	3.92	0	143
	hotel_dwell		✓	0.04	0.46	0	24	0.05	0.31	0	24
	owntransport_freq	Daily frequency & dwell time in own transport	✓	0.24	2.89	0	167	0.41	3.77	0	227
	owntransport_dwell		✓	0.12	0.63	0	24	0.11	0.60	0	24
Mobility	n_locations	Number locations in a day	✓	22.2	40.4	3	1585	23.1	42.7	3	1807
	avg_distance	Average distance traveled in a day (km.)	✓	7.42	6.70	0	126	7.59	7.54	0	170
	avg_location_entropy	Shannon entropy of frequency of visits	✓	1.90	1.11	0	1	1.84	1.16	0	5.87
	avg_time_entropy	Shannon entropy of dwell time at locations	✓	1.68	1.17	0	5.46	1.58	1.18	0	5.47
	n_unique_locations	Number of unique locations in a day	✓	7.6	15.7	1	573	8.94	18.6	1	417
	avg_time_spent	Average time spent at locations (in hours)	✓	4.20	3.64	0	24	4.08	3.61	0.06	24
	avg_rog	Average radius of gyration from home (in km.)	✓	6.11	4.28	0	125.1	6.21	4.71	0	132.1
	avg_speed	Average speed during the day (kmph)	✓	6.92	10.57	0	129	6.51	11.80	0	134
Accessibility	hospital_access	Distance from home to closest hospital (km.)	✗	1.63	0.84	0.02	2.62	1.58	0.89	0.21	4.1
	park_access	Distance from home to closest park	✗	0.40	0.29	0.04	2.67	1.24	0.21	0.03	4.62
	fitness_access	Distance to closest fitness facility	✗	0.55	0.37	0.02	2.04	0.76	0.42	0.03	2.91
	prescription_access	Distance to closest pharmacy	✗	0.45	0.28	0.02	2.05	0.61	1.25	0.02	2.69
	commute_access	Distance to closest commute	✗	0.15	0.13	0.02	3.05	0.23	0.18	0.02	2.92
	work_access	Distance from home to work	✗	1.90	3.34	0	39.4	1.86	3.37	0	41.1
Social Demographics	employment_type	Employment type of individual	✗	-	-	-	-	-	-	-	-
	employment_percent	Employment % in individual's census block group (cbg)	✗	0.82	0.09	0.44	1	0.85	0.09	0.59	1
	health_ins_percent	Health insurance % in individual's cbg	✗	0.99	0.04	0	1	0.98	0.04	0.07	1
	population	Population in individual's cbg	✗	1145	561	3	4696	1551	862	8	5254
	household_income	Average household income in cbg	✗	58321	31951	8654	250000	94193	50226	10278	250000
	median_age	Median age in individual's cbg	✗	37.4	9.30	10.8	79.9	35.7	7.53	18.9	73.8
	gross_rent	Gross rent in individual's cbg	✗	240	241	0	1384	395	246	0	2082
Health Outcome	hospitalization	An indicator if an individual spent 4 hours in a day at a medical facility (2 during 5AM - 12 AM, 2 during 12 AM - 5 AM)	✗	0.024	0.15	0	1	0.026	0.16	0	1
	hospitalization_night	An indicator if an individual spent two late night hours (12 AM - 5 AM)	✗	0.028	0.16	0	1	0.029	0.17	0	1

Table 2 Definition and Summary Statistics of Social Determinants and Health Events

3.4.2. Proposed Learner Let X_{TN} denote the dynamic numerical feature tensor (number of users $\times$ number of days in the observation period $\times$ number of dynamic numerical features), X_{TC} the dynamic categorical feature tensor (number of users $\times$ number of weeks[6] $\times$ number of dynamic categorical features), matrices X_{SN} and X_{SC} (number of users $\times$ number of categorical/numerical features) the static numerical, categorical individual features, respectively. To simplify the notation, in the following discussion, we will focus on a single individual's features denoted by $\mathbf{x}_{TC}$, $\mathbf{x}_{TN}$, $\mathbf{x}_{SC}$, and $\mathbf{x}_{SC}$ and their transformation to the probability of future hospitalization (i.e., the health risk).

1) **Embedding**: Embedding layers transforms one-hot encoded categorical features ($\mathbf{x}_{TC}$, $\mathbf{x}_{SC}$) to a continuous vector representation of a fixed dimension. Formally,

$$\mathbf{e}_{TC} = \mathbf{x}_{TC} W^e_{TC}; \mathbf{e}_{SC} = \mathbf{x}_{TC} W^e_{SC} \tag{3}$$

[6] Lifestyles are the only dynamic categorical features (weekday/weekend). Both dynamic and static categorical features are encoded using a one-hot encoding scheme.

where W^e_{TC} – number of dynamic categorical features $\times$ N^e_{TC}, W^e_{SC} – number of static categorical features $\times$ N^e_{SC} are the learnable weight parameters, N^e_{TC} N^e_{SC} are tune-able model hyper-parameters. Recall that in our setting, X_{TC} comprises of weekday and weekend lifestyles (L_i), both represented by top ten relevant activities (d^i_j, universe of activities W). Hence, an individual's weekday and weekend lifestyle can both be represented as a vectors of length $|W|$; that is, we learn two weight matrices of dimensionality $|W| \times N^e_{TC}$ to compute $\mathbf{e}_{TC}$. A similar procedure is followed to transform the other static categorical features (*employment_type*).

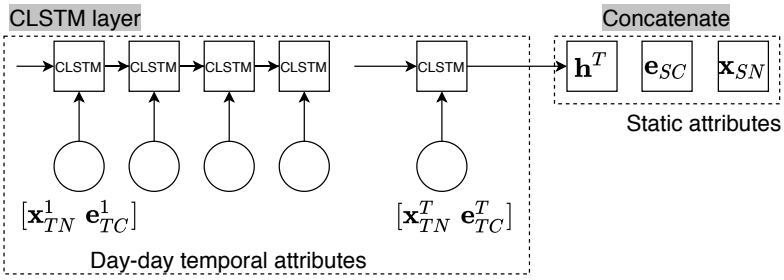


Figure 2 Illustrations of CLSTM and Concatenate layers.

2) **CLSTM layer**: CLSTM layer (illustrated in Figure 2) comprises of multiple CLSTM cells, each of which acts on different days of the $\mathbf{x}_{TN}$,$\mathbf{e}_{TC}$. Assume $\mathbf{x}^t_{TN}$,$\mathbf{e}^t_{TC}$ correspond to all numerical, embedded categorical dynamic features (contexts as defined in [13]) on an arbitrary day[7], each CLSTM cell performs the following transformations:

$$\begin{aligned} i^t &= \sigma(W_{iTC}\mathbf{e}^t_{TC} + W_{iTN}\mathbf{x}^t_{TN} + W_{ih}h^{t-1} + b_i) \\ f^t &= \sigma(W_{fTC}\mathbf{e}^t_{TC} + W_{fTN}\mathbf{x}^t_{TN} + W_{fh}h^{t-1} + b_f) \\ o^t &= \sigma(W_{oTC}\mathbf{e}^t_{TC} + W_{oTN}\mathbf{x}^t_{TN} + W_{oh}h^{t-1} + b_o) \\ c^t &= f^t * c^{t-1} + i^t * tanh(W_{cTC}\mathbf{e}^t_{TC} + W_{cTN}\mathbf{x}^t_{TN} + W_{ch}h^{t-1} + b_c) \\ h^t &= o^t * tanh(c^t) \end{aligned} \tag{4}$$

The above four equations detail modifications of the traditional LSTM cell where i,f and o are the input, output, and forget gates, respectively, to incorporate additional context $\mathbf{e}^t_{TC}$. After rearranging the terms, this is equivalent to considering a composite input $[\mathbf{x}^t_{TN}\ \mathbf{e}^t_{TC}]$. Each CLSTM cell transforms the concatenated input $[\mathbf{e}^t_{TC}\ \mathbf{x}^t_{TN}]$ into a hidden representation h^t (dimensions : number of individuals $\times$ N^e_T) with learnable shared[8] weight and bias parameters (W_*, b_*) and tune-able hyper-parameter N^e_T. Hence, the resulting representations from the CLSTM layer are

[7] $\mathbf{e}^t_{TC}$ is computed depending on whether the day is a weekday or weekend, since our lifestyles are derived for weekday/weekend rather than days.

[8] All the learnable weights W_* and bias parameters b_* are shared across different time steps (days in our model).

$\{h^t\}, t \in [1, T]$, where T is the number of days in our observation period.

3) **Concatenate**: Concatenate layers do not contain any learnable parameters and are simply used to combine different intermediate representations. We perform two concatenations, $[\mathbf{x}_{TN}^t \, \mathbf{e}_{TC}^t]$ as illustrated in Figure 2). Second, the concatenation of the hidden temporal representation obtained from the CLSTM layer ($\{h^t\}$), embedded static ($\mathbf{e}_{SC}$) and numerical features ($\mathbf{x}_{SN}$). Noting that $\mathbf{h}^T$, the hidden layer representation of the last day of observation captures temporal relations across the preceding days due to the recurrence nature of Equations 4, we combine this with $\mathbf{e}_{SC}$, $\mathbf{x}_{SN}$ to form $[\mathbf{h}^T \, \mathbf{e}_{SC} \, \mathbf{x}_{SN}]$, the final joint representation which comprises of both the dynamic and static features.

4) **Health risk**: We pass on the final representation into a fully connected dense layer, allowing for interactions between the temporal and static features, and assign the probability of hospitalization as

$$r = \sigma([W_T \, W_{SC} \, W_{SN} \, 1][\mathbf{h}^T \, \mathbf{e}_{SC} \, \mathbf{x}_{SN} \, b_r]^T) \tag{5}$$

where W_T, W_{SC}, W_{SN}, b_r are learnable parameters. For a given binary health outcome (hospitalization), to learn the various weights (W_* in Equations 3, 4, 5), we minimize the binary cross-entropy loss between the observed health outcome and $\mathbf{r}$, the vector of outcome probabilities. The rest of the hyper-parameters are tuned via cross-validation (details in Section 5).

