URBAN DESIGN 4 HEALTH

California Public Health Assessment Model: Final Data Development and Model Results

Prepared for:

Technical Advisory Committee

Strategic Growth Council

Sacramento Area Council of Governments

Prepared By:

Urban Design 4 Health, Inc., with support from Calthorpe Analytics

Date: June 30, 2015

Table of Contents

Introduction

The purpose of this memo is to summarize the final data development process, analytical methods, model results, and validation results for the UrbanFootprint (UF) physical activity and health models. The first draft was submitted to the Technical Advisory Committee (TAC) on June 30th, 2014, and a second draft was submitted to the TAC on October 31, 2014.

This memo finalizes the October 31st version, by removing preliminary versions of models and retaining only the final versions. The final models are unchanged from the October 31st version, based on TAC feedback to finalize all models as-is. The final models described in this report will next be programmed into UrbanFootprint and applied to a pilot case study for further testing.

Data development

The following sections provide details on data development tasks, including acquisition, review and cleaning.

California Household Travel Survey

The California Household Transportation Survey (CHTS) is conducted every 10 years to collect data on socioeconomic characteristics and travel behavior of California households. The 2010-2012 CHTS was recently released and includes data from 109,113 individuals within 42,131 households. These data are housed at the National Renewable Energy Laboratory (NREL). The CHTS consists of several survey components, including a telephone survey, a 24-hour travel diary, and three days wearing a global positioning satellite (GPS) device. Survey and diary data are available for all participants, while the GPS data is available only for a subsample of 8,202 persons in 3,871 households. These data were used to derive individual participation in transportation and recreational physical activity.

Components of the CHTS data set used in the analysis included the following:

- Geocoded household, trip, and activity location shapefile
- Household characteristics table
- Individual characteristics table
- Individual trip table
- Individual activity table

Application of exclusion criteria

The full CHTS data set was reduced by applying the following sequential exclusion criteria:

- 16,966 participants were located outside of the 30-county UrbanFootprint (UF) data extents.
- 296 participants were flagged by CALTRANS as having incomplete data.
- 9,689 participants reported being outside of the study area on the survey day.

- 6 participants had <1300 or >1600 minutes of travel + activity time based on trip table data.*
- 14,013 participants had activities with overlapping and/or nested activity times in the activity table.
- 3,464 participants recorded 24 hours of a single non-home activity, which upon manual review of a sub-sample appeared to actually be incomplete activity logs.
- 5,137 participants had <1000 or >1600 minutes of travel time based on trip table data + activity time based on activity table data.*
- 2,225 participants did not report their age.
- 2,101 participants were less than 5 years old.
- 1,486 participants home location did not intersect with a UF built environment grid cell.

*Because the CHTS outcome values are equal to the daily duration of specific travel and activity behaviors, under or over-reporting time engagement can result in invalid outcome values. Perfect accounting of daily time should result in 1,440 minutes of reported travel + activity episodes, though many logs have some amount of missing data (relatively common) or extra data (relatively rare). Imperfect time accounting was much more common in the activity table than in the trip table. Both activity and trip duration are available in the trip table, but only activity duration is available in the activity table. We noted that the trip and activity tables were often not entirely consistent with each other, but both contained uniquely important information, thus it was necessary to pull data from both tables. The exclusion range for activity table total time was based on the truncation range applied to other CHIS/CHTS variables (min= 25^{th} percentile value - 1.5 * interquartile range, max=75^{th} percentile value + 1.5 * interquartile range). Based on these threshold calculations, the lower bound for activity table total time was 991.5 minutes (which was rounded up to 1,000) while the upper bound was 1,707.5 minutes (which was rounded down to 1600, though this is somewhat irrelevant as the maximum time was 1,559 minutes). Because the interquartile range was equal to zero for trip table total time, the exclusion range was manually selected to exclude only six outliers identified by visual inspection (values of 719, 11,519, 20,159, 21,599, 25,919, and 27,359 minutes). The remainder of the trip table total times ranged from 1,319 to 1,459 minutes.

Note that we did not exclude those living at their current residence for less than 12 months. In the CHTS survey, tenure was recorded in years, and responses ranged from 1-97 years. Removing those responding 1 year would result in an additional exclusion of 2,961 participants. The decision not to remove those participants was made out of recognition that travel and physical activity patterns should adapt quickly to new built environments, unlike body weight and health outcomes.

After applying the above exclusion criteria, the final analytical sample consisted of 53,733 individuals. Participants were distributed by region as follows:

Bay Area	SACOG	SANDAG	San Joaq. Valley	SCAG	Total
14,739	3,134	2,777	8,890	24,193	53,733

When cross-classified by age and household income groups, the sample sizes are as follows:

	Household income groups:			
Age groups:	All	Low (<$50k)	Med ($50-100k)	High (>$100k)
All (5+)	53,733	16,933	17,068	19,732
Children (5-11)	4,829	1,670	1,378	1,781
Teens (12-17)	4,734	1,479	1,270	1,985
Adults (18-64)	35,695	10,593	11,283	13,819
Seniors (65+)	8,475	3,191	3,137	2,147

Imputing activity time for those with simultaneous activities

Several thousand travel survey records were identified that indicated a participant was engaged in two or more activities simultaneously over a specified period of time. For example, a participant may have entered three activities in their log, all beginning at 5pm and ending at 7pm, but with three different activity categories (e.g. cooking, watching tv, and socializing). To avoid having a participant's total travel/activity engagement sum to over 24 hours for the survey period, these simultaneous records were modified to split the time involvement evenly between multiple activities occurring simultaneously. For the above example, that would result in 40 minutes being assigned to each of the three activities, rather than 2 hours to each.

Active travel and recreational physical activity episode identification

Active trip legs were identified from the trip table according to the following criteria (note that each trip leg can only have one identified mode):

- Walk trips = any trip with a walk mode (mode = 1)
- Bike trips = any trip with a bike mode (mode = 2)
- Auto trips = any trip with a personal automobile mode (mode = 5-11)

Recreational physical activity episodes were identified according to the following criteria:

- The reported activity purpose at an activity location was:
    - At home, exercise (with or without equipment)/playing sports.
    - At work, exercise/sports.
    - At school, after school or non-class-related sports/physical activity.
    - At other place, outdoor exercise (playing sports/jogging, bicycling, walking, walking the dog, etc.).

    - At other place, indoor exercise (gym, yoga, etc.).
- Activity purpose was a loop trip, and further identified as a physically active loop trip according to the procedure described below.

Loop trip identification & cleanup

Loop trips are defined as trips that begin and end at the same location. Examples may include a walk around the neighborhood, biking for exercise, or a pleasure drive. In the CHTS survey, loop trips were typically recorded as activity episodes in the activity table along with an indication of travel mode in the trip table, though we found that the coding of loop trips in the two tables did not follow a consistent pattern. We also found that many times a loop trip was recorded as an activity episode but there was no clear information in the trip table to confirm the loop trip mode or in many cases that a loop trip actually occurred. Because loop trips made by walking or biking are relevant components of physical activity engagement, yet were not consistently coded in the activity and trip tables, loop trips warranted particular review and cleanup in the CHTS data.

The loop trip review resulted in the identification of the following coding irregularities:

- 1,372 loop trip or recreational activities were coded as 1 minute in duration in the activity table. Of these, 970 were preceded and followed by walk or bike trips in the trip table and the preceding trip end location was the same as the following trip end location. These 970 trips were recoded from walk/loop/walk or bike/loop/bike sequences to a single recreational physical activity episode. This coding scheme appeared to be the standard method for coding a loop trip.
- 4,384 loop trips were coded as being longer than 1 minute in duration in the activity table, but manual review suggested that many of these were not actually loop trips. Only 286 of these were recoded as recreational physical activity episodes according to the following criteria:
    - If the loop trip in the activity table was preceded and followed by a walk or bike trip and the preceding trip end location was the same as the loop trip activity location and the following trip end location, the walk/loop/walk or bike/loop/bike sequence was recoded to a single recreational physical activity episode.
    - If the loop trip in the activity table was preceded by a walk or bike trip and the preceding trip end location was the same as the loop trip activity location, the walk/loop or bike/loop sequence was recoded to a single recreational physical activity episode.
    - If the loop trip in the activity table was followed by a walk or bike trip and the loop trip activity location was the same as the following trip end location, the loop/walk or loop/bike sequence was recoded to a single recreational physical activity episode.
    - If the loop trip in the activity table was preceded and followed by a walk or bike trip, the preceding trip end location was the same as the following trip end location, and the place name for the loop trip activity matched the list of places below, the walk/loop/walk or bike/loop/bike sequence was recoded to a single recreational physical activity episode.

- LOOP, LOOP TRIP, WALK, NEIGHBORHOOD WALK, RICHMOND GREENWAY TRAIL, JOE RODATA TRAIL , AROUND SPRING LAKE, BIKE LOOP TRIP, RANCHO PARK , MONTAGUE PARK, WALK DESTINATION 2, BLACKBURN TALLEY PARK, NORTHSTAR PARK, BIKE/WALKING PATH, WALK DOG, TENNIS COURT, SMOTHERMAN PARK, AUBURN DISTRICT REGIONAL PARK, CENTRAL PARK, BIKE RIDE, ROOSEVELT PARK, DOG PARK, MCDOUGAL PARK, WALKING HIS DOG, LINCOLN PARK, TURN AROUND, DOG WALK, SIDEWALK

Identifying and truncating trip and activity duration outliers

Outliers in terms of walk/bike trip and recreational activity duration were identified and addressed by truncating high trip and activity times according to the following procedure:

- Walk/bike trip and recreational activity duration times were first log-transformed.
- Based on transformed walk/bike trip and recreational activity duration times, the upper threshold was set to equal the 75th percentile value + 1.5 * interquartile range for each variable. These upper thresholds (on the original, non-transformed scale) were calculated as follows:
    - Walk trip max threshold = 108.9 minutes / trip
    - Bike trip max threshold = 98.8 minutes / trip
    - Recreational activity max threshold = 623.2 minutes / episode

Walk/bike trip and recreational activity duration times greater than the maximum thresholds were recoded to equal the maximum threshold. This recoding process was applied to:

- 59 walk trips
- 38 bike trips
- 89 recreational activity episodes

Aggregate trip and activity data to persons, merge with person & household characteristics

The duration of walk, bike, and auto trips, and recreational physical activity episodes were then aggregated to daily totals per individual. The aggregated trip and activity durations were then merged with relevant characteristics from the person and household tables.

Covariate data development

The following covariates were developed from the CHTS data sets:

- Age in years
- Sex
    - 1=male
    - 2=female
- Race/ethnicity
    - 0=Hispanic
    - 1=White, non-Hispanic

    - 2=Black or African American, non-Hispanic
    - 4=Asian
    - 97= American Indian or Alaska native , Native Hawaiian or Pacific islander, or Other
- Educational attainment
    - 1=No high school diploma
    - 2=High school diploma
    - 3=Some college, no degree or vocational/associate's degree
    - 4=Bachelor's degree
    - 5=Graduate degree
- Adult employment status (only asked if participant was 16 or older)
    - 1=employed
    - 2=unemployed
- Adult home ownership
    - 1=own
    - 2=rent/other
- Household income
    - 1= <$10,000
    - 2= $10,000-$35,000
    - 3= $35,000-$50,000
    - 4= $50,000-$75,000
    - 5= $75,000-$100,000
    - 6= $100,000-$150,000
    - 7= >$150,000
- Household size
- Presence of children under 18 years old
    - 0 = no
    - 1 = yes
- Disability status
    - 0 = none
    - 1 = non-ambulatory disability
    - 2 = ambulatory disability
- For use in child & teen models only:
    - Maximum household educational attainment
    - Count of employed persons within household

Assess missing covariate data and impute missing covariates

Covariate data were reviewed to calculate the extent of missing data. The most commonly missing variable was household income, which was missing for over 7 percent of the sample. Ten percent of participants were missing at least one covariate. Note that the analysis & imputation of missing covariates was applied prior to joining the CHTS data to the UF built environment data, so the sample size for the covariate imputation included the 1,486 participants that were subsequently excluded because they did not join to UF built environment data.

Table 1; Covariate Data

Covariate	Sample size	Count of missing data	Percent missing
Age	55,222	0	0.0%
Household size	55,222	0	0.0%
Household vehicles	55,222	0	0.0%
Gender	55,222	49	0.1%
Employment status*	47,901	883	1.8%
Educational attainment	55,222	814	1.5%
Race/ethnicity	55,222	942	1.7%
Home ownership status	55,222	142	0.3%
Household income	55,222	4,119	7.5%
Any covariate	55,222	5,672	10.3%

* Employment status was not surveyed for those 15 and younger

Missing covariates were imputed in R using the "mi" (multiple imputation) library, which uses a Markov chain Monte Carlo method to predict plausible missing values (via regression) over multiple iterations, then combining the multiple imputed values into a single final value. Multiple imputation was applied for 30 iterations according to the following procedure:

- Missing home ownership status, household income, maximum household educational attainment, and the count of employed persons within a household were imputed using the mi() function based on non-missing household-level data.
- Missing gender, employment status, educational attainment, and race/ethnicity were imputed separately for each of the four age groups using the mi() function based on non-missing individual-level data.

All missing data was imputed by application of this procedure.

Transformations

Trip and activity duration distributions are commonly right skewed, with many short trips and few long trips. These distributions are also commonly zero-inflated, meaning that there are a large number of samples with no trips/activities. Both of these distributional issues require special treatment to ensure valid results in many statistical analyses.

The tables below summarize the following distributional characteristics for each of the CHTS outcome variables:

- Skewness – positive values indicate right skew, whereas negative values indicate left skew. The higher the z-score, the stronger the skewness.
- Kurtosis – indicates the height of the peak of the data relative to the tails, in comparison to a normal distribution. Low values indicate a flat top whereas high values indicate a sharp peak. The higher the z-score, the stronger the deviation from a normal peak.
- D'Agostino-Pearson omnibus test – this combines the skewness and kurtosis values to calculate a more comprehensive assessment of normality. The higher the chi^2 value, the more the distribution deviates from normality.

The normality metrics are highly sensitive to sample size, and even minor deviations can result in statistically significant findings of non-normality in large samples such as the CHTS data. As such, it is essentially impossible to generate non-significant normality metrics with the CHIS data, even after applying transformations.

Table 2: Distributional characteristics of CHTS outcome variables

Variable	transform-ation	skewness	skew z	kurotsis	kurt z	D'ago chi^2
Walk min/day (all)	none	5.5	200.8	43.8	NA	NA
Walk min/day (only >0)	none	2.5	55.4	10.3	38.3	4530.4
Bike min/day (all)	none	11.7	259.4	170.5	NA	NA
Bike min/day (only >0)	none	1.6	17.0	3.3	10.0	390.1
Auto min/day (all)	none	3.0	156.0	16.2	NA	NA
Auto min/day (only >0)	none	3.2	143.8	17.1	99.3	30531.2
Recreational PA min/day (all)	none	5.0	193.7	31.9	NA	NA
Recreational PA min/day (only >0)	none	2.4	59.6	6.9	37.0	4918.3

As expected, the trip and recreational physical activity duration distributions are right-skewed with high peaks and extensive deviation from normality. Removing the zero-inflated portion of the distribution helps reduce these deviations for the active transportation and recreational physical activity distributions, but the remaining non-zero data continues to deviate highly from normality. In response, log transformations were applied to the non-zero portions of the CHTS outcome variables.

Table 3: Distributional characteristics after transformations of CHTS outcome variables

variable	transform-ation	skewness	skew z	kurotsis	kurt z	D'ago chi^2
Walk min/day (only >0)	log	-0.5	-17.9	0.3	5.3	347.7
Bike min/day (only >0)	log	-0.4	-6.0	0.6	3.5	48.0
Auto min/day (only >0)	log	-0.3	-26.0	0.4	14.9	898.7
Recreational PA min/day (only >0)	log	-1.1	-36.5	4.4	31.4	2315.8

Although the log transformation did not eliminate the deviations from normality, the transformation did result in major improvement to all of the CHTS outcome distributions as compared to the non-transformed data, generating a reasonable approximation of normality to allow for valid model development.