4. Data

We combine several data sets: individual-level smartphone location data, census-block-level demographic data from the American Community Survey (2016), and HILFD public data of hospitals, emergency medical services, and urgent care facilities. The location data are curated with privacy compliance by a leading data collector via hundreds of commonly used mobile apps. The data cover one-quarter of the U.S. population across Android and iOS operating systems. Each data record corresponds to a location tracked with information about 1) Individual ID: an anonymized unique identifier of an individual's device, 2) Latitude, longitude and timestamp of a location visited; 3) Speed at which a visit was captured.

We analyzed data samples from Baltimore and D.C. (Baltimore – October, November 2018 and 2019; D.C. – April, May 2018 and October, November 2019). For each city, we only analyze the individuals who appear across all four months and at least ten days per month. We also eliminate those without reliable identification of *work* and *home* locations. The final sample comprises of 4,528 from Baltimore and 6,114 individuals from D.C. Tables 2, 3 and Figure 3 display the summary statistics of the social determinants of health computed from the location data, census block demographics, and public medical facilities.

Description	Baltimore				D.C.			
	Weekday		Weekend		Weekday		Weekend	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Daily activities	14.64	17.12	13.81	19.64	14.70	18.48	15.86	19.66
Unique daily activities	10.71	10.42	9.62	11.12	10.81	11.38	11.34	11.70
Activities at *home*	5.15	14.3	8.89	22.2	4.86	14.5	8.32	21.9
Activities at *work*	4.37	9.43	0.79	4.65	4.89	11.2	0.95	5.85
Activities at *publictransport*	2.25	6.03	2.30	7.20	2.29	9.94	2.57	10.9
Activities at *other*	4.86	15.4	5.13	17.2	4.73	15.2	4.99	16.4

Table 3 Summary statistics of the activity trajectories

Location Trajectories: Table 2 (**Mobility** row) shows the summary statistics of the raw location data. In Baltimore, there are on average 22 total locations (and 7 unique locations) per individual per day. The average speed is 6.92 kmph. For all other measures, we eliminate the locations captured at a speed > 5 kmph and dwell time < 5 minutes[9]. The average Haversine Distance between consecutive locations is 7.42 km and the average dwell time 2.2 hours.

Activity Trajectories : Table 2 (**Activity** and **Accessibility** rows), Table 3 detail the summary statistics of the activity trajectories (Section 3.1 for the transformation of the locations to activity trajectories). Table 2 shows that out of the 15 activity groups (Table 1), *home*, *work* and *publictransport* are the top three in both the average daily occurrences and time spent. When broken down by weekday and weekend (Table 3, Baltimore), *work* occurs less frequently (0.79) during the weekend compared to weekdays (4.37). In contrast, *home* occurs more frequently during weekends (5.15) than weekdays (8.89). To accommodate the differences[10] in these top activities, we learn weekday and weekend lifestyles separately. There are on average 14.64 total daily activities (9.62 unique) per individual on weekdays, characterized by its activity group and time range (*restaurant (2 - 5 PM)*); and 14 (10.71 unique) on weekends.

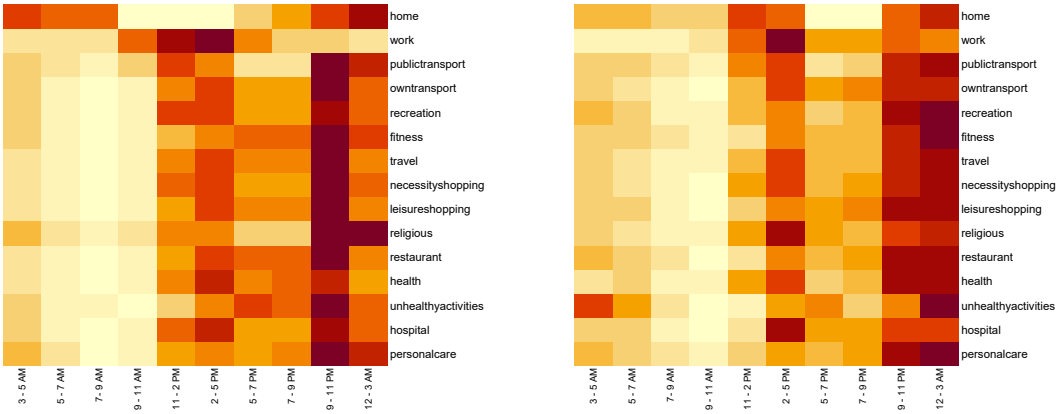
Figure 3 plots the heat map of the activities on weekends and weekdays, with a lighter color indicating a lower occurrence of an activity during the corresponding time. Figure 3a shows that *work* mostly occurs during 2 - 5 P.M., *home* 12 - 3 AM. In contrast, on weekends people tend to stay *home* during the same time window (Figure 3b). Also, leisure, shopping, and consumption activities, occur earlier (9 - 11 P.M.) on weekdays than weekends (12 - 3 A.M.).

Census Block Socio-economics: An individual's census block is assigned based on the closest census block by Haversine Distance to his/her *home*. Table 2 (**Social Demographics**) exhibits the summary statistics of the census block socio-economic factors.

Health Outcome: An individual's health outcome is defined as a hospitalization event observed in the location data over the last two months of the sampling period. A total of 111 (158) individuals had hospitalizations spanning both day and night and 127 (175) during night at Baltimore (D.C.).

[9] The time difference between consecutive locations is used to determine the dwell time spent at a location.

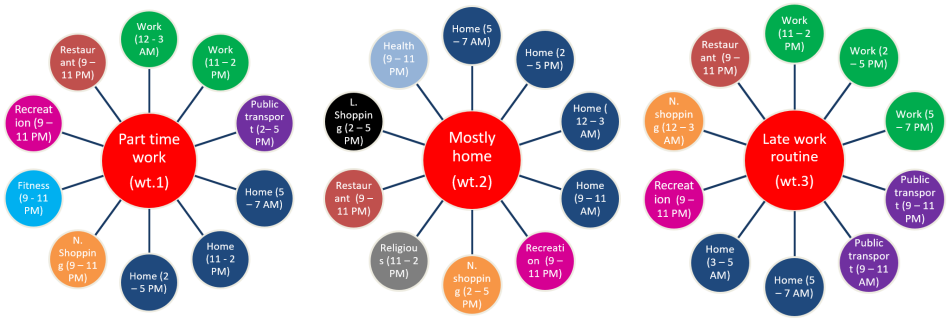
[10] We confirmed that these differences between weekends and weekdays for *work* and *home* are statistically significant at $p = 0.01$ based on a paired Wilcoxon test.



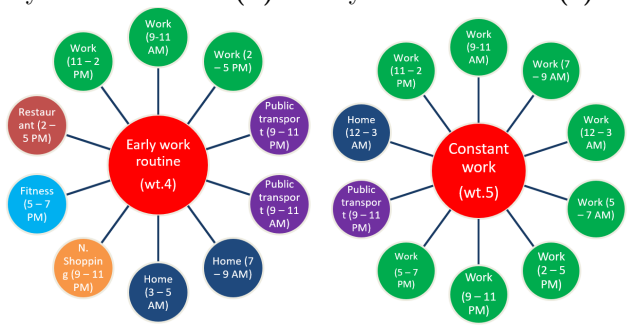
(a) Weekday **(b)** Weekend

Figure 3 **Row-normalized heatmap of activity occurrences (Baltimore). Darker reds indicate higher occurrences.**

5. Empirical Study



(a) Lifestyle 1 **(b)** Lifestyle 2 **(c)** Lifestyle 3



(d) Lifestyle 4 **(e)** Lifestyle 5

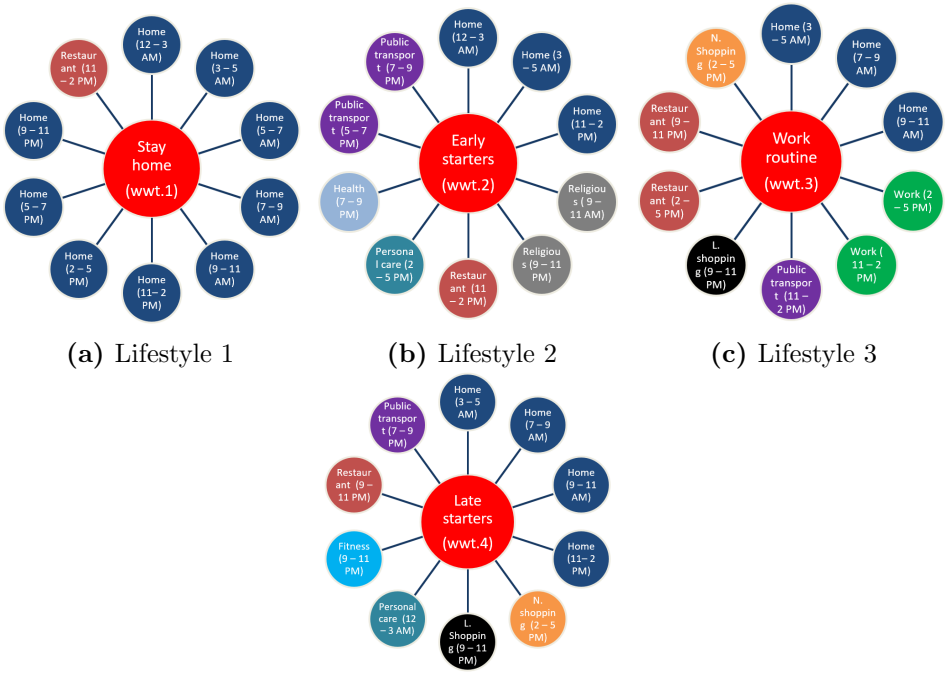
Figure 4 **Weekday Lifestyles (Baltimore)**

5.1. Lifestyles

We apply the proposed ATM-based methodology on the first two months of the location data during the sampling period to identify the lifestyles. Figures 4 and 5 present the top 10 activities representative of weekend and weekday lifestyles. The number of lifestyles (topics, K) are decided based on coherence [34]. Coherence measures how well-focused the top words (activities) describe

a specific lifestyle. We vary K between 3 to 25, and compute the average coherence over 50 runs to determine the number of topics for weekdays and weekends, respectively. The top 10 ($Y = 10$, Def. 3) relevant activities for each topic are then visualized for the highest coherent ATM model in figures 4 and 5. In total, we identify five weekday and four weekend lifestyles with different activities across different hours-of-the-day.

Weekday Lifestyles: Figure 4 visualizes the five identified weekday lifestyles and their corresponding activities for Baltimore residents. Lifestyle 3 (denoted by *wt.3*) characterizes a late work routine (*work* over 11 - 2 PM, 2 - 5 PM, 5 - 7 PM), commute via public transportation mornings and evenings (*publictransport* over 9 - 11 AM , 9 - 11 PM), late night dining at restaurants, grocery shopping, and recreation (*necessityshopping* 12 - 3 AM, *restaurant* 9 - 11 PM, *recreation* in 9 - 11 PM). In contrast, lifestyle 4 (*wt.4*, although with similar commute and consumption patterns, reveals an early work routine: *work* 9 - 11 AM, 11 - 2 PM, 2 - 5 PM, and fitness during evenings (5 - 7 PM). Both lifestyles feature a steady full-time work routine, and work-fitness balance. Lifestyle *wt.1*, in comparison, indicates a part-time job (*work* 11 - 2 PM, 12 - 3 AM).



(a) Lifestyle 1 **(b)** Lifestyle 2 **(c)** Lifestyle 3

(d) Lifestyle 4

Figure 5 Weekend Lifestyles (Baltimore)

Weekend Lifestyles: Figure 5 displays the top ten activities of the four weekend lifestyles for Baltimore residents. As expected, apart from lifestyle *wwt.3* (*work* 11 - 2 PM, 2 - 5 PM), all other lifestyles suggest a non-work routine. Lifestyle *wwt.1* characterizes an early start weekend routine

Weekday topic	Mean income	Mean Age	Mean population	Mean employment rate	Mean access to hospitals (km.)	Mean fitness activities per day (# of days/weekday)	Mean unhealthy activities per day (# of days/ weekday)	Mean personal care activities per day (# of days/weekday)
wt.1	62061	37.28	1146	0.83	0.82	0.81 (1.11)	0.14 (0.16)	0.37 (0.14)
wt.2	55058	37.39	1167	0.81	2.13	0.52 (0.62)	0.27 (0.25)	0.32 (0.12)
wt.3	57771	37.19	1107	0.82	1.62	0.41 (0.56)	0.13 (0.14)	0.28 (0.09)
wt.4	56904	37.86	1117	0.82	1.87	0.68 (0.92)	0.11 (0.13)	0.41 (0.16)
wt.5	59842	37.75	1178	0.85	1.37	0.37 (0.49)	0.24 (0.22)	0.27 (0.09)

Table 4 Weekday Lifestyle Demographics (Baltimore)

Weekend topic	Mean income	Mean Age	Mean population	Mean employment rate	Mean access to hospitals (km.)	Mean fitness activities per day (days/ weekend)	Mean unhealthy activities per day (days/weekend)	Mean personalcare activities per day (days/weekend)
wwt.1	56779	37.93	1161	0.80	1.89	0.42 (0.52)	0.96 (0.84)	0.72 (0.45)
wwt.2	54338	37.48	1076	0.82	2.12	0.69 (0.76)	0.67 (0.65)	0.86 (0.65)
wwt.3	58609	36.77	1148	0.85	1.25	0.45 (0.41)	0.61 (0.63)	0.65 (0.54)
wwt.4	60771	37.43	1143	0.82	0.98	0.79 (0.98)	0.52 (0.54)	0.94 (0.74)

Table 5 Weekend Lifestyle Demographics (Baltimore)

with visits to religious locations (*religious* 9 - 11 AM, 9 - 11 PM) and *restaurant* afterwards (11 - 2 PM). In contrast, lifestyle *wwt.4* indicates a late start routine, where the individuals mainly stay at *home* during these hours, with fitness and recreations later in the evening (*fitness* 9 - 11 PM, *personalcare* 12 - 3 AM). Besides work on weekends, individuals in lifestyle *wwt.3* regularly consume at restaurants (*restaurant* 2 - 5 PM, 9 - 11 PM). Lifestyle *wwt.1* stays at home weekends, dine at *restaurant* 11 - 2 PM, with limited fitness or leisure activities. In D.C., we identify five weekday and four weekend lifestyles. These are discussed in Appendix 7. Overall, the proposed lifestyle identification uncovers distinctive activity patterns from location data.

5.2. Lifestyle Demographics

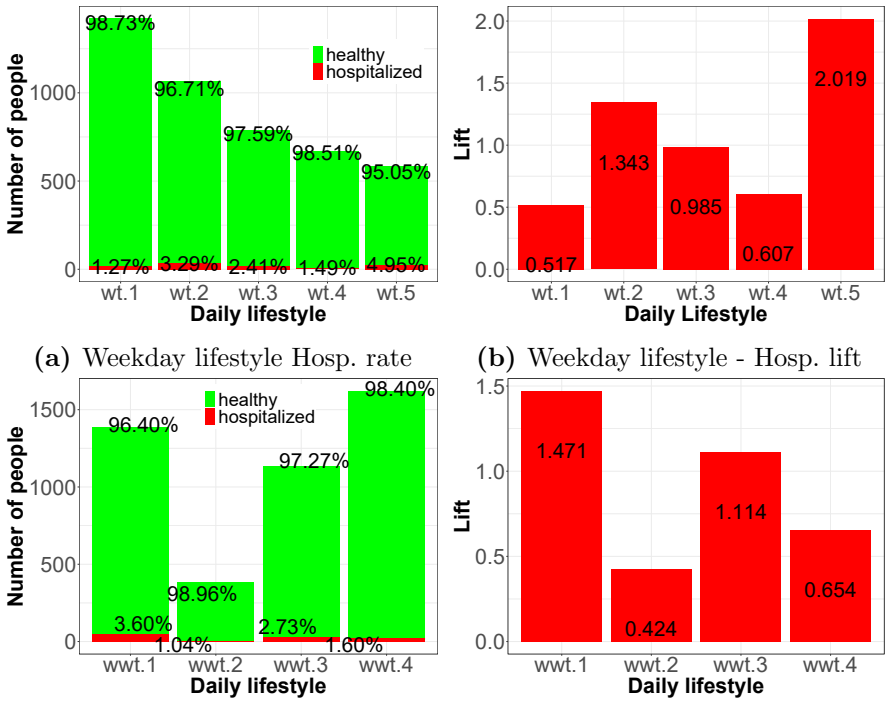
To delve deeper into the identified lifestyles, in Table 4, 5, we compute the average CBG level demographics - income, population, age, and employment rate. These demographics are further overlaid with the aggregate behavioral measures of fitness, personal care, and unhealthy activities of the individuals belonging to each life style. We display two behavioral measures - the mean number of activities per day, and the mean number of days (weekdays/weekends) there was at least one such activity across the observation period.

Several interesting findings emerge. Lower income population can present healthy lifestyles/outcomes (*wt.4* - high fitness, personal care activities and low unhealthy activities); while high-income population can present unhealthy ones (*wt.5* - high unhealthy activities, low fitness activities). Population with lower accessibility to healthcare or facilities can present healthy lifestyles/outcomes (*wwt.2*, low unhealthy activities, high fitness activities); while population with higher accessibility can present unhealthy ones (*wt.5* high unhealthy and low fitness activities).

5.3. Health Risk Quantification

We identify hospitalization from the two months of location data in 2019 and then link them to the social determinants.

Model-free Evidence: Figure 6 exhibits the histogram of the percentage of the 4,528 Baltimore residents with each lifestyle visiting medical facilities. Weekday lifestyles *wt.2* and *wt.5* have higher (3.29% and 4.95%, Figure 6a) than average (2.45%) percentages of individuals visiting medical facilities. In contrast, lifestyle *wt.4* has about half (1.49 %) the average percentage of hospitalizations. Similarly, Figure 6c reveals that weekend lifestyles *wwt.1* and *wwt.3* experience higher percentages of hospitalizations whereas lifestyle *wwt.2* half less likely.



(a) Weekday lifestyle Hosp. rate **(b)** Weekday lifestyle - Hosp. lift

(c) Weekend lifestyle Hosp. rate **(d)** Weekend lifestyle - Hosp. lift

Figure 6 (Baltimore) Association with Hospitalization : Model free analysis (*hospitalization*)

We quantify the relative rate of future hospitalizations per lifestyle against the average rate by using the lift score (Figures 6b, 6d). The top activities characterizing each lifestyle (Figures 4, 5) and their lift scores suggest that those who participate in *fitness* on weekends (*wwt.4*) or weekdays (*wt.4*) are less likely (0.65 and 0.60, respectively) to have hospitalizations on average. On the other extreme, those with either busy, volatile work routines (*wt.5*) or no work routine (*wt.2*) on weekdays, are 2.01 and 1.34 times more likely to have hospitalizations (Figures 6b, 4; and people who either work (*wwt.3*) or mostly stay at home (*wwt.1*) on weekends are 1.1 to 1.4 times more likely to have hospitalizations than average (Figures 6d, 5). Overall, the model-free evidence reveals heterogeneous rates of hospitalizations across different lifestyles.