Join built environment data

CHTS participants were spatially joined to UF built environment data in order to objectively describe the built environment around the home location of the CHTS participants. Using ESRI ArcMap (GIS software), geocoded CHTS household address points were spatially intersected with the UF grid shapefile. The result of this intersection is assignment of a unique ID corresponding to a UF grid cell to each CHTS household. Of the eligible CHTS participants, 1,486 participants were excluded because their household point did not intersect with the UF grid shapefile.

Details about the UF built environment surface are provided in a later section of this report called, "UrbanFootprint built environment data."

California Health Interview Survey

The California Health Interview Survey (CHIS) is the largest regularly-conducted health survey in the United States. CHIS data are housed at the UCLA Data Access Center (DAC). Between the years 2001-2009 the CHIS survey was conducted every other year. Beginning in 2011, the survey is now conducted continuously over two year periods. Data from the 2011-2012 cycle was recently released. Each CHIS survey includes between 40,000-50,000 households and consists of three unique survey instruments – one each for adults (18 and older), teens (12-17), and children (11 and younger). CHIS surveys are conducted by phone in all 58 California counties. These data provide individual health-related behaviors, health outcomes, and relevant covariates.

Because of the close temporal match to UF and CHTS data and comprehensive inclusion of physical activity variables, 2009 CHIS data were used for the following analyses. The 2009 CHIS survey data consists of:

- Adult survey (ages 18+), n = 47,614 individuals
- Adolescent survey (ages 12-17), n = 3,379 individuals
- Child survey (ages 0-11), n = 8,945 individuals

Application of exclusion criteria

The three CHIS data sets were reduced by applying the following sequential exclusion criteria:

- Adult
    - 7,981 participants were located outside of the 30-county UrbanFootprint (UF) data extents.
    - 184 participants reported being pregnant.
    - 359 participants reported being unable to walk.
    - 3,233 participants had geocoded home locations flagged as “medium” or “low” quality.
    - 724 participants did not join to UF built environment data.
- Adolescent
    - 527 participants were located outside of the 30-county UrbanFootprint (UF) data extents.
    - 283 participants did not join to adult covariate data.
    - 141 participants had geocoded home locations flagged as “medium” or “low” quality.
    - 61 participants did not join to UF built environment data.
- Child
    - 1,235 participants were located outside of the 30-county UrbanFootprint (UF) data extents.
    - 1,916 participants did not join to adult covariate data.
    - 2,289 participants were less than 5 years old.
    - 301 participants had geocoded home locations flagged as “medium” or “low” quality.

- 87 participants did not join to UF built environment data.

After applying the above exclusion criteria, the final analytical sample consisted of 40,617 individuals. Participants were distributed by region as follows:

Table 4: CHIS sample size by region

Bay Area	SACOG	SANDAG	San Joaq. Valley	SCAG	Total
9,446	4,006	5,239	4,988	16,938	40,617

The sample was cross-classified by age and household income groups, with sample sizes as follows:

Table 5: CHIS sample size by region, by age/income

	Household income groups:			
Age groups:	All	Low (<$35k)	Med ($35-100k)	High (>$100k)
All (5+)	40,617	17,669	11,173	11,775
Children (5-11)	3,117	1,159	757	1,201
Teens (12-17)	2,367	874	583	910
Adults (18-64)	23,515	9,188	6,537	7,790
Seniors (65+)	11,618	6,448	3,296	1,874

Outcome variable development

The following outcome variables were derived from CHIS data:

- Adult/senior physical activity variables
  - To address the zero-inflated distributions for the physical activity variables, each physical activity variable was divided into a binary component (no activity versus any activity) and a continuous component (minutes of activity for those with any activity)
  - Physical activity variables:
    - Weekly transportation walking
    - Weekly recreational walking
    - Weekly recreational moderate activity (excluding walking)
    - Weekly recreational vigorous activity
  - The four physical activity variables were also summed to calculate:
    - total minutes of physical activity
    - total metabolic equivalent (MET) minutes of physical activity, calculated as the sum of:
      - Minutes of transportation walking/week * 3.5[a]

[a] MET values based on Compendium of Physical Activities, 2011 update, for activity code 16060, walking for transportation. (https://sites.google.com/site/compendiumofphysicalactivities/)

        - Minutes of recreational walking/week * 3.5[b]
        - Minutes of recreational moderate activity/week * 4.0[c]
        - Minutes of recreational vigorous activity/week * 8.0[c]
- Child/teen physical activity
    - # of days with at least 60 minutes of PA during past week
    - Any walking/biking from school to home during past week
- Obesity
    - Adult/senior body mass index (BMI)
    - Child/teen body mass index percentile (BMI%)
    - Overweight or obese
        - For adults/seniors, BMI>=25
        - For adults/seniors, BM%>=85
    - Obese
        - For adults/seniors, BMI>=30
        - For adults/seniors, BM%>=95
- Adult/senior health status
    - Ever told by doctor of presence of high blood pressure
    - Ever told by doctor of presence of heart disease
    - Ever told by doctor of presence of type 2 diabetes
- General health (all age groups)
    - Self-reported poor/fair health (versus excellent/very good/good health)

Identifying and truncating continuous outcome variable outliers

Outliers in terms of adult/senior physical activity duration and BMI were addressed by truncating high activity times and both high and low BMI according to the following procedure:

- Physical activity duration times and BMI values were first log-transformed.
- Based on transformed physical activity duration times and BMI values, the upper threshold was set to equal the 75th percentile value + 1.5 * interquartile range for each variable. These upper thresholds (on the original, non-transformed scale) were calculated as follows:
    - Walking for transportation max threshold = 1084 minutes / week
    - Walking for recreation max threshold = 1230 minutes / week
    - Moderate physical activity max threshold = 1921 minutes / week
    - Vigorous physical activity max threshold = 1921 minutes / week
    - BMI max threshold = 50.0 (note that this was calculated to be 41.7, but increased based on external review of this threshold)

[b] MET values based on Compendium of Physical Activities, 2011 update, for activity code 17160, walking for pleasure. (https://sites.google.com/site/compendiumofphysicalactivities/)
[c] MET values based on Guidelines for Data Processing and Analysis of the International Physical Activity Questionnaire (IPAQ), November 2005. (http://www.ipaq.ki.se/scoring.pdf)

- Based on transformed BMI values, the lower threshold was set to equal the 25th percentile value - 1.5 * interquartile range for BMI. These lower thresholds (on the original, non-transformed scale) were calculated as follows:
    - BMI min threshold = 15.5

Physical activity duration times or BMI values greater than the maximum thresholds were recoded to equal the maximum threshold. BMI values less than the min threshold were recoded to equal the minimum threshold. This recoding process was applied to:

- 145 walking for transportation durations
- 68 walking for recreation durations
- 105 moderate physical activity durations
- 25 vigorous physical activity durations
- 269 BMI values

Covariate data development

The following covariates were developed from the CHIS data sets:

- Age in years
- Sex
    - 1=male
    - 2=female
- Race/ethnicity
    - 0=Hispanic
    - 1=white, non-Hispanic
    - 2=African American, non-Hispanic
    - 4=Asian
    - 97= American Indian/Alaska native, Native Hawaiian/Pacific islander, or other
- Adult/parent educational attainment
    - 1=No high school diploma
    - 2=High school diploma
    - 3=Some college, no degree or vocational/associate's degree
    - 4=Bachelor's degree
    - 5=Graduate degree
- Adult/parent employment status
    - 1=employed
    - 2=unemployed
- Adult/parent home ownership
    - 1=own
    - 2=rent/other

- Household income
    - 1= <$10,000
    - 2= $10,000-$35,000
    - 3= $35,000-$50,000
    - 4= $50,000-$75,000
    - 5= $75,000-$100,000
    - 6= $100,000-$150,000
    - 7= >$150,000
- Household size
    - Note that household size was truncated at a maximum of 10 people
- Presence of children under 18 years old (only relevant for adults)
    - 0 = no
    - 1 = yes
- Disability status (not available for teens)
    - 0 = none
    - 1 = any (ambulatory or non-ambulatory)

The child data set did not include parent employment status, household income, and parent home ownership variables. The adolescent data set did not include parent employment status, household income, and parent educational attainment variables. In both cases, these variables were derived from the adult survey conducted in the same child/adolescent household.

There was no missing covariate data in the CHIS data set.

Transformations

Physical activity duration and BMI distributions are commonly right skewed, with long physical activity durations and high BMI values less common than short physical activity durations and low BMI values. Physical activity distributions are also commonly zero-inflated, meaning that there are a large number of participants reporting no physical activity. Both of these distributional issues require special treatment to ensure valid results in many statistical analyses.

The tables below summarize the following distributional characteristics for each of the CHIS outcome variables:

- Skewness – positive values indicate right skew, whereas negative values indicate left skew. The higher the z-score, the stronger the skewness.
- Kurtosis – indicates the height of the peak of the data relative to the tails, in comparison to a normal distribution. Low values indicate a flat top whereas high values indicate a sharp peak. The higher the z-score, the stronger the deviation from a normal peak.

- D'Agostino-Pearson omnibus test – this combines the skewness and kurtosis values to calculate a more comprehensive assessment of normality. The higher the chi^2 value, the more the distribution deviates from normality.

The normality metrics are highly sensitive to sample size, and even minor deviations can result in statistically significant findings of non-normality in large samples such as the CHIS data. As such, it is essentially impossible to generate non-significant normality metrics with the CHIS data, even after applying transformations.

Table 6: Distributional characteristics of CHIS outcome variables

Variable	transform-ation	skewness	skew z	kurotsis	kurt z	D'ago chi^2
Walking for transport min/wk (all)	none	5.1	158.3	34.3	102.0	35443.5
Walking for recreation min/wk (all)	none	3.5	134.6	18.4	90.8	26371.1
Moderate PA min/wk (all)	none	4.2	146.0	23.3	95.3	30383.9
Vigorous PA min/wk (all)	none	6.1	169.6	57.5	109.9	40837.7
Walking for transport min/wk (>0 only)	none	3.8	93.6	17.9	60.3	12399.6
Walking for recreation min/wk (>0 only)	none	3.2	99.7	14.7	66.8	14408.5
Moderate PA min/wk (>0 only)	none	3.5	102.1	15.4	66.1	14798.2
Vigorous PA min/wk (>0 only)	none	4.0	75.0	23.4	49.9	8111.2
Body mass index	none	1.2	75.1	2.3	43.7	7541.6

As expected, the physical activity duration distributions are right-skewed with high peaks and extensive deviation from normality. Removing the zero-inflated portion of the distribution helps reduce these deviations for the active transportation and recreational physical activity distributions, but the remaining non-zero data continues to deviate highly from normality. The BMI distribution demonstrates significant but lesser deviations from normality. In response, log transformations were applied to the non-zero portions of the CHIS outcome variables.

Table 7: Distributional characteristics after transformations of CHIS outcome variables

Variable	transform-ation	skewness	skew z	kurotsis	kurt z	D'ago chi^2
Walking for transport min/wk (>0 only)	log	0.3	12.9	-0.1	-2.5	174.0
Walking for recreation min/wk (>0 only)	log	-0.1	-3.4	-0.2	-7.7	71.7
Moderate PA min/wk (>0 only)	log	0.2	10.0	-0.1	-3.9	115.8
Vigorous PA min/wk (>0 only)	log	0.0	-1.8	0.1	2.5	9.4
Body mass index	log	0.5	38.3	0.5	14.5	1677.6

Although the log transformation did not eliminate the deviations from normality, the transformation did result in major improvement to the CHIS physical activity outcome distributions and marginal improvement to the BMI distribution as compared to the non-transformed data, generating a reasonable approximation of normality to allow for valid model development.

Join built environment data

CHIS participants were spatially joined to UF built environment data in order to objectively describe the built environment around the home location of the CHIS participants. Using ESRI ArcMap (GIS software), geocoded CHIS household address points were spatially intersected with the UF grid shapefile. The result of this intersection is assignment of a unique ID corresponding to a UF grid cell to each CHIS household. Of the eligible CHIS participants, 872 participants were excluded because their household point did not intersect with the UF grid shapefile.

Details about the UF built environment data are provided in the next section of this report.

UrbanFootprint built environment data

Built environment data sets were developed by Calthorpe Associates for five California MPO regions: SACOG (Sacramento region), SANDAG (San Diego County), SCAG (including Los Angeles), the San Francisco Bay Area, and San Joaquin Valley regions. This coverage includes 30 counties in total. UF data were developed from a variety of data sources, including parcel land use data, transportation system data, Census data, and employment data. UF data were provided to UD4H at the grid-level (150m square units equal to 5.5 acres each). The UF data coverage consisted of 4,692,011 grid cells that contained at least one resident or employee.

Built environment variable development

The following built environment variables were derived from the UF data:

- Variables calculated within 1km airline buffers of each grid cell:
    - Intersection density = count of intersections / buffer area [sq mi]
    - Local street length = total centerline length of local and secondary roads [mi]
    - Major street length = total centerline length of freeway and arterial roads [mi]
    - Dwelling unit count
    - Residential density = dwelling unit count / (sum of residential + mixed-use parcel land area [acres])
    - Retail building floor area [sq ft]
    - Non-residential floor-to-area ratio (FAR) = (retail + office building floor area [sq ft]) / = (retail + office parcel land area [sq ft])
        - Higher FAR indicates a taller building and/or a building that occupies more parcel land area
    - Land use mix = entropy of 5 land use categories:

$$\frac{\sum_i^5 (\frac{b_i}{\sum_i^5 b_i} \times \ln \frac{b_i}{\sum_i^5 b_i})}{-\ln 5}$$

        - where b = building floor area for each of i land use categories:
            - 1 = Residential
            - 2 = Retail & services
            - 3 = Restaurant & entertainment
            - 4 = Office
            - 5 = Public administration
        - Land use mix can range from a low of 0 (no mix) to a high of 1 (perfect mix)
    - Bus stop count
    - Rail stop/station count
    - Any transit stop/station count
    - Park or open space land area [acres]

- The distance from each grid cell was calculated to the nearest qualifying feature. Maximum distances were truncated at 2km unless otherwise specified.
    - Retail use
    - Restaurant use
    - Educational use
    - Park or open space
    - Rail transit stop or station (light rail, subway/metro, or heavy passenger rail)
    - Bus transit stop (truncated at 1km)
    - Any transit stop or station
    - Freeway or arterial roadway (truncated at 5km)
- The presence or absence of the following was calculated:
    - Rail access = Rail transit stop or station within 2km
    - Bus access = Bus transit stop within 1km
    - Major road exposure = Freeway or arterial roadway within 500m
- The following regional accessibility variables were calculated:
    - Residential accessibility = sum of dwelling units within home-to-work attraction distance
    - Employment accessibility = sum of employees within home-to-work attraction distance

Transformations

Built environment variable distributions are commonly right skewed, with high variable values typically less common than low values. This distributional issue requires special treatment to ensure valid results in many statistical analyses.

The tables below summarize the following distributional characteristics for each of the UF built environment variables:

- Skewness – positive values indicate right skew, whereas negative values indicate left skew. The higher the z-score, the stronger the skewness.
- Kurtosis – indicates the height of the peak of the data relative to the tails, in comparison to a normal distribution. Low values indicate a flat top whereas high values indicate a sharp peak. The higher the z-score, the stronger the deviation from a normal peak.
- D'Agostino-Pearson omnibus test – this combines the skewness and kurtosis values to calculate a more comprehensive assessment of normality. The higher the chi^2 value, the more the distribution deviates from normality.

The normality metrics are highly sensitive to sample size, and even minor deviations can result in statistically significant findings of non-normality in large samples such as the CHIS and CHTS data.

Normality metrics were calculated separately for CHIS and CHTS samples after joining UF built environment data to CHIS/CHTS participants.

Table 8: UF variable normality metrics for the CHIS sample.

Variable	Transformation	skewness	skew z	kurotsis	kurt z	D'ago chi^2
Intersection density	none	1.6	90.4	6.9	70.5	13140.1
Dwelling unit count	none	5.2	160.4	46.8	107.9	37371.0
Residential density	none	6.5	175.0	62.3	112.0	43180.0
Retail floor area	none	5.8	167.5	53.5	109.9	40136.1
Non-residential FAR	none	10.8	206.9	468.6	132.8	60425.2
Retail distance	none	2.0	101.9	3.2	51.7	13066.8
Restaurant distance	none	1.2	76.0	0.3	8.8	5852.7
Land use mix	none	0.9	61.3	0.7	20.4	4170.5
Bus stop count	none	5.7	166.8	63.4	112.3	40437.2
Rail stop/station count	none	12.2	214.5	229.9	127.0	62120.4
Transit stop/station count	none	5.8	167.6	64.3	112.5	40744.0
Rail distance	none	-2.5	-115.5	5.2	63.4	17368.5
Bus distance	none	0.1	9.8	-1.7	NA	NA
Transit distance	none	0.6	45.2	-1.3	-420.2	178636.1
Local street length	none	-0.1	-9.6	0.0	-1.9	95.5
Major street length	none	2.5	116.6	7.8	73.3	18960.4
Freeway/arterial distance	none	0.9	59.1	-0.4	-19.7	3878.6
Education distance	none	-0.2	-17.3	-1.8	NA	NA
Park acres	none	3.4	133.8	19.2	92.5	26448.4
Park distance	none	1.6	90.9	2.3	43.5	10153.3
Residential accessibility	none	1.1	70.1	0.5	16.4	5189.8
Employment accessibility	none	1.0	66.5	0.2	7.9	4484.7

Table 9: UF variable normality metrics for the CHTS sample.

Variable	Transformation	skewness	skew z	kurotsis	kurt z	D'ago chi^2
Intersection density	none	1.5	82.3	5.9	63.1	10759.0
Dwelling unit count	none	4.9	149.8	40.5	100.3	32478.4
Residential density	none	5.8	159.7	48.7	102.9	36097.2
Retail floor area	none	5.6	157.1	48.5	102.9	35269.9
Non-residential FAR	none	4.8	147.6	45.2	101.9	32163.9
Retail distance	none	1.9	94.0	2.8	45.9	10952.1
Restaurant distance	none	1.1	67.4	-0.1	-3.3	4554.8
Land use mix	none	0.9	58.3	0.7	18.7	3753.0
Bus stop count	none	5.3	154.4	51.1	103.6	34577.8
Rail stop/station count	none	10.0	191.6	158.3	117.0	50420.7
Transit stop/station count	none	5.4	155.1	51.8	103.8	34822.8
Rail distance	none	-2.2	-103.7	3.8	53.4	13592.2
Bus distance	none	0.0	0.8	-1.7	NA	NA
Transit distance	none	0.5	35.1	-1.5	NA	NA
Local street length	none	-0.1	-9.0	0.0	-0.6	80.7
Major street length	none	2.6	112.1	8.4	71.2	17647.5
Freeway/arterial distance	none	0.9	59.4	-0.3	-12.3	3680.2
Education distance	none	-0.4	-26.0	-1.7	NA	NA
Park acres	none	4.4	142.5	37.9	99.3	30162.2
Park distance	none	1.6	86.0	2.2	41.0	9073.7
Residential accessibility	none	1.1	67.6	0.6	18.1	4895.5
Employment accessibility	none	1.0	63.8	0.3	9.9	4164.9

As expected, most of the built environment distributions (with the exception of local street length) are right-skewed, some have high peaks, and overall demonstrate extensive deviation from normality. In response, a variety of transformations were tested (log, square root, fourth root, and squared) and the transformation resulting in the lowest D'Agostino-Pearson chi^2 value was selected.

Table 10: UF normality metrics following transformation for the CHIS sample.

Variable	Transformation	skewness	skew z	kurotsis	kurt z	D'ago chi^2
Intersection density	square root	0.1	5.4	0.3	9.5	119.3
Dwelling unit count	fourth root	-0.4	-26.8	2.2	42.7	2538.1
Residential density	log	-0.1	-9.4	1.5	34.3	1261.8
Retail floor area	fourth root	-0.2	-17.1	0.5	15.3	525.9
Non-residential FAR	fourth root	-0.6	-45.0	1.4	33.4	3146.8
Retail distance	square root	0.7	50.7	-0.3	-15.5	2813.9
Restaurant distance	none	1.2	76.0	0.3	8.8	5852.7
Land use mix	square root	-0.2	-16.9	0.0	-1.6	289.4
Bus stop count	square root	1.0	65.4	1.9	39.8	5867.4
Rail stop/station count	fourth root	3.3	133.6	10.0	78.9	24062.8
Transit stop/station count	square root	1.0	66.3	2.0	41.2	6091.2
Rail distance	none	-2.5	-115.5	5.2	63.4	17368.5
Bus distance	fourth root	-1.1	-71.6	0.7	19.8	5512.7
Transit distance	fourth root	-0.6	-44.4	-0.2	-9.4	2063.0
Local street length	none	-0.1	-9.6	0.0	-1.9	95.5
Major street length	square root	1.2	74.8	0.3	11.3	5722.7
Freeway/arterial distance	square root	0.2	13.3	-0.8	-55.4	3241.0
Education distance	fourth root	-1.2	-76.5	0.3	9.6	5939.1
Park acres	fourth root	-0.3	-21.9	-0.8	-53.3	3324.0
Park distance	square root	0.4	27.8	-0.2	-10.5	884.5
Residential accessibility	fourth root	-0.2	-11.6	-0.6	-31.1	1098.8
Employment accessibility	fourth root	-0.2	-12.8	-0.8	-53.4	3021.1

Table 11: UF normality metrics following transformation for the CHTS sample.

Variable	Transformation	skewness	skew z	kurotsis	kurt z	D'ago chi^2
Intersection density	square root	0.0	1.1	0.3	10.4	109.3
Dwelling unit count	fourth root	-0.3	-19.1	2.2	40.4	1998.8
Residential density	log	0.0	0.9	1.5	32.5	1059.6
Retail floor area	fourth root	-0.2	-11.1	0.5	14.3	328.2
Non-residential FAR	fourth root	-0.6	-39.0	1.2	27.6	2278.4
Retail distance	square root	0.7	44.3	-0.4	-20.7	2396.4
Restaurant distance	none	1.1	67.4	-0.1	-3.3	4554.8
Land use mix	square root	-0.2	-13.8	-0.1	-3.0	200.3
Bus stop count	square root	1.2	69.5	2.2	40.5	6466.6
Rail stop/station count	fourth root	3.0	120.6	7.7	69.3	19346.7
Transit stop/station count	square root	1.2	70.3	2.3	41.8	6693.8
Rail distance	none	-2.2	-103.7	3.8	53.4	13592.2
Bus distance	fourth root	-1.2	-70.8	1.0	24.2	5598.5
Transit distance	fourth root	-0.7	-44.2	-0.1	-5.9	1989.5
Local street length	none	-0.1	-9.0	0.0	-0.6	80.7
Major street length	square root	1.2	71.1	0.4	12.3	5212.6
Freeway/arterial distance	square root	0.2	17.2	-0.8	-46.9	2495.1
Education distance	fourth root	-1.4	-76.5	0.5	15.8	6108.4
Park acres	fourth root	-0.3	-18.1	-0.8	-47.0	2534.5
Park distance	square root	0.4	26.4	-0.2	-9.7	790.3
Residential accessibility	fourth root	-0.1	-9.9	-0.5	-25.0	723.4
Employment accessibility	fourth root	-0.2	-11.2	-0.8	-49.2	2547.9

In nearly every case, either a square root or fourth root transformation was applied. A few exceptions were residential density, where a log transformation was preferred; distance to nearest restaurant and distance to nearest rail station, where the zero-inflated distribution was not improved through the application of any transformations; and local street length, which was already roughly normal.

Addressing multicollinearty between UF variables

A multi-stage process was applied to detect and address multicollinearity between UF built environment variables. First, UF variables were grouped into domains of conceptually similar variables and bivariate correlations were calculated within any domains containing at least two variables. Correlation tables for domains with at least two variables are provided below. Absolute values close to 1 indicate high correlation, whereas absolute values close to 0 indicate low correlation. Positive values indicate a positive association between the two variables and negative values indicate a negative association.

- Network connectivity
  - Intersection density
  - Local street length

	Int. density	Loc rd length
Int. density	1	-
Loc rd length	0.70	1

- Residential access
  - Dwelling unit count
  - Residential density

	Res units	Res density
Res units	1	-
Res density	0.92	1

- Commercial access
  - Retail floor area
  - Non-residential FAR
  - Retail distance
  - Restaurant distance

	retail sqft	FAR	retail dist.	rest. dist.
retail sqft	1	-	-	-
FAR	0.84	1	-	-
retail dist.	-0.72	-0.66	1	-
rest. dist.	-0.62	-0.59	0.60	1

- Mixed use
  - Land use mix
- Transit access
  - Rail count
  - Rail access within 2km
  - Rail distance
  - Bus count
  - Bus access within 1km
  - Bus distance
  - Transit count
  - Transit distance

	Rail count	Rail any	Rail dist.	Bus count	Bus any	Bus dist.	Transit count
Rail count	1	-	-	-	-	-	-
Rail any	0.98	1	-	-	-	-	-
Rail dist.	-0.87	-0.88	1	-	-	-	-
Bus count	0.33	0.32	-0.37	1	-	-	-
Bus any	0.15	0.15	-0.20	0.75	1	-	-
Bus dist.	-0.18	-0.19	0.23	-0.73	-0.79	1	-
Transit count	0.37	0.35	-0.41	1.00	0.74	-0.72	1
Transit dist.	-0.24	-0.24	0.29	-0.76	-0.82	0.97	-0.77

- Major road exposure
  - Major road length
  - Freeway/arterial distance
  - Major road exposure within 500m

	Maj rd length	Maj rd dist.	Maj rd any
Maj rd length	1	-	-
Maj rd dist.	-0.73	1	-
Maj rd any	0.68	-0.67	1

- School accessibility
  - Education distance
- Park access
  - Park acres
  - Park distance

	park acres	park dist.
park acres	1	-
park dist.	-0.60	1

- Regional accessibility
    - Residential accessibility
    - Employment accessibility

Res access	Emp access
1	-
0.99	1

Variance inflation factors (VIFs) were used to further detect the presence of multicollinearity. Variance inflation factors help quantify the amount of inflation in a variable's standard error that occurs due to multicollinearity in the model. A VIF of 1 indicates no multicollinearity, whereas higher VIFs indicate some multicollinearity. Although there is no standard VIF threshold to conclusively identify when multicollinearity has reached problematic levels, VIF values as low as 4 have been used as a guide.[d] Although VIF values help indicate the presence of multicollinearity, they cannot be used alone to identify which variables are collinear with each other. When entering all of the potential UF variables into two different CHTS models (one predicting whether the participant walked for transportation at all on the survey day, the other predicting whether the participant engaged in any recreational physical activity), the following VIFs were calculated for each variable, with VIF values greater than 4 highlighted in yellow:

[d] O'brien RM. A Caution Regarding Rules of Thumb for Variance Inflation Factors. Quality & Quantity. 41(5); 2007: 673-690.

Table 12: Variance inflation fractions (VIFs) for all original BE variables in CHTS transportation walking and recreation physical activity binary models. VIF values >4 (indicating high multicollinearity) are highlighted in yellow.

Variable	Transport walking	Recreational PA
Intersection density	2.0	2.2
Local street length	4.0	4.8
Dwelling unit count	16.5	16.4
Residential density	9.5	8.7
Retail floor area	12.2	11.6
Non-residential FAR	4.1	4.0
Retail distance	2.2	2.4
Restaurant distance	2.2	2.2
Land use mix	6.1	5.9
Rail count	31.9	32.8
Rail access	28.5	29.8
Rail distance	5.6	5.2
Bus count	487.0	523.2
Bus access	3.5	3.6
Bus distance	31.4	25.2
Transit count	514.6	555.8
Transit distance	38.7	32.1
Major street length	2.6	2.6
Freeway/arterial distance	3.0	2.9
Major road exposure	2.2	2.1
Education distance	1.3	1.3
Park acres	1.6	1.7
Park distance	1.5	1.7
Residential accessibility	54.3	48.2
Employment accessibility	55.2	49.9

The list of built environment variables was reduced by combining variables within domains with high correlation (> 0.60) into composite variables. Each composite variable was calculated by first normalizing the variable values to z-scores, then summing the normalized variables. The following composite variables were calculated in the first stage:

- Connectivity index = intersection density + local street length
- Residential access = dwelling unit count + residential density
- Commercial access = 2*retail floor area + 2*non-residential FAR - retail distance - restaurant distance
    - Note that because better access is provided by shorter distances, both distance variables were entered negatively. Thus, higher commercial access can be interpreted

as living at a location near more retail floor area, more intense (urban) non-residential land uses, and with closer proximity to the nearest retail and restaurant.
    - Relative differences between the association of each variable and the likelihood of any CHTS transportation walking were used to apply a simple weighting scheme to component variables
- Transit access = 2 * transit count – transit distance
    - Relative differences between the association of each variable and the likelihood of any CHTS transportation walking were used to apply a simple weighting scheme to component variables
- Park access = park acres - park distance
    - Note that because better access is provided by more park acres but shorter distance, the acreage variable was entered positively and the distance variable was entered negatively. Thus, higher park access can be interpreted as closer proximity to the nearest park and having more local park acreage.
- Major road index = major street length + any major road within 500m
- Regional access = residential accessibility + employment accessibility

Additionally, we dropped the following variables due to their high correlation with other variables within the same domain, problematic (zero-inflated) distributions, and/or the complication of using in a composite variables (e.g. combining a continuous and a binary variable, such as distance to nearest rail stop/station and any rail stop/station within 2km):

- Bus count and rail count (as total transit count was preferred)
- Distance to nearest rail stop/station (as rail stop/station access within 2km was preferred)
- Distance to nearest bus stop (as distance to nearest transit stop/station was preferred)
- Bus stop access within 1km (as distance to nearest transit stop/station was preferred)
- Distance to nearest freeway/arterial (as any major road exposure within 500m was preferred)

Following variable consolidation and elimination, 3 original variables and 7 composite variables remained. Bivariate correlations were calculated for all remaining variables, as demonstrated below. As with the VIFs, there is no definitive threshold above which multicollinearity becomes a problem, but we have chosen a threshold of 0.6 and greater to indicate the greatest threats of multicollinearity.

Table 13: Bivariate correlations between preliminary built environment variables. Correlations with an absolute value > 0.6 (indicating high correlations) are highlighted in yellow.

	mixed use	rail access	school dist.	res access	com access	park access	reg. access	connect index	transit access	maj rd index
mixed use	1	-	-	-	-	-	-	-	-	-
rail access	0.27	1	-	-	-	-	-	-	-	-
school dist.	-0.06	-0.12	1	-	-	-	-	-	-	-
res access	0.49	0.31	-0.22	1	-	-	-	-	-	-
com access	0.73	0.25	-0.15	0.79	1	-	-	-	-	-
park access	0.15	0.09	-0.23	0.32	0.21	1	-	-	-	-
reg. access	0.31	0.18	0.15	0.45	0.46	0.11	1	-	-	-
connect index	0.43	0.28	-0.20	0.79	0.68	0.26	0.35	1	-	-
transit access	0.42	0.33	-0.10	0.56	0.53	0.13	0.41	0.47	1	-
major rd. index	0.30	0.23	-0.08	0.25	0.27	0.10	0.21	0.20	0.25	1

One correlated cluster of four variables remained: land use mix, residential access, commercial access, and connectivity. Variance inflation factors confirm that most but not all of the problematic multicollinearity has been addressed, with high VIF values still present for residential access and commercial access. The next highest VIF values are for connectivity and mixed use, though these values are less than 4.

Table 14: Variance inflation fractions (VIFs) for preliminary BE variables in CHTS transportation walking and recreation physical activity binary models. VIF values >4 (indicating high multicollinearity) are highlighted in yellow.