Logit Analysis: To supplement the model-free evidence, we examine an individual's likelihood of having future hospitalization using a Logit model:

$$P(hospitalization\_i) = \frac{e^{X_i}}{1+e^{X_i}}$$
$$X_i = \alpha_i + \beta_1 lifestyle\_weekday_i + \beta_2 lifestyle\_weekday_i + \beta_3 X_i^{access} + \beta_4 X_i^{mobility} + \beta_5 X_i^{demog} + \beta_6 X_i^{community} \quad (6)$$

where $hospitalization\_i$ is 1 if an individual had at least one hospitalization during the two months in 2019, $lifestyle\_weekday_i$ and $lifestyle\_weekend_i$ are dummies indicating the individual's weekend and weekday lifestyles, X_i^{access}, $X_i^{mobility}$ are the average daily accessibility and mobility metrics, respectively, in Table 2; $X_i^{community}$ are the average mobility and accessibility metrics of the residents in the same census block group as the individual; and X_i^{demog} are the census block socio-economic factors. Table 6 (Columns 1 - 5) displays the maximum log-likelihood estimates of the lifestyles while controlling for different individual-level features. The coefficients indicate the odds of an individual with a specific lifestyle to have a future hospitalization over the average odds.

	Dep. variable : hospitalization				
	(1)	(2)	(3)	(4)	(5)
Weekend lifestyle 1 (wwt.1)	0.384** (0.174)	0.353** (0.176)	0.324** (0.182)	0.296* (0.197)	0.295* (0.198)
Weekend lifestyle 2 (wwt.2)	−0.804** (0.347)	−0.771** (0.348)	−0.805** (0.354)	−0.976** (0.402)	−0.982** (0.404)
Weekend lifestyle 4 (wwt.4)	−0.024 (0.182)	−0.010 (0.183)	0.114 (0.192)	0.213 (0.208)	0.214 (0.210)
Weekday lifestyle 1 (wt.1)	−0.685*** (0.189)	−0.691*** (0.190)	−0.665*** (0.200)	−0.654*** (0.209)	−0.648*** (0.211)
Weekday lifestyle 2 (wt.2)	0.121 (0.177)	0.076 (0.182)	−0.204 (0.175)	−0.138 (0.193)	−0.130 (0.195)
Weekday lifestyle 4 (wt.4)	−0.494** (0.241)	−0.463* (0.243)	−0.429* (0.250)	−0.395* (0.261)	−0.334* (0.262)
Weekday lifestyle 5 (wt.5)	0.998*** (0.146)	1.030*** (0.149)	0.933*** (0.155)	0.946*** (0.164)	0.928*** (0.167)
Accessibility metrics	✗	✓	✓	✓	✓
Mobility metrics	✗	✗	✓	✓	✓
Social Demographics	✗	✗	✗	✓	✓
Community Controls	✗	✗	✗	✗	✓
Observations	4,528	4,528	4,528	4,528	4,528
Log Likelihood	−665.904	−657.063	−586.244	−532.423	−527.367

*p<0.1; **p<0.05; ***p<0.01

Table 6 (Baltimore) Hospitalization Logit Analysis

Table 6 (Column 5) indicate that those with $wwt.4$, $wt.5$ have significantly higher odds of having a future hospitalization ($1.34 \approx \exp(0.296)$ and 2.57, respectively) than average, after controlling for other social determinants. Similarly, lifestyles $wwt.2$, $wt.1$ and $wt.4$ have significantly lower odds than average. These insights are qualitatively consistent with the model free evidence. Interestingly, we do not find any significant association between X_i^{access}, X_i^{demog} and future hospitalizations indicating that two individuals who live in the same neighborhood with similar social demographics, access to parks/fitness facilities, but with different lifestyles, will have different health risks. Also, we observe that an individual's community factors, $X_i^{community}$ do not have a significant correlation

to their future hospitalization. Overall, we find that an individual's day to day behavior, captured by the lifestyles have a dominant effect on their future health. These findings strongly align with a recent review article of social determinants of health in EHR and their impact on analysis and risk prediction by [4].

In Table 7, we introduce total dwell time at healthy (fitness, personal care) and unhealthy activities into the regression and observe that *regularity* of healthy activities matters (lifestyle $wt.1$, $wt.4$, $wwt.4$), instead of the total dwell time (e.g., two individuals with similar fitness/work hours per week, but different distribution of these activities across days may lead to different health risks) further highlighting the importance of mining the lifestyle patterns to quantify health risk. On the flip side, we also observe that total time spent at unhealthy activities is significantly correlated to future hospitalization. In D.C, the qualitative findings remain similar. In addition, we find that *regularity* of personal care activities associate with significantly lower odds (1.6) of future hospitalization.

	Dep. variable : hospitalization		
	(1)	(2)	(3)
Weekend lifestyle 1 (wwt.1)	0.295* (0.198)	0.295* (0.198)	0.295* (0.199)
Weekend lifestyle 2 (wwt.2)	−0.978** (0.404)	−1.010** (0.407)	−1.002** (0.406)
Weekend lifestyle 4 (wwt.4)	0.208 (0.210)	0.211 (0.210)	0.235 (0.211)
Weekday lifestyle 1 (wt.1)	−0.638*** (0.212)	−0.619*** (0.214)	−0.589*** (0.212)
Weekday lifestyle 2 (wt.2)	−0.126 (0.195)	−0.143 (0.196)	−0.143 (0.197)
Weekday lifestyle 4 (wt.4)	−0.334* (0.262)	−0.329* (0.262)	−0.339* (0.262)
Weekday lifestyle 5 (wt.5)	0.929*** (0.167)	0.915*** (0.167)	0.917*** (0.167)
total_fitness_dwell	−0.001 (0.009)		
total_personalcare_dwell		−0.001 (0.002)	
total_unhealthyactivities_dwell			0.003** (0.001)
Other social determinants	✓	✓	✓

*p<0.1; **p<0.05; ***p<0.01

Table 7 (Baltimore) Hospitalization : Additional Logit Analysis

Predictive Performance: As detailed in Sec 3.4, the proposed learner takes the individual features extracted from the location data in 2018 to predict the health risk (Eq 5), i.e., the probability of an individual having a hospitalization in 2019. In practice, given a series of risk scores, a domain expert would ideally set the minimum threshold to deem if an individual has surpasses an "at-risk" threshold. Hence, in Table 8, 9, we report the average cross-validated PRAUC and ROCAUC and corresponding standard deviation percentages to sweep all possible thresholds[11].

[11] We also include *hospitalization_alt*, an indicator of an individual spending 6 hours in a medical facility on any day in the 2 months in 2019, in addition to the 2 indicators in Table 2. The data are split into 70% training, 15% validation, and 15% test sets; and a ten-fold cross validation is performed. We perform a grid search to optimize several tune-able model hyper-parameters: dimensionality of the dynamic and static categorical embeddings, class weights, learning rate, number and size of various hidden layers.

City = Baltimore	**Day and Night Hospitalization** (*hospitalization*)		**Late Night Hospitalization** (*hospitalization_night*)		**6-hour hospital visit** (*hospitalization_alt*)	
Hospitalization rate	2.45%		2.80%		4.75%	
Model /Measure	PR AUC	AUC	PR AUC	AUC	PR AUC	AUC
RF (NLI & NAC)	0.05 (1.13 %)	0.63 (1.91 %)	0.06 (1.05 %)	0.61 (2.15%)	0.11 (1.76%)	0.65 (2.04%)
GB (NLI & NAC)	0.06 (1.14%)	0.64 (2.12%)	0.06 (1.32%)	0.63 (1.69%)	0.12 (1.10%)	0.65 (2.31%)
Lasso (ALLAGG)	0.14 (2.27 %)	0.73 (4.53%)	0.13 (2.05 %)	0.72 (4.70 %)	0.21 (2.02 %)	0.70 (4.58 %)
RF (ALAGG)	0.22 (2.54 %)	0.78 (4.06%)	0.21 (2.63%)	0.74 (4.90%)	0.29 (2.53%)	0.74 (4.69%)
GB (ALLAGG)	0.23 (2.47 %)	0.76 (4.06%)	0.21 (2.69%)	0.73 (4.64%)	0.27 (2.73%)	0.71 (4.44%)
LSTM (NL & NAC)	0.15 (1.97%)	0.72 (1.60%)	0.17 (1.39%)	0.74 (2.12%)	0.21 (1.22%)	0.72 (2.86%)
LSTM (NLI)	0.24 (1.05%)	0.79 (2.95%)	0.24 (1.26%)	0.76 (3.46%)	0.38 (1.71%)	0.80 (3.90%)
Full model	**0.28 (1.53%)**	**0.85 (3.67%)**	**0.29 (1.17%)**	**0.84 (3.95%)**	**0.42 (1.47%)**	**0.86 (3.67%)**