Variable	Transport walking	Recreational PA
mixed use	2.1	2.3
rail access	1.3	1.2
school distance	1.2	1.2
residential access	3.9	4.5
commercial access	3.8	4.4
park access	1.2	1.2
regional access	1.6	1.5
connectivity index	2.2	2.7
transit access	1.6	1.6
major road index	1.1	1.2

To address the remaining cluster of multicollinear variables, a walkability index variable was created as a composite of land use mix, residential access, commercial access, and connectivity. Relative differences between the association of each variable and the likelihood of any CHTS transportation walking were used to apply a simple weighting scheme to component variables, resulting in the following formula:

- Walkability index = 2*residential access + 1.5*connectivity + commercial access + 0.5 * land use mix

It does not appear necessary to bring transit access into the walkability index in order to address multicollinearity, although the effect of transit access should be monitored in the model results to ensure this is true. The correlation matrix for the seven final variables is as follows:

Table 15: Bivariate correlations between final seven built environment variables. The highest correlation (0.58) was highlighted in yellow.

	walk index	transit access	rail access	major rd. index	regional access	school distance	park access
walk index	1	-	-	-	-	-	-
transit access	0.58	1	-	-	-	-	-
rail access	0.32	0.33	1	-	-	-	-
major rd. index	0.27	0.25	0.23	1	-	-	-
regional access	0.46	0.41	0.18	0.21	1	-	-
school distance	-0.21	-0.10	-0.12	-0.08	0.15	1	-
park access	0.30	0.13	0.09	0.10	0.11	-0.23	1

The variance inflation factors confirm that problematic multicollinearity has been addressed:

Table 16: Variance inflation fractions (VIFs) for final BE variables in CHTS transportation walking and recreation physical activity binary models.

Variable	Transport walking	Recreational PA
walkability index	1.9	1.9
transit access	1.6	1.6
rail access	1.3	1.2
major road index	1.1	1.1
regional access	1.5	1.5
school distance	1.2	1.2
park access	1.1	1.1

Sample descriptive statistics

The following tables provide descriptive statistics for the CHIS and CHTS covariates, built environment variables, and outcome variables. Values in each table are either sample means or the percentage of participants in the variable category. For adult samples only, descriptive statistics are provided for the full sample, stratified by income groups, stratified by UF region, and stratified by disability status. These tables are followed by descriptive statistics comparing each of the four age groups without further stratification. In Table 19, 22, 25, and 28, the built environment variables are divided into two sections. The first (individual variables) consists of all built environment variables that were either incorporated into index variables or are entered directly into the models. These variables are in their original units. The second section (composite variables) consists of the five index variables described earlier. All index variables were re-scaled around a mean value of 0. As such, these variables are unitless and are only interpretable in terms of their value relative to 0. For example, the low income group has the highest mean walkability index of the three income groups, whereas the high income group has the lowest mean walkability index.

CHIS

Adult descriptive statistics – pooled, by income group, and by region

Table 17: CHIS adult covariate descriptive statistics. (mean values or percentages)

	Pooled	By Income group			By Region					By Disability Status		
Variable	All	Low	med	high	Bay Area	SACOG	SANDAG	SCAG	SJV	none	ambulatory	other
Age	46	44	47	48	48	46	47	46	44	45	52	46
Male %	43%	40%	41%	47%	44%	42%	42%	42%	43%	44%	43%	38%
Hispanic	23%	39%	16%	9%	10%	15%	21%	28%	37%	22%	26%	22%
White	56%	37%	63%	72%	64%	70%	63%	47%	52%	56%	50%	58%
Asian	14%	14%	13%	14%	20%	8%	9%	16%	5%	15%	16%	8%
Other	8%	9%	8%	5%	6%	8%	7%	9%	7%	7%	8%	12%
HS diploma or less	31%	53%	24%	11%	21%	31%	26%	33%	46%	28%	42%	37%
Some college	25%	27%	30%	19%	22%	30%	26%	25%	27%	23%	28%	32%
Bachelor's or higher	44%	20%	46%	70%	57%	39%	48%	41%	27%	48%	30%	31%
Employed %	64%	49%	69%	77%	68%	63%	65%	62%	60%	70%	52%	41%
Home owner %	65%	39%	74%	87%	67%	69%	66%	61%	65%	68%	54%	57%
HH income < $50k	39%	100%	0%	0%	28%	38%	36%	43%	51%	33%	55%	57%
HH income $50-150k	28%	0%	100%	0%	27%	32%	29%	27%	28%	29%	24%	24%
HH income > $150k	33%	0%	0%	100%	46%	30%	35%	30%	20%	38%	21%	19%
HH size	3.0	3.1	2.8	3.0	2.7	2.8	2.9	3.1	3.3	3.1	2.5	2.9
Any children present	43%	44%	38%	45%	39%	40%	40%	44%	51%	46%	40%	28%
Ambulatory disability	15%	22%	13%	8%	12%	17%	15%	15%	17%	0%	100%	0%
Other disability	11%	16%	10%	7%	11%	12%	11%	11%	12%	0%	0%	100%

Table 18: CHIS adult outcome descriptive statistics. (mean values or percentages)

	Pooled	By Income group			By Region					By Disability Status		
Variable	All	low	med	high	Bay Area	SACOG	SANDAG	SCAG	SJV	none	ambulatory	other
Walking – transp. (% with any)	48%	53%	45%	46%	54%	42%	46%	49%	42%	49%	41%	51%
Walking - rec. (% with any)	62%	56%	63%	68%	65%	61%	65%	62%	56%	65%	51%	59%
Moderate PA (% with any)	57%	52%	59%	62%	59%	61%	60%	55%	58%	60%	47%	57%
Vigorous PA (% with any)	32%	25%	32%	41%	37%	33%	33%	31%	27%	36%	16%	29%
Walking – transp. (min/wk)	54.3	66.6	49.7	43.6	59.9	47.6	49.4	55.5	49.9	54.2	47.5	63.7
Walking - rec. (min/wk)	83.6	74.0	84.5	94.2	93.4	79.9	91.4	81.5	67.6	87.2	66.1	83.2
Moderate PA (min/wk)	109.6	103.3	108.9	117.6	109.9	130.1	115.0	101.0	116.1	111.3	100.3	110.9
Vigorous PA (min/wk)	58.2	46.4	57.3	73.0	64.9	57.3	59.9	56.1	52.1	65.2	28.8	51.1
BMI	26.8	27.6	26.8	26.0	26.0	27.5	26.8	26.7	28.4	26.3	29.1	27.1
Overweight %	33%	32%	32%	34%	31%	34%	33%	33%	35%	33%	33%	31%
Obese %	23%	28%	24%	18%	18%	28%	23%	23%	33%	20%	25%	38%
High BP %	25%	28%	25%	22%	23%	26%	25%	26%	29%	21%	30%	44%
Heart disease %	5%	6%	4%	4%	4%	5%	4%	5%	5%	3%	6%	12%
Type 2 diabetes %	6%	8%	6%	4%	5%	5%	6%	7%	7%	4%	7%	14%
Poor health %	18%	32%	12%	6%	15%	15%	14%	20%	23%	10%	30%	47%

Table 19: CHIS adult built environment descriptive statistics. (mean values or percentages)

	Pooled	By Income group			By Region					By Disability Status		
Variable	All	low	med	high	Bay Area	SACOG	SANDAG	SCAG	SJV	none	ambul-atory	other
Individual variables												
Res. density (units/acre)	9.1	10.5	8.5	7.8	10.8	5.8	9.5	9.8	5.4	8.9	9.0	9.9
Dwelling unit count	4260	4872	4037	3725	5220	2497	4122	4821	2131	4215	4186	4653
Retail floor area (sq ft)	760600	908900	708100	629800	714200	386100	538000	1092000	251000	749700	769100	821300
Non-residential FAR	0.21	0.23	0.20	0.20	0.18	0.15	0.15	0.30	0.07	0.21	0.20	0.22
Retail distance (m)	338	263	359	410	418	514	198	191	688	347	326	298
Restaurant distance (m)	549	451	541	670	816	581	436	350	809	565	501	507
Land use mix index	0.08	0.08	0.07	0.07	0.08	0.08	0.07	0.09	0.04	0.08	0.08	0.08
Intersect. density (int/sq mi)	95	107	93	83	97	80	84	103	86	93	98	101
Local street length (mi)	29	31	29	28	31	24	27	32	25	29	29	30
Transit count	31	38	29	26	47	27	30	32	7	31	32	34
Transit distance (m)	764	666	779	868	711	870	504	643	1447	774	763	698
Rail access %	9%	12%	8%	6%	13%	7%	11%	8%	2%	8%	10%	10%
Major street length (mi)	1.2	1.4	1.1	0.9	1.3	0.8	1.8	1.3	0.4	1.2	1.2	1.3
Major road exposure %	22%	26%	22%	18%	29%	16%	27%	20%	16%	22%	24%	23%
Park area (acres)	33.8	26.0	34.4	42.6	59.3	31.3	57.9	17.6	19.2	34.8	30.6	31.4
Park distance (m)	449	482	450	410	279	454	374	485	718	444	475	451
Res. accessibility (houses)	531200	557200	498200	528300	433000	176700	344100	836300	154900	536100	499900	540300
Employ. accessibility (jobs)	693000	726100	645300	693900	566000	172200	445400	1118000	160100	701500	643100	702900
School distance (m)	1173	1170	1156	1191	828	629	353	1858	757	1181	1145	1159
Composite variables												
Walkability index	0.06	0.85	-0.14	-0.71	0.37	-2.01	-0.29	1.29	-2.66	-0.02	0.10	0.51
Transit access	0.27	0.74	0.14	-0.19	0.80	-0.08	0.78	0.56	-1.93	0.22	0.30	0.54
Major road exposure	0.06	0.25	0.04	-0.16	0.34	-0.28	0.46	0.03	-0.48	0.04	0.06	0.18
Park access	0.12	-0.10	0.13	0.36	1.12	0.31	0.83	-0.47	-0.61	0.15	0.00	0.07
Regional access	0.00	0.00	-0.15	0.13	-0.03	-2.10	-0.45	1.24	-1.98	0.04	-0.22	0.03

Descriptive statistics by age group

Table 20: CHIS covariate descriptive statistics by age group. (mean values or percentages)

Variable	seniors	adults	teens	children
Age	75	46	15	8
Male %	38%	43%	52%	51%
Hispanic	8%	23%	35%	34%
White	78%	56%	44%	42%
Asian	8%	14%	13%	15%
Other	6%	8%	9%	10%
HS diploma or less*	33%	31%	35%	33%
Some college*	28%	25%	22%	21%
Bachelor's or higher*	39%	44%	43%	46%
Employed %*	17%	64%	70%	67%
Home owner %*	80%	65%	69%	65%
HH income < $50k	56%	39%	37%	37%
HH income $50-150k	28%	28%	25%	24%
HH income > $150k	16%	33%	38%	39%
HH size	1.7	3.0	4.2	4.3
Any children present*	3%	43%	100%	100%
Ambulatory disability	34%	15%	-	-
Other disability	17%	11%	-	-
Any disability	-	-	-	6%

* For teens/children, statistics are for surveyed parent

Table 21: CHIS outcome descriptive statistics by age group. (mean values or percentages)

Variable	seniors	adults	teens	children
Walking – transp. (% with any)	37%	48%	-	-
Walking - rec. (% with any)	56%	62%	-	-
Moderate PA (% with any)	57%	57%	-	-
Vigorous PA (% with any)	17%	32%	-	-
Walking - trans. (min/wk)	44	54	-	-
Walking - rec. (min/wk)	85	84	-	-
Moderate PA (min/wk)	128	110	-	-
Vigorous PA (min/wk)	32	58	-	-
Days/week > 60 min PA	-	-	3.5	3.7
Walk/bike from school %	-	-	46%	36%
BMI / BMI%	26.3	26.8	59%	62%
Overweight %	35%	33%	15%	16%
Obese %	19%	23%	11%	22%
High BP %	59%	25%	-	-
Heart disease %	21%	5%	-	-
Type 2 diabetes %	15%	6%	-	-
Poor health %	23%	18%	9%	6%

Table 22: CHIS built environment descriptive statistics by age group. (mean values or percentages)

Variable	seniors	adults	teens	children
Individual variables				
Residential density (units/acre)	8.7	9.1	8.0	8.1
Dwelling unit count	4069	4260	3839	3845
Retail floor area (sq ft)	763000	760600	617900	660600
Non-residential FAR	0.21	0.21	0.18	0.19
Retail distance (m)	347	338	362	349
Restaurant distance (m)	540	549	591	573
Land use mix index	0.08	0.08	0.06	0.07
Intersection density (int/sq mi)	91	95	92	93
Local street length (mi)	29	29	29	29
Transit count	31	31	29	28
Transit distance (m)	779	764	793	805
Rail access %	7%	9%	21%	20%
Major street length (mi)	1.1	1.2	1.1	1.1
Major road exposure %	20%	22%	8%	7%
Park area (acres)	35.8	33.8	33.2	33.9
Park distance (m)	466	449	443	444
Res. accessibility (houses)	518000	531200	506900	512900
Employ. accessibility (jobs)	671100	693000	657400	669700
School distance (m)	1150	1173	1145	1182
Composite variables				
Walkability index	-0.18	0.06	-0.29	-0.17
Transit access	0.18	0.27	0.14	0.08
Major road exposure	-0.06	0.06	0.00	-0.02
Park access	0.11	0.12	0.14	0.12
Regional access	-0.05	0.00	-0.15	-0.09

CHTS

Adult descriptive statistics – pooled, by income group, and by region

Table 23: CHTS adult covariate descriptive statistics. (mean values or percentages)

	Pooled	By income			By region					By disability status		
variable	All	low	med	high	Bay Area	SACOG	SANDAG	SCAG	SJV	none	ambulatory	other
Age	46	44	46	47	47	47	46	45	45	45	53	49
Male %	49%	47%	49%	51%	50%	48%	49%	49%	49%	50%	43%	48%
Hispanic %	25%	47%	21%	11%	14%	13%	27%	30%	32%	25%	26%	27%
White %	62%	40%	67%	75%	69%	77%	63%	56%	60%	62%	59%	59%
Asian %	8%	5%	7%	10%	11%	5%	5%	7%	3%	8%	2%	3%
Other race %	6%	8%	6%	4%	5%	5%	4%	6%	5%	5%	12%	11%
HS diploma or less %	26%	50%	22%	11%	17%	22%	25%	28%	39%	25%	37%	47%
Some college %	28%	31%	33%	21%	24%	32%	27%	29%	32%	28%	37%	30%
Bachelors or higher %	46%	19%	44%	68%	60%	46%	47%	44%	29%	47%	26%	24%
Employed %	72%	57%	75%	81%	76%	72%	73%	72%	65%	75%	28%	31%
Home owner %	77%	53%	82%	92%	80%	83%	76%	76%	77%	78%	60%	66%
HH income < $50k	30%	100%	0%	0%	19%	26%	30%	32%	44%	28%	58%	60%
HH income $50-100k	32%	0%	100%	0%	29%	38%	32%	32%	33%	32%	26%	24%
HH income > $100k	39%	0%	0%	100%	52%	36%	38%	36%	24%	40%	16%	15%
HH size	3.2	3.3	3.1	3.2	3.0	3.1	3.3	3.3	3.3	3.2	2.7	2.8
Any children present	35%	35%	32%	38%	34%	34%	38%	35%	36%	36%	17%	18%
HH vehicles	2.1	1.6	2.2	2.5	2.1	2.2	2.2	2.2	2.1	2.2	1.6	1.6
Ambulatory disability	3%	6%	2%	1%	2%	4%	3%	3%	5%	0%	100%	0%
Other disability	3%	6%	2%	1%	2%	4%	2%	3%	5%	0%	0%	100%

Table 24: CHTS adult outcome descriptive statistics. (mean values or percentages)

	Pooled	By income			By region					By disability status		
Variable	All	low	med	high	Bay Area	SACOG	SANDAG	SCAG	SJV	none	ambulatory	other
Walk travel (min/day)	5.0	7.7	3.6	4.1	7.6	2.8	3.8	4.3	3.6	4.9	6.3	7.2
Bike travel (min/day)	1.2	1.0	1.0	1.6	2.1	1.7	1.2	0.9	0.5	1.3	0.4	0.8
Auto travel (min/day)	75	63	77	82	74	79	71	77	71	76	55	49
Recreational PA (min/day)	18	13	18	22	20	20	24	17	14	18	10	8
Walk travel (any)	15%	20%	11%	13%	22%	9%	12%	12%	10%	14%	17%	19%
Bike travel (any)	2%	2%	2%	3%	4%	3%	2%	2%	1%	2%	1%	2%
Auto travel (any)	83%	74%	85%	88%	84%	85%	84%	83%	81%	84%	61%	61%
Recreational PA (any)	17%	12%	17%	21%	22%	18%	21%	16%	12%	18%	9%	11%

Table 25: CHTS adult built environment descriptive statistics. (mean values or percentages)

	Pooled	By income			By region					By disability status		
variable	All	low	med	high	Bay Area	SACOG	SANDAG	SCAG	SJV	none	ambulatory	other
Individual variables												
Res. density (units/acre)	9.0	10.2	8.7	8.4	11.7	5.8	9.8	9.1	5.2	9.0	9.7	9.7
Dwelling unit count	4,276	4,688	4,168	4,048	5,684	2,567	4,125	4,430	2,104	4,272	4,358	4,318
Retail floor area (sq ft)	753,300	865,500	736,700	681,000	772,700	356,200	549,400	1,002,000	241,500	751,000	789,800	787,600
Non-residential FAR	0.21	0.22	0.20	0.20	0.19	0.15	0.15	0.28	0.07	0.21	0.21	0.20
Retail distance (m)	367	302	380	407	416	543	194	216	695	369	332	339
Restaurant distance (m)	608	483	585	723	872	608	445	393	806	614	525	508
Land use mix	0.08	0.08	0.07	0.07	0.08	0.08	0.06	0.09	0.04	0.07	0.08	0.08
Int. density (int/sq mi)	98	109	96	91	104	80	89	103	85	98	103	105
Local road length (mi)	30	31	30	30	33	24	28	31	24	30	30	30
Transit count	29	35	28	26	43	26	32	28	7	29	31	33
Transit distance (m)	843	762	864	888	737	948	534	723	1,418	844	851	805
Rail access %	10%	12%	9%	8%	17%	7%	14%	8%	2%	10%	10%	11%
Major road length (mi)	1.1	1.4	1.1	0.9	1.3	0.8	1.8	1.2	0.4	1.1	1.2	1.2
Major road exposure %	21%	24%	21%	19%	29%	15%	26%	19%	14%	21%	23%	23%
Park area (acres)	30	21	30	36	49	36	55	18	19	30	25	24
Park distance (m)	465	510	486	412	301	433	381	489	715	462	513	500
Res. accessibility (houses)	525,600	537,000	500,200	537,800	440,500	195,400	322,900	774,000	168,200	528,000	497,300	480,200
Employ. accessibility (jobs)	686,500	695,200	646,500	712,500	590,300	188,200	409,800	1,028,000	175,900	690,500	637,300	612,200
School distance (m)	1,254	1,219	1,237	1,294	833	625	345	1,880	757	1,260	1,182	1,139
Composite variables												
Walkability index	0.1	0.7	-0.1	-0.3	0.7	-2.0	-0.2	1.0	-2.7	0.0	0.3	0.3
Transit access	0.0	0.4	-0.1	-0.3	0.6	-0.3	0.7	0.3	-1.9	0.0	0.1	0.2
Major road exposure	0.0	0.2	0.0	-0.1	0.4	-0.4	0.4	0.0	-0.5	0.0	0.1	0.1
Park access	0.0	-0.2	0.0	0.2	0.9	0.4	0.8	-0.5	-0.6	0.0	-0.2	-0.1
Regional access	0.0	-0.1	-0.1	0.2	0.0	-1.9	-0.6	1.0	-1.8	0.0	-0.3	-0.3

Descriptive statistics by age group

Table 26: CHTS covariate descriptive statistics by age group. (mean values or percentages)

Variable	seniors	adults	teens	children
Age	73	46	15	8
Male %	49%	49%	52%	50%
Hispanic	14%	25%	36%	41%
White	76%	62%	53%	46%
Asian	5%	8%	7%	8%
Other	6%	6%	4%	5%
HS diploma or less*	12%	12%	17%	20%
Some college*	26%	25%	24%	22%
Bachelor's or higher*	61%	62%	60%	58%
Employed %*	27%	72%	-	-
Home owner %*	88%	77%	78%	70%
HH income < $50k	38%	30%	31%	35%
HH income $50-100k	37%	32%	27%	29%
HH income > $100k	25%	39%	42%	37%
HH size	2.314	3.2	4.4	4.7
% with any child <18*	4%	35%	100%	100%
HH vehicles	1.835	2.1	2.2	2.0
Ambulatory disability	11%	3%	1%	0%
Other disability	8%	3%	1%	1%

* For teens/children, statistics are for surveyed parents and education is max in household

Table 27: CHTS outcome descriptive statistics by age group. (mean values or percentages)

Variable	seniors	adults	teens	children
Walking – transp. (% with any)	10%	15%	24%	21%
Biking – transp. (% with any)	1.1%	2.2%	4.4%	2.4%
Auto travel (% with any)	74%	83%	77%	81%
Recreational PA (% with any)	17%	17%	22%	25%
Walking – transp. (min/day)	3.4	5.0	7.3	5.1
Biking – transp. (min/day)	0.6	1.2	1.5	0.8
Auto travel (min/day)	61	75	48	51
Recreational PA (min/day)	18	18	35	34

Table 28: CHTS built environment descriptive statistics by age group. (mean values or percentages)

Variable	seniors	adults	teens	children
Individual variables				
Residential density (units/acre)	8.4	9.0	7.7	8.3
Dwelling unit count	4021	4276	3707	3955
Retail floor area (sq ft)	720300	753300	608200	654700
Non-residential FAR	0.20	0.21	0.18	0.18
Retail distance (m)	402	367	386	374
Restaurant distance (m)	612	608	652	629
Land use mix index	0.22	0.22	0.21	0.21
Intersection density (int/sq mi)	93	98	93	98
Local street length (mi)	29	30	29	30
Transit count	27	29	24	26
Transit distance (m)	881	843	883	866
Rail access %	8%	10%	7%	9%
Major street length (mi)	1.1	1.1	1.1	1.1
Major road exposure %	20%	21%	21%	21%
Park area (acres)	29.7	29.8	29.3	28.2
Park distance (m)	493	465	464	466
Residential accessibility	1269	1254	1283	1263
Employment accessibility	520000	525600	511900	511000
School distance (m)	679000	686500	671300	666800
Composite variables				
Walkability index	-0.33	0.06	-0.31	-0.09
Transit access	-0.17	-0.02	-0.23	-0.13
Major road exposure	-0.04	0.00	-0.04	-0.01
Park access	-0.08	0.01	-0.03	-0.03
Regional access	-0.01	0.02	-0.03	-0.07

Model development procedures

Models were generated for the following outcomes. For each outcome, one model was generated for each age group with all income groups pooled together, and three additional models were generated for adults after stratifying the sample into the three income groups:

Table 29: Preliminary models by age

	Age cohort				Adjusting for:			
Data set	Adult/ senior	Teen/ child	Outcome	Regression	Dem /SES	Tr. walk	PA	BMI
CHIS	both		Walking - trans. (min/wk)	Two-part	x			
CHIS	both		Walking - rec. (min/wk)	Two-part	x	x		
CHIS	both		Moderate PA (min/wk)	Two-part	x	x		
CHIS	both		Vigorous PA (min/wk)	Two-part	x	x		
CHIS		both	Days/week > 60 min PA	Poisson	x			
CHIS		both	Likelihood to walk/bike from school	Binary	x			
CHIS	both	both	BMI / BMI%	Linear	x		x	
CHIS	both	both	Likelihood to be overweight/obese	Binary	x		x	
CHIS	both	both	Likelihood to be obese	Binary	x		x	
CHIS	both		Likelihood to have high BP	Binary	x		x	x
CHIS	both		Likelihood to have heart disease	Binary	x		x	x
CHIS	both		Likelihood to have type 2 diabetes	Binary	x		x	x
CHIS	both	both	Likelihood to have poor health	Binary	x		x	x
CHTS	both	both	Walking for transport. (min/day)	Two-part	x			
CHTS	adults		Biking for transport. (min/dau)	Two-part	x			
CHTS	both	both	Auto travel (min/day)	Two-part	x			
CHTS	both	both	Recreational PA (min/day)	Two-part	x	x		

Two-part regression indicates that the model was split into two parts. In the first, binary logistic regression was used to model the likelihood of any activity versus no activity. In the second, linear regression was used to model the amount of activity in minutes for only the portion of the sample with any activity.

All demographic and socioeconomic covariates listed previously were tested for inclusion in every model. Additionally, all recreational physical activity models adjusted for whether a participant engaged in any transportation walking. All BMI/overweight/obesity models adjusted for total metabolic equivalent (MET)-minutes of physical activity (the sum of MET-minutes of transportation walking, recreation walking, moderate PA, and vigorous PA). Finally, the four health outcome models adjusted for total minutes of physical activity and BMI (for adults/seniors) or BMI% (for teens/children).

Two separate modeling processes were completed for each model. First, each built environment variable was added one-at-a-time to a model that adjusted for all applicable covariates. This provided information about each built environment variable when isolated from potential impacts of multicollinearity. Next, all built environment variables were added simultaneously to a model that adjusted for all applicable covariates. Finally, a variable selection process was conducted to remove covariates and built environment variables that, when removed, improved the Aikake information criterion (AIC, a metric indicating relative model fit) by at least 1.[e] This was accomplished by calculating the change in AIC that would occur from removing each individual variable, removing the variable that would result in the largest reduction in AIC (so long as the reduction was >1), then repeating until removing the variable that would result in the largest reduction in AIC would only reduce the AIC by <=1.

[e] Burnham KP, Anderson DR. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, 2nd ed. Springer-Verlag. 2002.

The following table shows an example of raw model output for one CHTS outcome (the likelihood of any transportation walking for adults), including all demographic/socioeconomic covariates and all built environment variables entered simultaneously:

Table 30: Raw model output for CHTS outcome -- likelihood of any transportation walking for adults

Variable name	Definition	Coefficient	p value
(Intercept)	Model intercept	-0.02786	0.833377
gender2	Gender = female	0.072842	0.026013
age	Age	-0.01098	9.60E-15
racehisp1	Race/eth = white, non-Hispanic	0.107283	0.020898
racehisp2	Race/eth = African American, non-Hispanic	0.07508	0.365301
racehisp4	Race/eth = Asian	-0.00081	0.990775
racehisp97	Race/eth = American Indian/Alaska native, Native Hawaiian/Pacific islander, or other	-0.04655	0.702832
emply2	Employed = no	0.235686	5.87E-10
educa2	Education = High school diploma	-0.22711	0.000758
educa3	Education = Some college, no degree, vocational or associate's degree	-0.20524	0.002634
educa4	Education = Bachelor's degree	0.043382	0.552193
educa5	Education = Graduate degree	0.292059	0.000123
own2	Home owner = no	0.221444	1.10E-07
hhsize	Household size	0.132621	2.08E-17
hhveh	Number of household vehicles available	-0.67511	5.18E-185
incom2	Income = $10,000-$35,000	-0.36747	9.11E-07
incom3	Income = $35,000-$50,000	-0.52909	1.72E-09
incom4	Income = $50,000-$75,000	-0.56595	6.91E-11
incom5	Income = $75,000-$100,000	-0.46971	1.58E-07
incom6	Income = $100,000-$150,000	-0.37637	2.70E-05
incom7	Income = > $150,000	-0.0683	0.455768
child_any1	Any children <18 in household = yes	-0.27355	2.44E-10
disabled1	Disability status = non-ambulatory	-0.15502	0.084862
disabled2	Disability status = ambulatory	-0.2485	0.008276
walk_index	Walkability index	0.071057	5.20E-35
transit_access	Transit access index	0.029562	8.34E-06
rail_any	Presence of a rail stop/station within 2km	0.312423	1.44E-10
majrd_index	Major road exposure index	-0.00415	0.635124
regional_access	Regional access index	0.046249	1.01E-05
school_distance	Distance to nearest school	-0.06163	1.88E-18
park_access	Park access index	0.085254	8.34E-15

Final model results

The following tables summarize the coefficient direction and strength of association for each built environment variable in the final models. Summary model tables are presented for each age cohort for pooled income groups, and for adults after further stratifying by income group. All tables include coefficient information for the following variables (where z() indicates a z-scaled component of a composite variable):

- **Walkability index** (composite variable) = z[z(dwelling unit count) + z(residential density)] + z[z(retail floor area) + z(non-residential FAR) - z(distance to nearest retail) - z(distance to nearest restaurant)] + z(land use mix) + z[z(intersection density) + z(local street length)]
- **Transit access** (composite variable) = z(transit count) - z(distance to nearest transit stop)
- **Rail access** = Presence or absence of a rail transit stop/station within 2km
- **Major road index** (composite variable) = z(major street length) + z(any major road within 500m)
- **Regional access** (composite variable) = z(regional residential accessibility) + z(regional employment accessibility)
- **School distance** = distance to nearest educational land use (truncated at 2km)
- **Park access** (composite variable) = z(park acres) – z(distance to nearest park)

For the models that adjust for physical activity and/or BMI, the tables also include the following variables:

- **Any transportation walking** = participant participated in >0 minutes of transportation walking
- **Total PA (MET-minutes)** (adult/senior only) = sum of metabolic equivalent (MET)-minutes of transportation walking, recreational walking, moderate recreational physical activity, and vigorous recreational physical activity
- **Days/week >60 min PA** (child/teen only) = count of days per week (from 0-7) that participant typically engages in at least 60 minutes of physical activity
- **Body mass index** (adult/senior only)
- **Body mass index %** (child/teen only) = age & gender adjusted body mass index. As with adult body mass index, a lower value indicates lower body weight relative to height.

For every model outcome, the table summarizes the coefficient direction and strength of association for each built environment variable in the final model. For certain model outcomes, the table also provides the coefficient direction and strength of association for physical activity and body weight variables also included in the model. For example, for the CHIS adult type 2 diabetes model built environment variables are both directly associated with diabetes *and* indirectly associated with diabetes through their association with physical activity and body mass index.

A 1-page summary of the symbols used in the model results table is provided in Appendix E.

The key for the table symbols and colors is as follows:

Table 31: Symbol and color definitions related to model coefficients

+++	Positive association, p < 0.001 (strong statistically significant)
++	Positive association, p < 0.05 (statistically significant)
+	Positive association, p > 0.05 (not statistically significant)
-	Negative association, p > 0.05 (not statistically significant)
--	Negative association, p < 0.05 (statistically significant)
---	Negative association, p < 0.001 (strong statistically significant)
	Variable was tested but not selected for inclusion in the model
NA	Variable was not tested for the model

Face validation

Each cell in each model table also provides an indicator of whether the direction of the coefficient met prior expectations. The table below summarizes the expected direction of the association between each built environment variable and specific outcomes based on literature reviews[f,g,h,i] and prior experience. For example, we expect higher walkability index values to be associated with more active transport (utilitarian physical activity), less auto transportation, lower body weight, and better health conditions, while the literature does not provide enough evidence for the impact on recreational physical activity. The symbols used in the walkability index column in the table below represent these relationships with a "+", "-" or a "?" respectively, where the "?" indicates that there is not clear evidence available for establishing a prior expectation.

[f] McCormack GR, Shiell A. In search of causality: a systematic review of the relationship between the built environment and physical activity among adults. *Int J Behav Nutr Phys Act*. 2011;8(1):125.
[g] Van Holle V, Deforche B, Van Cauwenberg J, et al. Relationship between the physical environment and different domains of physical activity in European adults: a systematic review. *BMC Public Health*. 2012;12(1):807.
[h] Durand CP, Andalib M, Dunton GF, Wolch J, Pentz MA. A systematic review of built environment factors related to physical activity and obesity risk: implications for smart growth urban planning. *Obes Rev*. 2011;12(5):e173–82.
[i] Feng J, Glass TA, Curriero FC, Stewart WF, Schwartz BS. The built environment and obesity: a systematic review of the epidemiologic evidence. *Health & Place*. 2010;16(2):175–90.

Table 32: Expected direction for the association of each built environment variable with various outcomes.