Table 8 (Baltimore) Hospitalization prediction

City = DC	**Day and Night Hospitalization** (*hospitalization*)		**Late Night Hospitalization** (*hospitalization_night*)		**6-hour Hospital visit** (*hospitalization_alt*)	
Hospitalization rate	2.58%		2.87%		5.71%	
Model/Measure	PR AUC	AUC	PR AUC	AUC	PR AUC	AUC
RF (NLI & NAC)	0.07 (1.01 %)	0.63 (2.11 %)	0.06 (0.96 %)	0.65 (1.85%)	0.12 (1.89%)	0.66 (2.11%)
GB (NLI & NAC)	0.06 (1.08%)	0.62 (2.04%)	0.07 (1.14%)	0.66 (1.96%)	0.11 (1.61%)	0.68 (2.41%)
Lasso (ALLAGG)	0.17 (2.75 %)	0.76 (4.13 %)	0.16 (2.84 %)	0.76 (4.96 %)	0.24 (2.85 %)	0.78 (4.97 %)
RF (ALLAGG)	0.23 (2.96 %)	0.80 (4.82%)	0.22 (2.69%)	0.80 (4.90%)	0.30 (2.57%)	0.81 (4.98%)
GB (ALLAGG)	0.24 (2.07 %)	0.79 (4.27%)	0.22 (2.27%)	0.81 (4.64%)	0.31 (2.63%)	0.82 (4.82%)
LSTM (NLI & NAC)	0.16 (1.24%)	0.74 (1.99%)	0.17 (1.91%)	0.75 (2.62%)	0.26 (1.75%)	0.77 (2.72%)
LSTM (NLI)	0.23 (1.97%)	0.81 (2.71%)	0.21 (2.01%)	0.80 (3.22%)	0.35 (1.75%)	0.81 (3.54%)
Full model	**0.30 (1.42%)**	**0.87 (3.12%)**	**0.32 (1.39%)**	**0.89 (3.41%)**	**0.44 (1.87%)**	**0.90 (3.80%)**

Table 9 (D.C.) Hospitalization Prediction

We compare our learner's predictive performance with several baselines' to investigate 1) importance of jointly representing multiple facets of an individual using a sequential model; 2) performance lift provided by individual behavioral features – lifestyles and day-to-day activities. To support 1), we employ non-sequential learners, Random Forest (RF), regularized logistic regression (LASSO), and Gradient Boosting (GB) with aggregated static, dynamic numerical and categorical features (ALLAGG) (Table 2)[12]. To support 2), we design ablations of the proposed learner and baselines without the dynamic lifestyles (NLI) and activity features (NAC).

Table 8 suggest that the proposed learner outperforms both types of baselines considered. The proposed model for health risk quantification has an AUPRC of 0.28 and AUC of 0.85. The best performing non-sequential model performs worse than the proposed learner (0.23 compared to 0.28), indicating the importance of modelling temporal correlations across features via LSTMs. Ablations of the non-sequential models and the proposed learner suggest lifestyles and day-to-day activities, in aggregate and dynamic form, provide a performance lift. The best performing non-sequential model (GB) has a PR AUC increment from 0.06 to 0.23; sequential models from 0.15 to

[12] The dynamic features for e.g. daily unique locations are aggregated across days as the average daily unique locations. The categorical features are encoded as one-hot dummies. Several parameters of the baselines are optimized by performing a grid search. We report the average ten-fold cross-validated PR AUC and AUC. SMOTE [3] is used to account for the class imbalance for the non-sequential baselines.

0.28. Finally, the ablation of the proposed learner without the CLSTM cell performs worse (0.24 PR AUC) than the full model (16 % increase in PR AUC), indicating the importance of lifestyles as the contexts to the dynamic features. These results remain qualitatively similar for the D.C. residents (Table 9, 30% increase from ablations).

6. Conclusion

We develop a framework to identify individual social determinants of health and quantify their impact on future hospitalizations from population scale, granular location data. Specifically, building on topic models, we first identify an individual's lifestyles; then supplement them with additional accessibility and socio-economic features. Through an array of analyses, we quantify the strong connection between the social determinants of health, particularly lifestyles, and future hospitalization by leveraging sequential deep learning models. This research broadens the prior literature by proposing a comprehensive framework that leverages cutting edge machine learning methods to explore novel, population-scale, behavior-rich big data with a panoramic view of consumer lifestyles beyond the traditional EHRs. It thus offers generalizable insights regarding the relative impact of various social determinants of health. It further offers critical guidance to policy making and marketing communication to advocate healthy lifestyles, healthy products and services, regularities in lifestyles, to mitigate the rocketing healthcare costs, and to promote public health of immense societal benefits.

References

[1] V. Agarwal, M. Smuck, C. Tomkins-Lane, and N. H. Shah. Inferring physical function from wearable activity monitors: analysis of free-living activity data from patients with knee osteoarthritis. *JMIR mHealth and uHealth*, 6(12), 2018.

[2] D. Ben-Zeev, C. J. Brenner, M. Begale, J. Duffecy, D. C. Mohr, and K. T. Mueser. Feasibility, acceptability, and preliminary efficacy of a smartphone intervention for schizophrenia. *Schizophrenia bulletin*, 40(6):1244–1253, 2014.

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[4] M. Chen, X. Tan, and R. Padman. Social determinants of health in electronic health records and their impact on analysis and risk prediction: A systematic review. *JAMIA*, 27(11):1764–1773, 2020.

[5] S. E. Chiuve, T. T. Fung, E. B. Rimm, F. B. Hu, M. L. McCullough, M. Wang, M. J. Stampfer, and W. C. Willett. Alternative dietary indices both strongly predict risk of chronic disease. *The Journal of nutrition*, 142(6):1009–1018, 2012.

[6] W. C. Cockerham, T. Abel, and G. Lüschen. Max weber, formal rationality, and health lifestyles. *Sociological Quarterly*, 34(3):413–425, 1993.

[7] P. N. Dawadi, D. J. Cook, and M. Schmitter-Edgecombe. Modeling patterns of activities using activity curves. *Pervasive and mobile computing*, 28:51–68, 2016.

[8] N. Eagle, A. S. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the national academy of sciences*, 106(36):15274–15278, 2009.

[9] K. Farrahi and D. Gatica-Perez. Discovering routines from large-scale human locations using probabilistic topic models. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(1):1–27, 2011.

[10] W. Freeman, A. Weiss, and K. Heslin. Overview of us hospital stays in 2016: Variation by geographic region. *Rockville, MD: Agency for Healthcare Research and Quality*, 2018.

[11] L. García-Olmos, R. Aguilar, D. Lora, M. Carmona, A. Alberquilla, R. García-Caballero, L. Sánchez-Gómez, and C. Group. Development of a predictive model of hospitalization in primary care patients with heart failure. *Plos one*, 14(8):e0221434, 2019.

[12] A. Ghose, B. Li, and S. Liu. Mobile targeting using customer trajectory patterns. *Management Science*, Forthcoming, 2018.

[13] S. Ghosh, O. Vinyals, B. Strope, S. Roy, T. Dean, and L. Heck. Contextual lstm (clstm) models for large scale nlp tasks. *arXiv preprint arXiv:1602.06291*, 2016.

[14] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *nature*, 453(7196):779, 2008.

[15] D. M. Greer, D. J. Baumgardner, F. D. Bridgewater, D. A. Frazer, C. L. Kessler, E. S. LeCounte, G. R. Swain, and R. A. Cisler. Milwaukee health report 2013: Health disparities in milwaukee by socioeconomic status. 2014.

[16] C. B. Hilton, A. Milinovich, C. Felix, N. Vakharia, T. Crone, C. Donovan, A. Proctor, and A. Nazha. Personalized predictions of patient outcomes during and after hospitalization using artificial intelligence. *NPJ Digital Medicine*, 3(1):1–8, 2020.

[17] T. Hu, R. Song, Y. Wang, X. Xie, and J. Luo. Mining shopping patterns for divergent urban regions by incorporating mobility data. In *Proceedings of the 25th ACM CIKM*, pages 569–578, 2016.

[18] ICSI. Going beyond clinical walls : Solving complex problem. 2004.

[19] I. Joumard, C. André, C. Nicq, and O. Chatal. Health status determinants: lifestyle, environment, health care resources and efficiency. *Environment, Health Care Resources and Efficiency*, (627), 2010.

[20] Y. Li, A. Pan, D. D. Wang, X. Liu, K. Dhana, O. H. Franco, S. Kaptoge, E. Di Angelantonio, M. Stampfer, W. C. Willett, et al. Impact of healthy lifestyle factors on life expectancies in the us population. *Circulation*, 138(4):345–355, 2018.

[21] G. Loewenstein, T. Brennan, and K. G. Volpp. Asymmetric paternalism to improve health behaviors. *Jama*, 298(20):2415–2417, 2007.

[22] B. Logan, J. Healey, M. Philipose, E. M. Tapia, and S. Intille. A long-term evaluation of sensing modalities for activity recognition. In *International conference on Ubiquitous computing*, pages 483–500. Springer, 2007.

[23] M. Macha, B. Li, N. Zhang Foutz, and A. Ghose. Perils of location tracking? personalized and interpretable privacy preservation in consumer mobile trajectories. *Working Paper, Carnegie Mellon University*, 2019.

[24] R. D. Miller and T. Frech. The productivity of health care and pharmaceuticals: quality of life, cause. 2002.

[25] D. Molitor, P. Reichhart, M. Spann, and A. Ghose. Measuring the effectiveness of location-based advertising: A randomized field experiment. 2019.

[26] S. Morse, M. C. Gonzalez, and N. Markuzon. Persistent cascades: Measuring fundamental communication structure in social networks. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 969–975. IEEE, 2016.

[27] J. Nixon and P. Ulmann. The relationship between health care expenditure and health outcomes. *The European Journal of Health Economics*, 7(1):7–18, 2006.

[28] L. Pappalardo, S. Rinzivillo, and F. Simini. Human mobility modelling: exploration and preferential return meet the gravity model. *Procedia Computer Science*, 83:934–939, 2016.

[29] D. Ramachandram and G. W. Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108, 2017.

[30] S. Robben, M. Pol, and B. Kröse. Longitudinal ambient sensor monitoring for functional health assessments: a case study. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pages 1209–1216, 2014.

[31] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. *arXiv preprint arXiv:1207.4169*, 2012.

[32] B. L. Rosenberg, J. A. Kellar, A. Labno, D. H. Matheson, M. Ringel, P. VonAchen, R. I. Lesser, Y. Li, J. B. Dimick, A. A. Gawande, et al. Quantifying geographic variation in health care outcomes in the united states before and after risk-adjustment. *PLoS One*, 11(12), 2016.