	BE variable with presumed association for outcomes						
Outcomes	Walkability index	Transit access	Rail access %	Major road exposure	Regional accessibility	School distance	Park access
Active transport	+	+	+	-	+	-	+
Recreational phys. Activity	?	?	?	-	?	-	+
Auto transport.	-	-	-	+	+	+	?
Body weight and health outcomes related to inactivity and obesity	-	-	-	+	?	+	-

In the following results tables, any counterintuitive associations (those not matching the sign of the prior expectation) are indicated in red text with a footnote to provide commentary on what may be causing the counterintuitive association. For any coefficients where there was insufficient evidence to provide a prior expectation, a superscript "?" is used in the results to indicate that the result was not compared to prior expectations. The key for interpreting the face validation symbols is provided immediately below, followed by additional explanation on investigating the counterintuitive associations.

Table 33: Symbol and color definitions related to face validation.

--[?]	Superscript "?" indicates that a prior expectation was not established for this coefficient.
--[b]	Red text indicates counterintuitive association, and footnote provides potential explanation for the counterintuitive association.
--[b]	Counterintuitive association was caused by multicollinearity problem.
	Variable was dropped from the final model due to multicollinearity problem.
--[b]	Cumulative impact of two-part model was in expected direction.

Multicollinearity

While the vast majority of the model coefficients met these expectations, those that did not are called out in red in the following tables. As part of the analysis of unexpected findings, there were indications that the walkability index, transit access, rail access, and regional access variables are causing multicollinearity problems in some models. Eight out of 37 original counterintuitive coefficients (based on application of the methodology described above) appeared to be caused by multicollinearity problems, as the coefficient in the preliminary model (where each built environment was entered one-at-a-time) was in the expected direction. Therefore, the observed counterintuitive direction of

association created through multiple regression output was most likely a function of interactions between the independent predictors simultaneously included in the model.

It also appeared that multicollinearity may be reducing the strength of these same variables in several models, causing one or more to be dropped from the final model when applying the variable selection methods described on page 46.

In response to the multicollinearity problems described above, we have developed and implemented the following approach. In cases where there was evidence that multicollinearity was a problem we have selectively added problematic variables (transit access, rail access or regional access) into the walkability index.

The process we applied was as follows:

1. We compared the coefficients and p-values for the four potentially multicollinear built environment variables (walk index, walkability, transit access, rail access, regional access) from the preliminary version of each model (where each built environment variable was entered one-at-a-time) to the final version.
2. For any of the four potentially multicollinear built environment variables, if the sign of the variable's coefficient in the preliminary model was opposite of the sign in the final model, we assumed this was caused by multicollinearity, and applied one of two remedies:
    a. If the variable coefficient's p-value in the preliminary model was < 0.05, we added the variable to the walk index with a sign that matches that found in the preliminary model. We weighted the components of the revised walkability index according to the relative t-values for each component variable in the preliminary model.
        i. For example, in the preliminary binary logistic model of any senior transportation walking, the walkability index had a strong positive association ($p<0.001$) as did transit access ($p<0.001$). However, when they are added simultaneously, the walkability index continues to have a strong positive association ($p<0.001$), but the sign on transit access became negative with a non-significant association ($p=0.21$). Following the variable selection methodology described on page 46, both variables were retained in the final model. As the only reason for the change in sign for the transit access coefficient appears to be due to multicollinearity, we propose to create a new walkability index variable that includes the original walkability index and transit access variables. Based on the higher t-value for the original walkability index ($t=12.7$) in the preliminary model as compared to transit access ($t=6.6$), we would assign a weight of $12.7/6.6=1.92$ to the walkability index and a weight of 1.0 to transit access. The new walkability index is thus equal to 1.92 * z(original walkability index) + 1 * z(transit access). The new walkability index had a positive association in the revised final model ($p<0.001$).

b. If the variable coefficient's p-value in the preliminary model was > 0.05, we dropped the variable from the final model.
    i. For example, in the preliminary binary logistic model of any senior recreational physical activity, the walkability index had a positive association but was not statistically significant (p=0.53). However, when added simultaneously with other correlated variables (transit access, rail access, and regional access), the sign on the walkability index became negative and the association strengthened (p=0.1). Following the variable selection methodology described on page 46, the walkability index was retained in the final model. As the only reason for the change in sign for the transit access coefficient appears to be due to multicollinearity, As the only reason for including the walkability index in the final model appears to be due to multicollinearity, we propose instead to remove the walkability index from the final model, which had a positive association but was non-significant (p=0.53) in the preliminary model.

3. For any of the four potentially multicollinear built environment variables, if the variable coefficient's p-value in the preliminary model was < 0.05 but the variable was dropped from the final model, we assumed this was caused by multicollinearity, and we added the variable to the walk index with a sign that matches that found in the preliminary model. We weighted the components of the revised walkability index according to the relative t-values for each component variable in the preliminary model.
    a. For example, in the preliminary linear model of senior transportation walking minutes, the walkability index had a positive association (p=0.016) as did transit access (p=0.019). However, when they are added simultaneously, they both continue to have a positive association, but the p-value for walkability dropped to 0.53 and for transit access dropped to 0.13. Following the variable selection methodology described on page 46, only the transit access variable was retained in the final model, and the p-value dropped again to 0.019 after removing the walkability index. As the only reason for dropping the walkability index from the final model appears to be due to multicollinearity, we propose instead to create a new walkability index variable that includes both the original walkability index and transit access variables. Based on the slightly higher t-value for the original walkability index (t=2.41) in the preliminary model as compared to transit access (t=2.35), we would assign a weight of 2.41/2.35=1.026 to the walkability index and a weight of 1.0 to transit access. The new walkability index is thus equal to 1.026 * z(original walkability index) + 1 * z(transit access). The new walkability index had a positive association in the revised final model (p=0.007).

This approach has been applied, and final results are provided on the following pages. Further analysis and details on the multicollinearity problems and potential solutions are provided in Appendix B.

Additional symbolization related to face validation and multicollinearity are defined in the table below.

Table 34: Symbol and color definitions related to multicollinearity.

any biking for transportation	The purple shading indicates a model that was revised to deal with multicollinearity problems. All other model results are identical between the two versions.
WI	Variable was added to the walkability index to address multicollinearity problem.
dropped	Variable was dropped from the final model to address multicollinearity problem.

A 1-page summary of all symbols used in the model results table is provided in Appendix E.

Two-part model interpretation

Interpretation of coefficient signs in two-part models requires an understanding not only of the signs in each individual part, but also on the cumulative impact when applying both parts of the model. In 10 out of 37 cases (which are highlighted in green), the counterintuitive sign occurred in one part of a two-part model, but the cumulative impact was in the expected direction. An example of this situation is that higher walkability was associated with greater likelihood of adults making any bike trips but also with less time spent biking. This suggests that areas with a high density of destinations may encourage additional people to bicycle, but the close proximity of destinations in these areas leads to shorter average travel distances. Although we predict that the average time spent biking per adult that does any cycling will be lower in areas with high walkability, the fact that we predict more people to do any cycling in areas of high walkability results in a net increase in the time spent biking when averaged across all adults.

Final approach to dealing with counterintuitive associations

Based on TAC feedback during meeting #4 and subsequent discussions, the final approach to dealing with counterintuitive results was as follows:

- Multicollinearity corrections were applied according to the methodology described above.
- All other counterintuitive results were accepted as-is.

CHTS models

Table 35: CHTS adult models.

Outcome	Any trans. walking	Walkability index	Transit access	Rail access	Major road exposure	Regional access	Distance to school	Park access
any walking for transportation	NA	+++	+++	+++		+++	---	+++
minutes/day transport walking for those with any	NA	+++	WI	++		+		++
any biking for transportation	NA	+++	++	++	-	WI	---	+++
minutes/day transportation biking for those with any	NA	--[j]				+++		
any automobile transportation	NA	---	-	---	++	-[k]	+++	+[?]
minutes/day automobile transport for those with any	NA	---	-	WI	-[l]	+++	+++	
any recreational PA	+++	--[?]						+++
minutes/day of recreational PA for those with any	---			+[?]				

Table 36: CHTS senior models.

Outcome	Any trans. walking	Walkability index	Transit access	Rail access	Major road exposure	Regional access	Distance to school	Park access
any walking for transportation	NA	+++	WI	++		++	---	+++
minutes/day transport walking for those with any	NA	++	WI					+
any automobile transportation	NA	---		---	++	++		
minutes/day automobile transport for those with any	NA	---	WI	dropped		dropped		
any recreational PA	+++	dropped	++[?]	--[?]		+[?]		++
minutes/day of recreational PA for those with any	---		+[?]	+[?]				

[j] Those living in more walkable areas have more opportunities for making bike trips but those that bike don't have as far to go to reach their destinations.
[k] Having better regional accessibility may mean less need to drive (due to other modal options) but longer & more congested travel for those that do drive.
[l] Living near a major road could mean more auto trips are made but those trips are relatively shorter.

Table 37: CHTS teen models.

Outcome	Any trans. walking	Walkability index	Transit access	Rail access	Major road exposure	Regional access	Distance to school	Park access
any walking for transportation	NA	+++	WI	WI	-	+	---	+++
minutes/day transport walking for those with any	NA	+++	WI					++
any automobile transportation	NA	+[m]			+	--[n]	+++	
minutes/day automobile transport for those with any	NA	---		+[o]	+	dropped		
any recreational PA	+++		++[?]			--[?]		++
minutes/day of recreational PA for those with any	---						-	

Table 38: CHTS children models.

Outcome	Any trans. walking	Walkability index	Transit access	Rail access	Major road exposure	Regional access	Distance to school	Park access
any walking for transportation	NA	+++	WI	+	--	+++	---	++
minutes/day transport walking for those with any	NA	++				dropped	++[p]	
any automobile transportation	NA	-	dropped	--	+			
minutes/day automobile transport for those with any	NA	--				+	+	
any recreational PA	+++		--[?]	+[?]				
minutes/day of recreational PA for those with any	---							

[m] Living in an area with high local accessibility provides opportunities for more teens to make auto trips, but they are relatively shorter auto trips.
[n] Having better regional accessibility may mean less need to drive (due to other modal options) but longer & more congested travel for those that do drive.
[o] ???
[p] As distance to school increases, those that walk to school must walk further, leading to more minutes of walking

CHIS models

Table 39: CHIS adult models.

Outcome	Any trans. walking	Total PA (MET-minutes)	BMI	Walkability index	Transit access	Rail access	Major road exposure	Regional access	Distance to school	Park access
any walking for transportation	NA	NA	NA	+++	+++	+++	---	+++	---	++
minutes/wk transport walking for those with any	NA	NA	NA	++		+			--	
any walking for recreation	+++	NA	NA	-?		-?	-	++?		++
minutes/wk recreational walking for those with any	+++	NA	NA				+[q]		--	
any recreational moderate PA	+++	NA	NA	---?	WI	WI		WI		
minutes/wk moderate PA for those with any	++	NA	NA	---?	WI	WI	+[r]	-?		-[s]
any recreational vigorous PA	+++	NA	NA	--?	dropped	-?		dropped		++
minutes/wk vigorous PA for those with any	-	NA	NA				--	+?		
body mass index	NA	---	NA	---	WI		++	WI	--[t]	---
likelihood of being overweight or obese	NA	---	NA	---	-	dropped	+	WI	--[t]	---
likelihood of being obese	NA	---	NA	---	-		++	WI	-[t]	-
likelihood of having high blood pressure	NA	---	+++		--				+	-
likelihood of having heart disease	NA	--	+++							-
likelihood of having type 2 diabetes	NA	--	+++	++[u]				--?		--
likelihood of having poor self-reported health	NA	---	+++	+[u]	+[v]	-	+			--

[q] Walking along/across/near major roads may involve more delay in street crossings.
[r] Jogging/biking along/across/near major roads may involve more delay in street crossings.
[s] Better park access may mean people visit parks or recreate near parks more frequently but recreate there for shorter episodes.
[t] ???
[u] Increased walkability had a net negative impact on physical activity (based on a positive association with utilitarian PA and a negative association with recreational PA), which is consistent with worse health outcomes. In one-at-a-time model, p=0.37 for diabetes and p=0.17 for self-reported health.
[v] Increased transit access may be associated with a more urbanized location with lower socioeconomic status, leading to a worse self-assessment of health.

Table 40: CHIS senior models.

Outcome	Any trans. walking	Total PA (MET-minutes)	BMI	Walkability index	Transit access	Rail access	Major road exposure	Regional access	Distance to school	Park access
any walking for transportation	NA	NA	NA	+++	+	++	-	+++	--	
minutes/wk transport walking for those with any	NA	NA	NA	++	-[w]	+		-[x]		
any walking for recreation	+++	NA	NA	dropped	++[?]	--[?]	-	+[?]		++
minutes/wk recreational walking for those with any	-	NA	NA	++[?]						+
any recreational moderate PA	+++	NA	NA	---[?]						+
minutes/wk moderate PA for those with any	+	NA	NA	--[?]	+[?]	-[?]		-[?]		--[y]
any recreational vigorous PA	+++	NA	NA	-[?]				+[?]		++
minutes/wk vigorous PA for those with any	--	NA	NA	-[?]		--[?]	+[z]		-	
body mass index	NA	---	NA			-	+	---[?]		
likelihood of being overweight or obese	NA	---	NA	---			+	WI		
likelihood of being obese	NA	---	NA			-		--[?]		
likelihood of having high blood pressure	NA	---	+++		-		++	+[?]		
likelihood of having heart disease	NA	---	++					-[?]		
likelihood of having type 2 diabetes	NA	---	+++	++[aa]	--			-[?]		-
likelihood of having poor self-reported health	NA	---	+	+[aa]	+[bb]			WI	+	-

[w] Better local transit may mean more people making shorter walk-to-transit trips.
[x] Better accessibility may mean more frequent but shorter walking trips.
[y] Better park access may mean people visit parks or recreate near parks more frequently but recreate there for shorter episodes.
[z] Walking along/across/near major roads may involve more delay in street crossings.
[aa] Increased walkability had a net negative impact on physical activity (based on a positive association with utilitarian PA and a negative association with recreational PA), leading to a net positive association on body weight and poor health conditions. In one-at-a-time model, p=0.44 for diabetes and p=0.006 for general health.
[bb] Increased transit access may be associated with a more urbanized location with lower socioeconomic status, leading to a worse self-assessment of health.

Table 41: CHIS teen models.

Outcome	Days/week >60 min PA	BMI%	Walkability index	Transit access	Rail access	Major road exposure	Regional access	Distance to school	Park access
days/week with at least 60 minutes PA	NA	NA	-?				-?	--	-cc
likelihood of walking/biking from school	NA	NA	+++	++	++	-	dropped	--	
body mass index percentile	+	NA				+			-
likelihood of being overweight or obese	-	NA	-						
likelihood of being obese	--	NA	-			+			
likelihood of having poor self-reported health	---	+++	-	+dd		-dd	+?		+cc

Table 42: CHIS children models.

Outcome	Days/week >60 min PA	BMI%	Walkability index	Transit access	Rail access	Major road exposure	Regional access	Distance to school	Park access
days/week with at least 60 minutes PA	NA	NA	--?	-?			WI	-	
likelihood of walking/biking from school	NA	NA	+++	+			WI	---	+
body mass index percentile	-	NA	-	+ee			-?		
likelihood of being overweight or obese	--	NA	-	++ee			-?		
likelihood of being obese	-	NA	-	++ee		+	-?		
likelihood of having poor self-reported health	--	+							

[cc] ???

[dd] Increased transit access and major road exposure may be associated with a more urbanized location with lower socioeconomic status, leading to a worse self-assessment of health.

[ee] Increased transit access may be associated with a more urbanized location with lower socioeconomic status and worse food access, leading to worse body weight and health.

Adult model results stratified by income group

Model results were also generated for the adult cohort after further stratifying each age group by household income group. Income groups were defined as follows:

- Low: < $50,000
- Medium: $50,000 - $100,000
- High: > $100,000

As with the previous section, two versions of the results are provided: one showing results before addressing multicollinearity problems, and a second version after resolving multicollinearity problems. For the version after resolving the multicollinearity problems, two sub-sets of results were generated. The first provides results based on variable selection being conducted once for all income groups pooled together, then the model coefficients were re-fit for the same selection of variables in each income group-specific model. In the second, variable selection was conducted uniquely for each income group. The benefit of conducting variable selection uniquely for each income group is that variables important for only one income group or that have associations in opposite directions for different income groups are more likely to be included in models. The downside is that the sample size for each income group is smaller than for the pooled sample, potentially resulting in variables being excluded solely because of the lower sample size.