[33] S. Saeb, M. Zhang, C. J. Karr, S. M. Schueller, M. E. Corden, K. P. Kording, and D. C. Mohr. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior. *Journal of medical Internet research*, 2015.

[34] C. Sievert and K. Shirley. Ldavis: A method for visualizing and interpreting topics. *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70, 2014.

[35] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In *Proceedings of the 10th ACM SIGKDD*, 2004.

[36] F.-T. Sun, Y.-T. Yeh, H.-T. Cheng, C. Kuo, and M. Griss. Nonparametric discovery of human routines from sensor data. In *2014 IEEE international conference on pervasive computing and communications (PerCom)*, pages 11–19. IEEE, 2014.

[37] T. D. Wachs, M. Georgieff, S. Cusick, and B. S. McEwen. Issues in the timing of integrated early interventions: contributions from nutrition, neuroscience, and psychological research. 2015.

[38] N. E. Williams, T. A. Thomas, M. Dunbar, N. Eagle, and A. Dobra. Measures of human mobility using mobile phone records enhanced with gis data. *PloS one*, 10(7):e0133630, 2015.

[39] J. Zheng and L. M. Ni. An unsupervised framework for sensing individual and cluster behavior patterns from human mobile data. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 153–162, 2012.

7. Appendix

7.1. Location to Activity trajectories

Home and Work activity groups : Given T_i of an individual, we find the location where a consumer spends the most time from 1 AM - 5 AM on all days, map this location across T_i to activity *home*. A similar design to identify home location has been earlier used in the location data literature to address privacy concerns [23]. To assign work, we exclude the *home* and locations within a 200m buffer around it (*not-home*) and proceed in the following sequence.

1. If the average time spent per day, across the observation period, at a *not-home* location is greater than 5 hours, indicating full-time work, we assign that location as consumer's *work*.

2. If the average time spent at a top *not-home* location is greater than 2 hours, indicating part-time work, we assign this location as *work* in T_i. If multiple locations satisfy this condition, indicating multiple part-time vocations, we assign all of these locations as *work*.

3. If a consumer spends less than 30 minutes at 3 or more *not-home* locations, indicating delivery behaviour, we assign such locations for a certain day as *work* to construct T_i.

4. Finally, if we are not able to identify a secondary *not-home* location where a consumer spent significant time in, we assume that the consumer does not have a steady job.

Other activity groups : To map the rest of the locations in T_i to activities, we identify a point of interest closest to the location using the Google Places API[13]. The API returns a list of decorations (refer to second column of Table 1) which can help capture an individual's behavior (*gym*, *amusement_park*,*hair_care*), their consumption (*restaurant*, *meal_takeaway*,*cafe*) and their leisure activities (*art_gallery*, *spa*, *bowling_alley*) for each location. We aggregate decorations with similar semantics to construct thirteen more activities (first column in Table 1) in addition to home and work. To add a temporal context, we append each mapped activity with a coarser timestamp of t^i_j, c^i_j : 12 - 2 AM, 3 - 5 AM, 5 - 7 AM, 7 - 9 AM, 9 - 11 AM, 11 - 2 PM, 2 - 5 PM, 5 - 7 PM, 7 - 9 PM, 9 - 12 PM.

7.2. Author Topic Models Primer:

The Author Topic model (ATM), introduced by [31] is a probabilistic generative model for documents that extends LDA [35] to include authorship of documents. In ATM, each author is associated with a multinomial distribution over topics and each topic, like LDA, is associated with a multinomial distribution over words. By modeling the interests of authors, ATM enables us to establish what topics an author writes about, which authors are likely to have written documents similar to an observed document, and which authors produce similar work.

Figure 7 illustrates the generative process with a graphical model using plate notation. Shaded and unshaded circles indicate observed and latent variables respectively. An arrow indicates a

[13] Google Places `https://developers.google.com/places/web-service/supported_types`

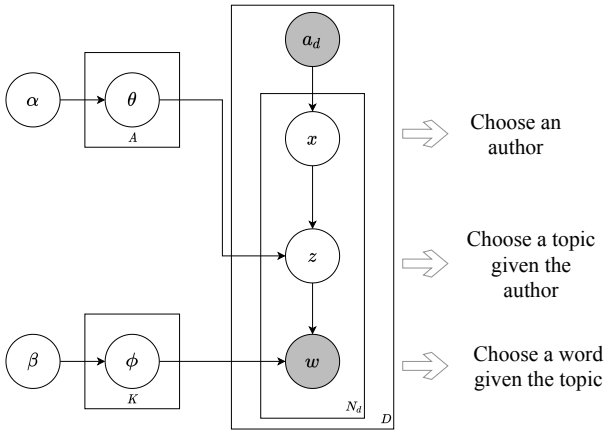


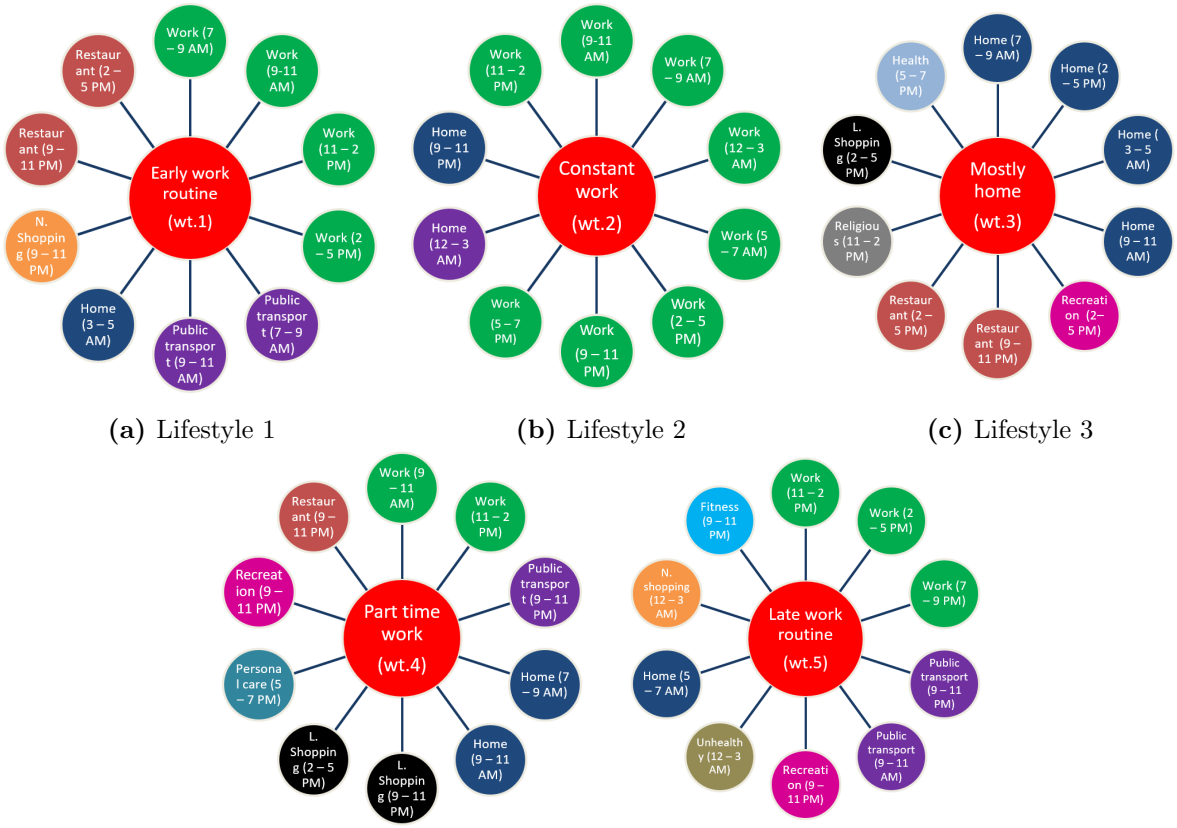
Figure 7 Probabilistic Graphical model of Author Topic Model using plate notation.

conditional dependency between variables and plates (the boxes in Figure 7) indicate repeated sampling with the number of repetitions given by the variable in the bottom. In ATM, we observe both w and a_d, the set of authors of document d. When generating a document, an author is chosen at random ($Uniform(a_d)$) for each individual word in the document. The author picks a topic from their multinomial distribution over topics ($A \times K$ matrix, denoted by θ), and then samples a word from the multinomial distribution over words associated with that topic ($W \times K$ matrix, denoted by ϕ). This process is repeated for all words in the document until all the documents are created. Formally, the distributions of the unobserved variables are

$$\begin{aligned} P(\theta|\alpha) &\sim Dirichlet(\alpha) \\ P(\phi|\beta) &\sim Dirichlet(\beta) \\ P(z|x,\theta^{(x)}) &\sim Multinomial(\theta^{(x)}) \\ P(w|z,\phi_{(z)}) &\sim Multinomial(\phi_{(z)}) \\ P(x|a_d) &\sim Uniform(a_d) \end{aligned} \tag{7}$$

Note that the last equation simplifies to $x = a_d$; $|a_d| = 1$ in our lifestyle identification since each document would only comprise of a single consumer's day to day activities.

Gibbs Sampling and Estimation : The main objectives of ATM inference are to estimate the probability of generating w from topic k, $\phi_k^{(w)}$ and the probability of assigning topic k to a word generated by author a, $\theta_k^{(a)}$. More generally, for a given training corpus D_{train}, we need to estimate an approximation of the posterior distribution $P(\theta, \phi|z, x, D_{train}, \alpha, \beta)$, where $P(\theta, \phi|\alpha, \beta) = P(\theta|\alpha)P(\phi|\beta)$. Following the approach suggested in [31], we first obtain an empirical sample based estimate of $P(z, x|D_{train}, \alpha, \beta)$ using Gibbs sampling for 1000 iterations (chaining). Next, we compute posterior estimates by leveraging the fact that Dirichlet and multinomial are conjugate distributions (Refer Eq 6).



(a) Lifestyle 1 **(b)** Lifestyle 2 **(c)** Lifestyle 3

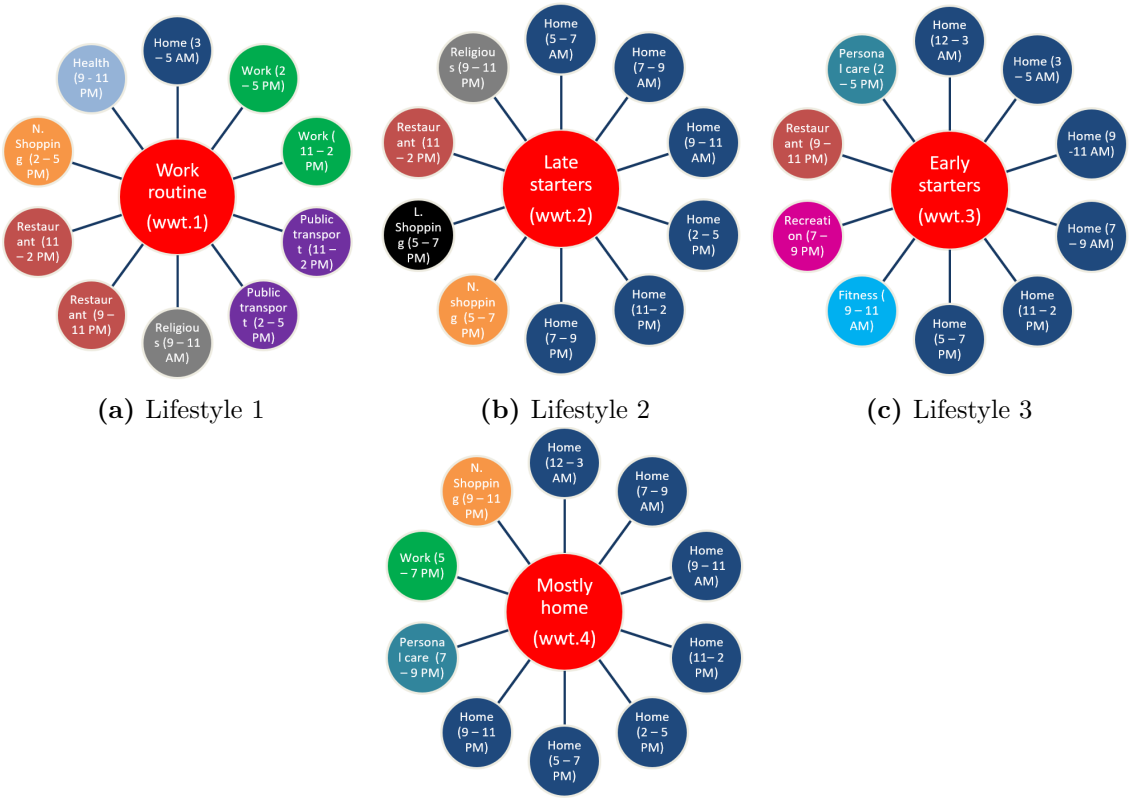
(d) Lifestyle 4 **(e)** Lifestyle 5

Figure 8 DC Weekday Lifestyles

7.3. D.C. Residents

Weekday Lifestyles: Figure 8 visualizes the five identified weekday lifestyles and their corresponding activities for D.C. residents. Lifestyle 5 (denoted by *wt.5*) characterizes a late work routine (*work* over 11 - 2 PM, 2 - 5 PM, 7 - 9 PM), commute via public transportation mornings and evenings (*publictransport* over 9 - 11 AM , 9 - 11 PM), late night recreation, unhealthy activities, fitness and necessity shopping (*recreation* in 9 - 11 PM, *fitness* 9 - 11 PM, *necessityshopping* 12 - 3 AM, *unhealthy.activities* 12 - 3 AM,). In contrast, lifestyle 1 (*wt.1*, with similar commute pattern, reveals an early work routine: *work* 7 - 9 AM, 9 - 11 AM, 11 - 2 PM, 2 - 5 PM, and consumption at restaurants during evenings and nights (*restaurant* 2 - 5 PM, 9 - 11 PM). Both lifestyles feature a steady full-time work routine, and work-fitness balance. Lifestyle *wt.4*, in comparison, indicates a part-time job (*work* 9 - 11 PM, 11 - 2 AM). Lifestyle 3 (*wt.3*) reveals a mostly at home routine while lifestyle 2 (*wt.2*) indicates multiple full-time/part-time jobs.

Weekend Lifestyles: Figure 9 displays the top ten activities of the four weekend lifestyles for D.C. residents. Different from Baltimore, we observe that two lifestyles *wwt.1* (*work* 11 - 2 PM, 2 -



(a) Lifestyle 1 **(b)** Lifestyle 2 **(c)** Lifestyle 3

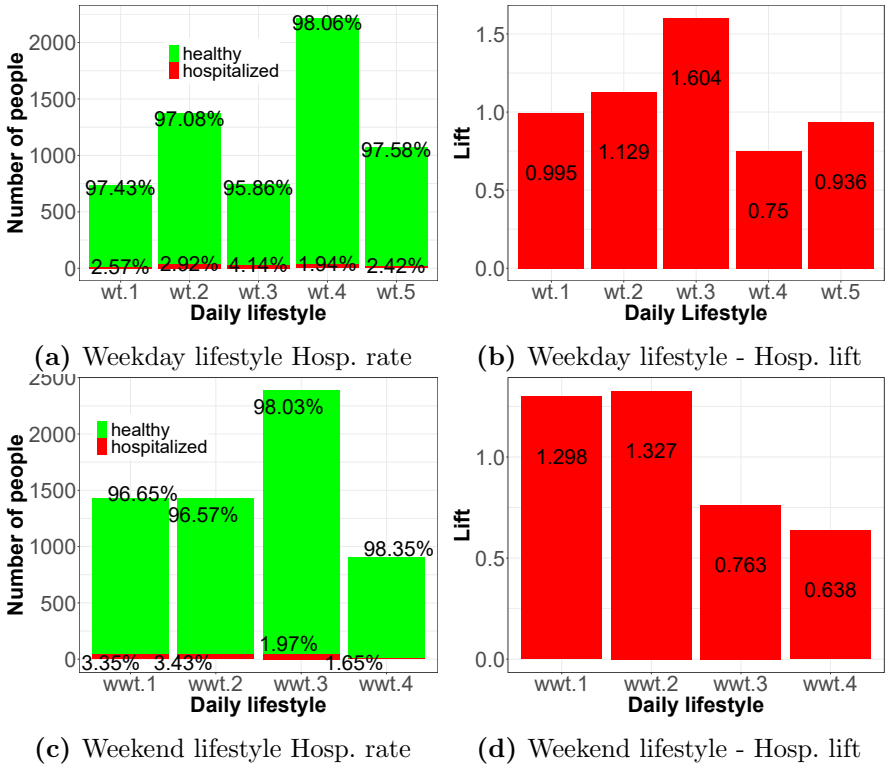
(d) Lifestyle 4

Figure 9 DC Weekend Lifestyles

5 PM) and *wwt.4* (*work* 5 - 7 PM) with work activities, all other lifestyles suggest a non-work routine. Lifestyle *wwt.3* characterizes an early start weekend routine with fitness activities (*fitness* 9 - 11 AM) and *personal.care* afterwards (2 - 5 PM). In contrast, lifestyle *wwt.2* indicates consumption at restaurants later in the morning (*restaurant* 11 - 2 PM), with shopping and religious activities later in the evening (*leisure.shopping, necessity.shopping* 5 - 7 PM, *religious* 9 - 11 PM). Besides work on weekends, individuals in lifestyle *wwt.1* regularly consume at restaurants (*restaurant* 11 - 2 PM, 9 - 11 PM).

Model-free Evidence: Figure 10 exhibits the histogram of the percentage of the 6,114 Baltimore residents with each lifestyle visiting medical facilities. Weekday lifestyles *wt.3* and *wt.2* have higher (4.14% and 2.92%, Figure 10a) than average (2.58%) percentages of individuals visiting medical facilities. In contrast, lifestyle *wt.4* has fewer than (1.94 %) average percentage of hospitalizations. Similarly, Figure 10c reveals that weekend lifestyles *wwt.1* and *wwt.2* experience higher percentages of hospitalizations whereas lifestyle *wwt.4* are sixty percent less likely.

In Figures 10b, 10d), we present the lift scores of weekend and weekday lifestyles. The top activities characterizing each lifestyle (Figures 8, 9) and their lift scores suggest that those who participate in *personal.care* activities on weekends (*wwt.4*) or weekdays (*wt.4*) are less likely (0.63



(a) Weekday lifestyle Hosp. rate **(b)** Weekday lifestyle - Hosp. lift

(c) Weekend lifestyle Hosp. rate **(d)** Weekend lifestyle - Hosp. lift

Figure 10 (D.C.) Association with Hospitalization : Model free analysis (*hospitalization*)

and 0.75, respectively) to have hospitalizations on average. On the other extreme, those with either busy, volatile work routines (*wt.2*) or no work routine (*wt.3*) on weekdays, are 1.12 and 1.60 times more likely to have hospitalizations (Figures 10b, 8; and people who either work (*wwt.1*) or are late starters (*wwt.2*) on weekends are 1.29 to 1.32 times more likely to have hospitalizations than average (Figures 10d, 9). Overall, the model-free evidence, similar to Baltimore residents, reveals heterogeneous rates of hospitalizations across different lifestyles.