In the tables that follow, the following symbols are used to call out important differences across the models:

-- +	Indicates a difference in sign direction across income cohorts.
--	A variable that was added to the model after conducting variable selection uniquely for each income group.
--	A variable that was dropped from the model after conducting variable selection uniquely for each income group, with the symbol indicating the earlier result.
--	A variable that was dropped from the model after conducting variable selection uniquely for each income group, with the symbol indicating the earlier result. The variable was significantly associated ($p<0.05$) with the outcome in the preliminary model, suggesting that it was dropped due to multicollinearity.

Table 43: CHTS adult models.

Outcome	Income cohort	Any trans. walking	Walkability index	Transit access	Rail access	Major road exposure	Regional access	Distance to school	Park access
any walking for transportation	All	NA	+++	+++	+++		+++	---	+++
	Low	NA	+++	++	+		+++	---	+
	Med	NA	+++	+	+++	-	+	---	+++
	High	NA	+++	++	+++		+	---	+++
minutes/day transport walking for those with any	All	NA	+++	WI	++		+		++
	Low	NA	+++	WI	+	+	+		+
	Med	NA	++	WI	+		+++		+
	High	NA	+++	WI	++	-	+	-	+
any biking for transportation	All	NA	+++	++	++	-	WI	---	+++
	Low	NA	++	+	+	--	WI	-	+
	Med	NA	+	+++	+	-	WI	--	++
	High	NA	+++	+	+++	+	WI	---	++
minutes/day transportation biking for those with any	All	NA	--				+++		
	Low	NA	+				++	--	
	Med	NA	-				++		+
	High	NA	--		-		++	++	-
any automobile transportation	All	NA	---	-	---	++	-	+++	+
	Low	NA	---	-	---	++	--	+	+
	Med	NA	---	-	---	+	++	+	+
	High	NA	-	-	---	+	-	+++	+
minutes/day automobile transport for those with any	All	NA	---	-	WI	-	+++	+++	
	Low	NA	--	-	WI	+	+++	+	
	Med	NA	---	-	WI	-	+++	+++	-
	High	NA	---	-	WI	-	+++	++	+

Outcome	Income cohort	Any trans. walking	Walkability index	Transit access	Rail access	Major road exposure	Regional access	Distance to school	Park access
any recreational PA	All	+++	--						+++
	Low	+++	-			+			++
	Med	+++	--	-	+		+		+++
	High	+++	-			--		-	+++
minutes/day of recreational PA for those with any	All	---			+				
	Low	---	++		++				-
	Med	---			-		--		
	High	---	-	+	-	-			

Table 44: CHIS adult models

Outcome	Income cohort	Any trans. walking	MVPA	BMI	Walkability index	Transit access	Rail access	Major road exposure	Regional access	Distance to school	Park access
any walking for transportation	All	NA	NA	NA	+++	+++	+++	---	+++	---	++
	Low	NA	NA	NA	+++	+++	++	---	+++	--	+
	Med	NA	NA	NA	+++	++	++	+	++	-	+
	High	NA	NA	NA	+++	+	++	---	+++	--	+
minutes/wk transport walking for those with any	All	NA	NA	NA	++		+			--	
	Low	NA	NA	NA	++		-			-	
	Med	NA	NA	NA	+		+			--	-
	High	NA	NA	NA	+		++			--	
any walking for recreation	All	+++	NA	NA	-		-	-	++		++
	Low	+++	NA	NA	-		--	-	++	+	+
	Med	+++	NA	NA	-		-	-	-		+
	High	+++	NA	NA	-	-	+	-	+		++
minutes/wk recreational walking for those with any	All	+++	NA	NA				+		--	
	Low	+	NA	NA				+		-	
	Med	+	NA	NA	-		+	+		+	
	High	+++	NA	NA				+		--	
any recreational moderate PA	All	+++	NA	NA	---	WI	WI		WI		
	Low	+++	NA	NA	---	WI	WI		WI	+	
	Med	+++	NA	NA	---	WI	WI		WI		
	High	+++	NA	NA	---	WI	WI		WI	-	
minutes/wk moderate PA for those with any	All	++	NA	NA	---	WI	WI	+	-		-
	Low	+	NA	NA	---	WI	WI	+	+		-
	Med	+	NA	NA	-	WI	WI	+	--		+
	High	+	NA	NA	---	WI	WI	-	-	--	-

Outcome	Income cohort	Any trans. walking	MVPA	BMI	Walkability index	Transit access	Rail access	Major road exposure	Regional access	Distance to school	Park access
any recreational vigorous PA	All	+++	NA	NA	--	Dropped	-		dropped		++
	Low	+++	NA	NA	-	dropped	+		dropped	+	+
	Med	+++	NA	NA	--	dropped	-		dropped		+
	High	+++	NA	NA	-	dropped	-		dropped		++
minutes/wk vigorous PA for those with any	All	-	NA	NA				--	+		
	Low	-	NA	NA	-			-	+		+
	Med	-	NA	NA	++			-	+	+	
	High	+	NA	NA				+	+		
body mass index	All	NA	---	NA	---	WI		++	WI	--	---
	Low	NA	---	NA	---	WI		+	WI	--	-
	Med	NA	---	NA	---	WI	-	+	WI	-	-
	High	NA	---	NA	---	WI	+	++	WI	-	---
likelihood of being overweight or obese	All	NA	---	NA	---	-	dropped	+	WI	--	---
	Low	NA	---	NA	---	-	dropped	-	WI	-	-
	Med	NA	---	NA	---	-	dropped	+	WI	-	--
	High	NA	---	NA	--	-	dropped	++	WI	+	---
likelihood of being obese	All	NA	---	NA	---	-		++	WI	-	-
	Low	NA	---	NA	---	+		+	WI	--	-
	Med	NA	---	NA	--	-	-	+	WI	-	-
	High	NA	---	NA	-	-		++	WI	+	--
likelihood of having high blood pressure	All	NA	---	+++		--				+	-
	Low	NA	-	+++	-	-	++			+	+
	Med	NA	---	+++	++	--				+	-
	High	NA	-	+++		-	--			+	-

Outcome	Income cohort	Any trans. walking	MVPA	BMI	Walkability index	Transit access	Rail access	Major road exposure	Regional access	Distance to school	Park access
likelihood of having heart disease	All	NA	--	+++							-
	Low	NA	--	++				+	-		-
	Med	NA	-	+	+	-	--				-
	High	NA	-	+				-			-
likelihood of having type 2 diabetes	All	NA	--	+++	++				--		--
	Low	NA	--	+++	+		+		-		-
	Med	NA	-	+++	++		-		-		--
	High	NA	-	+++	-		-		-		-
likelihood of having poor self-reported health	All	NA	---	+++	+	+	-	+			--
	Low	NA	---	+++	++	+	-	+			-
	Med	NA	---	+++	-	+	-	+	+		--
	High	NA	---	+++	+	+	-	+	++		-

CHTS survey trip validation with CHTS GPS data

The CHTS included a sub-sample of participants who wore a GPS device for three days. Of the 53,733 CHTS participants included in the final analytical sample, GPS data were available for 5,272 participants, just under 10% of the full sample. GPS devices were not worn by any children in the 5-11 year age group.

CALTRANS cleaned the GPS data and identified discrete trips from the GPS data, including origin, destination, trip duration, and travel mode. Travel mode was not self-reported by participants, but was derived from travel speed, distance, and other characteristics. The table below summarizes the average minutes of daily walking and biking from the travel diary and GPS for GPS sample participants, along with a comparison to travel diary data for non-GPS sample participants.

Table 45. Mean daily minutes of walking and biking for GPS sample participants.

Outcome	GPS sample size	Travel diary		GPS
		Non-GPS sample	GPS sample	
Adult walk minutes / day	4,488	4.6	8.0	8.1
Adult bike minutes / day	4,488	1.1	2.5	1.7
Senior walk minutes / day	559	3.2	6.4	9.6
Teen walk minutes / day	225	7.3	7.6	9.5

In general, GPS-derived walk and bike travel time per day per person was similar between the travel diary and GPS for GPS sample participants. Adult walk time matched nearly exactly, while bike minutes were over-reported by 0.8 minutes/day on the travel diary as compared to the GPS. Seniors under-reported walking on the travel diary by the most, an average of 3.2 minutes less per day than recorded on the GPS device. Teens also under-reported by 1.9 minutes as compared to the GPS.

It is also interesting to note that the self-reported diary walk and bike travel time was much higher for adults and seniors in the GPS sample as compared to those not in the GPS sample. This disparity could be explained by more accurate reporting due to the presence of the GPS device, an added effort to be physically active because of the presence of the GPS device, or self-selection of more active participants being more willing to use the GPS device.

The table below provides a comparison between the distributions of self-reported and GPS-derived walking and biking activity for the GPS sample. In every case, more CHTS participants had any GPS-derived activity than self-reported activity, but the amount of activity per person with any activity was much greater when self-reported.

Table 46. GPS sample distributions of walking and biking activity.

			Outcome distribution for those with any activity					
Outcome	Source	% with any activity	Min	25%	Median	Mean	75%	Max
Adult walking minutes/day	Diary	25%	1	14	24	32.6	43.5	231
	GPS	58%	0.1	2.3	7.8	14	19	215.9
Adult biking minutes/day	Diary	4%	3	26	48	56.5	77	197
	GPS	7%	0.4	7.8	17.5	25.4	34.2	214.7
Senior walking minutes/day	Diary	28%	2	10	20	27.3	30	135
	GPS	62%	0.1	3.8	10.6	15.2	22	92.1
Teen walking minutes/day	Diary	20%	2	11	20.5	31.5	46.8	165
	GPS	56%	0.1	2.2	8.7	17	21.3	151.8

Cross-validation metrics

Validation metrics describing the fit of CHIS and CHTS models are provided below. Each table indicates:

- Sample size used for model fitting
- Mean outcome value for the model sample (either a percentage for binary logistic regression models or a numeric value for linear or Poisson regression models)
- Cross-validation metrics (using 10-fold cross-validation):
    - Accuracy (for binary logistic regression models): the predicted value from a binary logistic regression model is a likelihood probability ranging from 0 to 1. Cross-validation accuracy was calculated by rounding the predicted probability to 0 or 1, comparing to the observed outcome (either 0 or 1), and calculating the percent of the predictions that matched the observations. Note that accuracy depends on the mean outcome value. For example, it is easier to correctly predict an outcome that occurs 99% of the time (i.e. simply guessing that the outcome always occurs would be 99% accurate) rather than an outcome that only occurs 50% of the time (i.e. simply guessing that the outcome always occurs would only be 50% accurate).
    - Mean absolute error (for linear and Poisson regression models): for each prediction, the observed value was subtracted from the predicted value and converted to an absolute value, then the mean value of these absolute differences was calculated. This metric

indicates the magnitude of the average error for each model, and is in the same units as the outcome value.

- Correlation: this was derived by calculating the Pearson's correlation between predicted and observed outcomes. Correlation ranges from -1 to 1, where -1 is perfect negative correlation, 0 is no correlation, and 1 is perfect correlation.

Note: Validation metrics were generated only for the earlier version of the models before applying the corrections for multicollinearity problems. A review of validation metrics for the final models (after implementing the approach for dealing with multicollinearity) indicated that they would not differ substantively from those included in this report.

CHIS models

Table 47: CHIS adult models

			cross-validation metrics		
Outcome	sample size	mean sample outcome	accuracy	mean abs. error	correlation
any walking for transportation	23,515	48%	59%	NA	0.22
minutes/day transport walking for those with any	11,357	112	NA	76	0.20
any walking for recreation	23,515	62%	63%	NA	0.20
minutes/wk recreational walking for those with any	14,578	135	NA	85	0.19
any moderate PA	23,515	57%	60%	NA	0.18
minutes/wk moderate PA for those with any	13,499	191	NA	129	0.18
any vigorous PA	23,515	32%	69%	NA	0.30
minutes/wk vigorous PA for those with any	7,601	180	NA	115	0.15
body mass index	21,763	26.9	NA	4.0	0.39
likelihood of being overweight or obese	21,763	56%	67%	NA	0.37
likelihood of being obese	21,763	23%	77%	NA	0.29
likelihood of having high blood pressure	21,763	26%	77%	NA	0.41
likelihood of having heart disease	21,763	4.8%	95%	NA	0.24
likelihood of having type 2 diabetes	21,763	6.1%	94%	NA	0.32
likelihood of having poor self-reported health	21,763	18%	85%	NA	0.51

Table 48: CHIS senior models

			cross-validation metrics		
outcome	sample size	mean sample outcome	accuracy	mean abs. error	correlation
any walking for transportation	11,618	37%	65%	NA	0.21
minutes/day transport walking for those with any	4,277	120	NA	84	0.17
any walking for recreation	11,618	56%	64%	NA	0.30
minutes/wk recreational walking for those with any	6,453	152	NA	97	0.19
any moderate PA	11,618	57%	62%	NA	0.25
minutes/wk moderate PA for those with any	6,626	225	NA	158	0.14
any vigorous PA	11,618	17%	83%	NA	0.26
minutes/wk vigorous PA for those with any	1,985	186	NA	119	0.15
body mass index	11,375	26.2	NA	3.5	0.39
likelihood of being overweight or obese	11,375	54%	65%	NA	0.32
likelihood of being obese	11,375	19%	81%	NA	0.32
likelihood of having high blood pressure	11,375	59%	63%	NA	0.26
likelihood of having heart disease	11,375	21%	79%	NA	0.26
likelihood of having type 2 diabetes	11,375	15%	85%	NA	0.26
likelihood of having poor self-reported health	11,375	23%	80%	NA	0.47

Table 49: CHIS teen models

			cross-validation metrics		
outcome	sample size	mean sample outcome	accuracy	mean abs. error	correlation
days/week with at least 60 minutes PA	2,367	3.5	NA	1.8	0.28
likelihood of walking/biking from school	2,323	46%	61%	NA	0.28
body mass index percentile	2,220	59%	NA	24%	0.27
likelihood of being overweight or obese	2,220	25%	75%	NA	0.26
likelihood of being obese	2,220	10%	90%	NA	0.20
likelihood of having poor self-reported health	2,220	9.0%	91%	NA	0.25

Table 50: CHIS children models

			cross-validation metrics		
Outcome	sample size	mean sample outcome	accuracy	mean abs. error	correlation
days/week with at least 60 minutes PA	3,117	3.7	NA	1.9	0.33
likelihood of walking/biking from school	3,117	36%	63%	NA	error
body mass index percentile	2,314	62%	NA	29%	0.25
likelihood of being overweight or obese	2,885	38%	65%	NA	error
likelihood of being obese	2,885	22%	76%	NA	error
likelihood of having poor self-reported health	2,885	5.4%	94%	NA	error

CHTS models

Table 51: CHTS adult models

			cross-validation metrics		
outcome	sample size	mean sample outcome	accuracy	mean abs. error	correlation
any walking for transportation	35,695	15%	86%	NA	0.38
minutes/day transport walking for those with any	5,194	34	NA	20	0.28
any biking for transportation	35,695	2%	98%	NA	0.17
minutes/day transportation biking for those with any	801	55	NA	29	0.25
any automobile transportation	35,695	83%	84%	NA	0.32
minutes of auto transportation for those with any	29,644	90	NA	49	0.14
any recreational PA	35,695	17%	83%	NA	0.21
minutes/day of recreational PA for those with any	6,166	104	NA	59	0.26

Table 52: CHTS senior models

			cross-validation metrics		
outcome	sample size	mean sample outcome	accuracy	mean abs. error	correlation
any walking for transportation	8,475	10%	90%	NA	0.34
minutes/day transport walking for those with any	842	34	NA	20	0.29
any automobile transportation	8,475	74%	76%	NA	0.39
minutes of auto transportation for those with any	6,248	82	NA	48	0.17
any recreational PA	8,475	17%	83%	NA	0.26
minutes/day of recreational PA for those with any	1,472	102	NA	63	0.34