Logit Analysis: Table 10 (Column 4) indicate that those with $wwt.1$, $wt.3$, $wt.2$ have significantly higher odds of having a future hospitalization (1.31, 1.61, and 1.49 respectively) than average, after controlling for other social determinants. Similarly, lifestyles $wwt.4$ and $wt.4$ have significantly lower odds than average. These insights are qualitatively consistent with the model free evidence. Similar to Baltimore residents, we do not find any significant association between X_i^{access}, X_i^{demog} and future hospitalizations indicating that two individuals who live in the same neighborhood with similar social demographics, access to parks/fitness facilities, but with different lifestyles, will have different health risks. In Table 11, we introduce total dwell time at healthy (fitness, personal care) and unhealthy activities into the regression and observe that *regularity* of personal care activities matters (lifestyle $wt.4$, $wwt.4$), instead of the total dwell time. In contrast to Baltimore, we do not observe that total time spent at unhealthy activities is significantly correlated to future hospitalization.

	Dependent variable: hospitalization				
	(1)	(2)	(3)	(4)	(5)
Weekend lifestyle 1 (wwt.1)	0.326**	0.292*	0.275*	0.272*	0.274*
	(0.150)	(0.151)	(0.159)	(0.164)	(0.165)
Weekend lifestyle 2 (wwt.2)	0.426**	0.425**	0.358*	0.315	0.314
	(0.175)	(0.176)	(0.181)	(0.184)	(0.185)
Weekend lifestyle 4 (wwt.4)	−0.473**	−0.509**	−0.417*	−0.412*	−0.410*
	(0.221)	(0.221)	(0.235)	(0.235)	(0.237)
Weekday lifestyle 2 (wt.2)	0.294	0.344*	0.407*	0.401*	0.401*
	(0.193)	(0.195)	(0.198)	(0.206)	(0.210)
Weekday lifestyle 3 (wt.3)	0.387**	0.387**	0.414**	0.481**	0.483**
	(0.176)	(0.177)	(0.185)	(0.192)	(0.193)
Weekday lifestyle 4 (wt.4)	−0.359*	−0.337*	−0.350*	−0.327*	−0.327*
	(0.153)	(0.155)	(0.166)	(0.170)	(0.171)
Weekday lifestyle 5 (wt.5)	0.079	0.010	0.215	0.195	0.196
	(0.188)	(0.191)	(0.203)	(0.207)	(0.218)
Accessibility metrics	✗	✓	✓	✓	✓
Mobility metrics	✗	✗	✓	✓	✓
Social Demographics	✗	✗	✗	✓	✓
Community Controls	✗	✗	✗	✗	✓
Observations	6,114	6,114	6,114	6,114	6,114
Log Likelihood	−484.149	−478.313	−453.994	−429.825	−398.986
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 10 (D.C.) Hospitalization Logit Analysis

	Dependent variable: hospitalization		
	(1)	(2)	(3)
Weekend lifestyle 1 (wwt.1)	0.273*	0.274*	0.282*
	(0.167)	(0.167)	(0.172)
Weekend lifestyle 2 (wwt.2)	0.312	0.314	0.314
	(0.186)	(0.187)	(0.193)
Weekend lifestyle 4 (wwt.4)	−0.411*	−0.411*	−0.411*
	(0.237)	(0.242)	(0.243)
Weekday lifestyle 2 (wt.2)	0.402*	0.404*	0.402*
	(0.211)	(0.211)	(0.213)
Weekday lifestyle 3 (wt.3)	0.484**	0.492**	0.496**
	(0.196)	(0.198)	(0.199)
Weekday lifestyle 4 (wt.4)	−0.329*	−0.327*	−0.328*
	(0.174)	(0.178)	(0.177)
Weekday lifestyle 5 (wt.5)	0.201	0.198	0.197
	(0.217)	(0.218)	(0.218)
total_fitness_dwell	−0.002		
	(0.009)		
total_personalcare_dwell		−0.014	
		(0.014)	
total_unhealthyactivities_dwell			0.003
			(0.004)
Other social determinants	✓	✓	✓
Observations	6,114	6,114	6,114
Log Likelihood	−378.824	−377.489	−378.868
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 11 (D.C.) Hospitalization : Additional Logit Analysis

7.4. Robustness Checks

Sensitivity in Health Outcome : In Table 8 and 9, we reported the predictive performance of alternate definitions of health outcomes *hospitalization_night*, *hospitalization_alt* . To further showcase the robustness of our findings, we replicate our logit analysis for both Baltimore and D.C. residents (Tables 12, 13). We observe that both the qualitative and quantitative findings remain consistent with our key hospitalization variable *hospitalization*.

	Dependent variable:							
	hospitalization_alt				*hospitalization_night*			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Weekend lifestyle 1 (wwt.1)	0.220*	0.216*	0.197*	0.170	0.213	0.205	0.101	0.149
	(0.118)	(0.120)	(0.125)	(0.130)	(0.150)	(0.151)	(0.158)	(0.163)
Weekend lifestyle 2 (wwt.2)	−0.938**	−0.920**	−0.818**	−0.784*	−0.915**	−0.894**	−0.881**	−0.798*
	(0.190)	(0.191)	(0.197)	(0.206)	(0.263)	(0.264)	(0.270)	(0.277)
Weekend lifestyle 4 (wwt.4)	−0.088	−0.093	−0.004	0.034	−0.018	−0.013	0.094	0.130
	(0.116)	(0.117)	(0.124)	(0.129)	(0.150)	(0.150)	(0.158)	(0.163)
Weekday lifestyle 1 (wt.1)	−0.383***	−0.377***	−0.315**	−0.288**	−0.273*	−0.282*	−0.230	−0.191
	(0.121)	(0.121)	(0.128)	(0.134)	(0.151)	(0.152)	(0.159)	(0.166)
Weekday lifestyle 2 (wt.2)	0.014	−0.010	−0.292**	−0.287**	0.093	0.086	−0.189	−0.108
	(0.129)	(0.130)	(0.136)	(0.143)	(0.162)	(0.164)	(0.170)	(0.179)
Weekday lifestyle 4 (ww.4)	−0.377**	−0.356**	−0.311**	−0.292*	−0.516**	−0.492**	−0.442**	−0.382*
	(0.160)	(0.162)	(0.168)	(0.177)	(0.218)	(0.219)	(0.225)	(0.240)
Weekday lifestyle 5 (wt.5)	0.714***	0.716***	0.627***	0.626***	0.966***	0.964***	0.886***	0.884***
	(0.114)	(0.116)	(0.122)	(0.126)	(0.136)	(0.138)	(0.144)	(0.149)
Accessibility metrics	✗	✓	✓	✓	✗	✓	✓	✓
Mobility metrics	✗	✗	✓	✓	✗	✗	✓	✓
Social Demographics	✗	✗	✗	✓	✗	✗	✗	✓
Observations	4,528	4,528	4,528	4,528	4,528	4,528	4,528	4,528
Log Likelihood	−1,117.051	−1,108.370	−1,006.016	−923.317	−779.343	−775.067	−700.783	−643.706

Note: *p<0.1; **p<0.05; ***p<0.01

Table 12 (Baltimore) Logit Analysis : Robustness check for Hospitalization

	Dependent variable:							
	hospitalization_alt				*hospitalization_night*			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Weekend lifestyle 1 (wwt.1)	0.226**	0.233**	0.241**	0.197*	0.245*	0.254*	0.282*	0.308**
	(0.102)	(0.103)	(0.107)	(0.111)	(0.142)	(0.143)	(0.148)	(0.155)
Weekend lifestyle 2 (wwt.2)	0.489**	0.493**	0.415**	0.373**	0.345**	0.342**	0.330*	0.320*
	(0.124)	(0.125)	(0.129)	(0.132)	(0.168)	(0.168)	(0.173)	(0.179)
Weekend lifestyle 4 (wwt.4)	−0.429**	−0.437**	−0.420**	−0.413**	−0.456**	−0.462**	−0.384*	−0.415*
	(0.138)	(0.138)	(0.146)	(0.150)	(0.191)	(0.192)	(0.201)	(0.210)
Weekday lifestyle 2 (wt.2)	−0.402**	0.430**	−0.458**	−0.496**	−0.435**	−0.424**	−0.467**	−0.460**
	(0.135)	(0.137)	(0.139)	(0.143)	(0.185)	(0.187)	(0.190)	(0.197)
Weekday lifestyle 3 (wt.3)	0.279**	0.267**	0.332**	0.381***	0.512***	0.508***	0.541***	0.582***
	(0.128)	(0.128)	(0.134)	(0.138)	(0.167)	(0.167)	(0.173)	(0.181)
Weekday lifestyle 4 (wt.4)	−0.370**	−0.373**	−0.408***	−0.421***	−0.420***	−0.412**	−0.422**	−0.445**
	(0.105)	(0.106)	(0.113)	(0.117)	(0.160)	(0.162)	(0.171)	(0.181)
Weekday lifestyle 5 (wt.5)	0.187	0.188	0.196	0.207	0.205	0.215	0.225	0.211
	(0.136)	(0.138)	(0.146)	(0.150)	(0.173)	(0.175)	(0.184)	(0.191)
Accessibility metrics	✗	✓	✓	✓	✗	✓	✓	✓
Mobility metrics	✗	✗	✓	✓	✗	✗	✓	✓
Social Demographics	✗	✗	✗	✓	✗	✗	✗	✓
Observations	6,114	6,114	6,114	6,114	6,114	6,114	6,114	6,114
Log Likelihood	−1,286.838	−1,283.772	−1,233.189	−1,160.998	−757.339	−755.076	−730.246	−674.633

Note: *p<0.1; **p<0.05; ***p<0.01

Table 13 (D.C.) Logit Analysis : Robustness check for Hospitalization