Table 53: CHTS teen models

			cross-validation metrics		
outcome	sample size	mean sample outcome	accuracy	mean abs. error	correlation
any walking for transportation	4,734	24%	76%	NA	0.27
minutes/day transport walking for those with any	1,146	30	NA	17	0.32
any automobile transportation	4,734	77%	77%	NA	0.25
minutes of auto transportation for those with any	3,633	62	NA	39	0.13
any recreational PA	4,734	22%	78%	NA	0.17
minutes/day of recreational PA for those with any	1,039	157	NA	81	0.19

Table 54: CHTS children models

			cross-validation metrics		
outcome	sample size	mean sample outcome	accuracy	mean abs. error	correlation
any walking for transportation	4,829	21%	80%	NA	0.32
minutes/day transport walking for those with any	1,029	24	NA	14	0.38
any automobile transportation	4,829	81%	82%	NA	0.32
minutes of auto transportation for those with any	3,903	63	NA	39	0.14
any recreational PA	4,829	25%	75%	NA	0.18
minutes/day of recreational PA for those with any	1,185	140	NA	76	0.19

Adult models by income cohort

Table 55: CHIS adult models by income cohort

				cross-validation metrics		
Outcome	income cohort	sample size	mean sample outcome	accuracy	mean abs. error	correlation
any walking for transportation	low	9,188	53%	60%	NA	0.24
	med	6,537	45%	59%	NA	0.19
	high	7,790	46%	58%	NA	0.18
minutes/day transport walking for those with any	low	4,868	126	NA	84	0.18
	med	2,918	111	NA	78	0.17
	high	3,571	95	NA	65	0.15
any walking for recreation	low	9,188	56%	59%	NA	0.16
	med	6,537	63%	64%	NA	0.20
	high	7,790	68%	68%	NA	0.17
minutes/wk recreational walking for those with any	low	5,143	132	NA	85	0.16
	med	4,111	134	NA	85	0.20
	high	5,324	138	NA	85	0.22
any moderate PA	low	9,188	52%	58%	NA	0.19
	med	6,537	59%	60%	NA	0.14
	high	7,790	62%	63%	NA	0.15
minutes/wk moderate PA for those with any	low	4,820	197	NA	141	0.18
	med	3,846	185	NA	123	0.18
	high	4,833	190	NA	122	0.20
any vigorous PA	low	9,188	25%	76%	NA	0.31
	med	6,537	32%	69%	NA	0.26
	high	7,790	41%	62%	NA	0.23
minutes/wk vigorous PA for those with any	low	2,327	183	NA	122	0.18
	med	2,061	182	NA	118	0.18
	high	3,213	177	NA	107	0.14
body mass index	low	8,152	27.6	NA	4.5	0.37
	med	6,139	26.8	NA	3.9	0.39
	high	7,472	26.1	NA	3.4	0.42
likelihood of being overweight or obese	low	8,152	60%	69%	NA	0.37
	med	6,139	56%	66%	NA	0.35
	high	7,472	52%	67%	NA	0.39
likelihood of being obese	low	8,152	28%	72%	NA	0.28
	med	6,139	24%	76%	NA	0.29
	high	7,472	18%	82%	NA	0.25

				cross-validation metrics		
Outcome	income cohort	sample size	mean sample outcome	accuracy	mean abs. error	correlation
likelihood of having high blood pressure	low	8,152	29%	76%	NA	0.44
	med	6,139	26%	76%	NA	0.39
	high	7,472	23%	79%	NA	0.37
likelihood of having heart disease	low	8,152	5.8%	94%	NA	0.27
	med	6,139	4.2%	96%	NA	0.19
	high	7,472	4.1%	96%	NA	0.19
likelihood of having type 2 diabetes	low	8,152	8.4%	92%	NA	0.34
	med	6,139	5.9%	94%	NA	0.31
	high	7,472	3.8%	96%	NA	0.22
likelihood of having poor self-reported health	low	8,152	33%	74%	NA	0.47
	med	6,139	12%	88%	NA	0.39
	high	7,472	6%	94%	NA	0.31

Table 56: CHTS adult models by income cohort

				cross-validation metrics		
outcome	income cohort	sample size	mean sample outcome	accuracy	mean abs. error	correlation
any walking for transportation	low	10,593	20%	83%	NA	0.44
	med	11,283	11%	89%	NA	0.33
	high	13,819	13%	87%	NA	0.33
minutes/day transport walking for those with any	low	2,085	39	NA	23	0.27
	med	1,283	32	NA	19	0.29
	high	1,826	31	NA	18	0.25
any biking for transportation	low	10,593	2%	98%	NA	0.14
	med	11,283	2%	98%	NA	0.18
	high	13,819	3%	97%	NA	0.18
minutes/day transportation biking for those with any	low	209	51	NA	28	0.27
	med	228	52	NA	28	0.28
	high	364	59	NA	29	0.35
any automobile transportation	low	10,593	74%	77%	NA	0.39
	med	11,283	85%	85%	NA	0.20
	high	13,819	88%	88%	NA	0.18
minutes of automobile transportation for those with any	low	7,820	85	NA	48	0.12
	med	9,630	90	NA	50	0.15
	high	12,194	93	NA	50	0.13
any recreational PA	low	10,593	12%	88%	NA	0.19
	med	11,283	17%	83%	NA	0.19
	high	13,819	21%	79%	NA	0.17
minutes/day of recreational PA for those with any	low	1,289	105	NA	61	0.34
	med	1,959	102	NA	59	0.25
	high	2,918	104	NA	59	0.25

Predictive validation metrics

To simulate the effect of changing built environment characteristics, two predictive modeling scenarios were developed:

1. Base scenario: Variable values were identical to values used for model fitting.
2. Change scenario: All covariates were held constant, while built environment variables were revised to simulate the effect of presumed "healthful" changes in the built environment.

Change in built environment variables was derived by first calculating a 1 decile difference (between the 5th and 6th deciles) for each built environment variable for the CHIS and CHTS samples, then adding (or subtracting) that value from the base built environment variable for each participant. Because this approach did not work for the two dichotomous built environment variables, the difference for these variables was calculated as 1/10th of the percent of participants currently "exposed" to the built environment variable. For example, 9.7% of CHTS participants currently have rail access within 2km, so the change scenario difference was calculated to be 0.97%. The table below summarizes theses values for the CHTS participants.

Table 57: Predictive modeling scenarios

BE variable	Presumed health influence	5th decile value	6th decile value	Percent exposed	Difference
Residential density (units/acre)	+	6.8	7.7	NA	+ 0.9
Dwelling unit count	+	3,247	3,851	NA	+ 604
Retail floor area (sq ft)	+	403,588	567,971	NA	+ 164,383
Non-residential FAR	+	0.14	0.18	NA	+ 0.04
Retail distance (m)	-	150	212	NA	- 62
Restaurant distance (m)	-	300	335	NA	- 35
Land use mix index	+	0.20	0.23	NA	+ 0.04
Intersection density (int/sq mi)	+	86.3	103.3	NA	+ 17.0
Local street length (mi)	+	30.5	32.9	NA	+ 2.5
Transit count	+	12	20	NA	+ 8
Transit distance (m)	-	245	343	NA	- 98
Rail access %	+	NA	NA	9.7%	+ 0.97%
Major street length (mi)	-	0.0	0.1	NA	- 0.1
Major road exposure %	-	NA	NA	21.3%	-2.13%
Park area (acres)	+	13.3	19.7	NA	+ 6.4
Park distance (m)	-	298	393	NA	- 95
Residential accessibility	+	420,631	516,653	NA	+ 96,022
Employment accessibility	+	523,042	703,760	NA	+ 180,718
School distance (m)	-	300	335	NA	- 35

Upon applying the differences to each participant's base built environment variable values, if the application of a "negative" built environment value resulted in a negative value, the change value was recoded to equal zero. For example, a participant with 0 miles of major road within 1km would continue to have 0 miles in the change scenario, rather than -0.1 miles.

The tables below indicate the mean observed outcome, mean base scenario predicted outcome, mean change scenario predicted outcome, and the absolute and percent differences between the base and change predicted outcomes. Because the same data were used for fitting and testing the models, we should expect the mean sample observed outcome and mean base predicted outcomes to be very similar.

Note: Validation metrics were generated only for the earlier version of the models before applying the corrections for multicollinearity problems. A review of validation metrics for the final models (after implementing the approach for dealing with multicollinearity) indicated that they would not differ substantively from those included in this report.

CHIS models

The largest error between the mean sample observed outcome and mean base predicted outcome was 2.6% for adult diabetes, and in the majority of cases, the error was less than 1%.

Table 58: CHIS adult models

Outcome	mean sample observed outcome	mean base predicted outcome	mean change predicted outcome	absolute change (base predicted – change predicted)	% change
minutes of transportation walking (weekly)	54.3	54.3	58.0	3.7	6.8%
minutes of recreational walking (weekly)	83.6	83.7	84.5	0.8	0.9%
minutes of moderate PA (weekly)	109.6	109.2	105.1	-4.1	-3.7%
minutes of vigorous PA (weekly)	58.2	58.0	59.2	1.2	2.0%
body mass index	26.9	26.8	26.7	-0.2	-0.7%
likelihood of being overweight or obese	56.4%	56.4%	54.7%	-1.7%	-3.0%
likelihood of being obese	23.4%	23.3%	22.3%	-1.0%	-4.3%
likelihood of having high blood pressure	25.8%	25.7%	24.9%	-0.8%	-3.0%
likelihood of having heart disease	4.8%	4.7%	4.6%	-0.1%	-2.1%
likelihood of having type 2 diabetes	6.1%	5.9%	5.6%	-0.3%	-5.0%
likelihood of having poor self-reported health	17.8%	17.6%	17.5%	-0.2%	-0.9%

Table 59: CHIS senior models

outcome	mean sample observed outcome	mean base predicted outcome	mean change predicted outcome	absolute change (base predicted – change predicted)	% change
minutes of transportation walking (weekly)	44.1	44.1	46.2	2.1	4.8%
minutes of recreational walking (weekly)	84.6	84.6	88.1	3.4	4.1%
minutes of moderate PA (weekly)	128.4	128.1	124.6	-3.5	-2.7%
minutes of vigorous PA (weekly)	31.7	31.4	31.7	0.3	0.9%
body mass index	26.2	26.2	26.2	-0.1	-0.3%
likelihood of being overweight or obese	54.1%	54.1%	53.3%	-0.8%	-1.4%
likelihood of being obese	19.0%	18.8%	18.4%	-0.5%	-2.5%
likelihood of having high blood pressure	58.9%	59.0%	58.5%	-0.5%	-0.8%
likelihood of having heart disease	21.3%	21.3%	21.0%	-0.3%	-1.2%
likelihood of having type 2 diabetes	15.1%	15.0%	14.5%	-0.5%	-3.5%
likelihood of having poor self-reported health	22.5%	22.3%	22.6%	0.3%	1.4%

Table 60: CHIS teen models

outcome	mean sample observed outcome	mean base predicted outcome	mean change predicted outcome	absolute change (base predicted – change predicted)	% change
days/week with at least 60 minutes PA	3.5	3.5	3.4	-0.1	-1.8%
likelihood of walking/biking from school	46.4%	46.4%	48.8%	2.5%	5.3%
body mass index percentile	58.8%	58.8%	58.4%	-0.4%	-0.6%
likelihood of being overweight or obese	24.8%	24.8%	24.4%	-0.5%	-1.8%
likelihood of being obese	10.5%	10.5%	10.2%	-0.3%	-2.7%
likelihood of having poor self-reported health	9.0%	9.0%	9.8%	0.8%	8.5%

Table 61: CHIS children models

outcome	mean sample observed outcome	mean base predicted outcome	mean change predicted outcome	absolute change (base predicted – change predicted)	% change
days/week with at least 60 minutes PA	3.7	3.7	3.6	-0.1	-1.8%
likelihood of walking/biking from school	36.2%	36.2%	39.4%	3.2%	8.8%
body mass index percentile	61.6%	61.6%	61.4%	-0.2%	-0.3%
likelihood of being overweight or obese	37.8%	37.8%	37.8%	0.0%	0.0%
likelihood of being obese	21.7%	21.7%	21.2%	-0.5%	-2.2%
likelihood of having poor self-reported health	3.9%	3.9%	3.9%	0.0%	0.5%

CHTS models

Table 62: CHTS adult models

outcome	mean sample observed outcome	mean base predicted outcome	mean change predicted outcome	absolute change (base predicted – change predicted)	% change
minutes of transportation walking (daily)	5.0	5.0	6.0	1.0	19.1%
minutes of transportation biking (daily)	1.2	1.1	1.3	0.2	17.8%
minutes of automobile transportation (daily)	74.9	75.2	74.3	-0.9	-1.2%
minutes of recreational PA (daily)	17.9	17.9	18.4	0.5	2.7%

Table 63: CHTS senior models

outcome	mean sample observed outcome	mean base predicted outcome	mean change predicted outcome	absolute change (base predicted – change predicted)	% change
minutes of transportation walking (daily)	3.4	3.4	4.2	0.8	22.1%
minutes of automobile transportation (daily)	60.6	60.8	59.5	-1.2	-2.1%
minutes of recreational PA (daily)	17.8	18.2	19.7	1.4	7.8%

Table 64: CHTS teen models

outcome	mean sample observed outcome	mean base predicted outcome	mean change predicted outcome	absolute change	% change
minutes of transportation walking (daily)	7.3	7.3	8.6	1.3	17.6%
minutes of automobile transportation (daily)	47.9	48.2	47.0	-1.1	-2.4%
minutes of recreational PA (daily)	34.5	34.6	35.3	0.8	2.2%

Table 65: CHTS children models

outcome	mean sample observed outcome	mean base predicted outcome	mean change predicted outcome	absolute change (base predicted – change predicted)	% change
minutes of transportation walking (daily)	5.1	5.2	5.9	0.7	13.4%
minutes of automobile transportation (daily)	50.6	50.7	50.3	-0.4	-0.8%
minutes of recreational PA (daily)	34.3	34.2	33.0	-1.1	-3.3%

Adult models by income cohort

Table 66: CHIS adult models by income cohort

outcome	income cohort	mean sample observed outcome	mean base predicted outcome	mean change predicted outcome	absolute change	% change
minutes of transportation walking (weekly)	low	66.6	66.7	71.1	4.5	6.7%
minutes of transportation walking (weekly)	med	49.7	49.4	52.7	3.3	6.6%
minutes of transportation walking (weekly)	high	43.6	43.5	46.9	3.4	7.8%
minutes of recreational walking (weekly)	low	74.0	74.0	74.8	0.8	1.1%
minutes of recreational walking (weekly)	med	84.5	84.7	84.8	0.2	0.2%
minutes of recreational walking (weekly)	high	94.2	94.5	95.6	1.1	1.2%
minutes of moderate PA (weekly)	low	103.3	102.4	99.0	-3.4	-3.3%
minutes of moderate PA (weekly)	med	108.9	108.6	103.5	-5.1	-4.7%
minutes of moderate PA (weekly)	high	117.6	116.7	112.5	-4.2	-3.6%
minutes of vigorous PA (weekly)	low	46.4	46.3	46.9	0.6	1.3%
minutes of vigorous PA (weekly)	med	57.3	57.1	57.7	0.6	1.1%
minutes of vigorous PA (weekly)	high	73.0	72.8	74.8	2.0	2.8%
body mass index	low	27.6	27.6	27.5	-0.1	-0.4%
body mass index	med	26.8	26.8	26.7	-0.2	-0.6%
body mass index	high	26.1	26.1	25.9	-0.2	-0.8%
likelihood of being overweight or obese	low	60.3%	60.3%	59.4%	-1.0%	-1.6%
likelihood of being overweight or obese	med	56.3%	56.3%	54.4%	-1.8%	-3.3%
likelihood of being overweight or obese	high	52.2%	52.2%	50.1%	-2.1%	-4.0%
likelihood of being obese	low	28.0%	27.9%	27.2%	-0.8%	-2.8%
likelihood of being obese	med	23.9%	23.9%	23.0%	-0.8%	-3.5%
likelihood of being obese	high	18.1%	18.0%	16.8%	-1.2%	-6.7%
likelihood of having high blood pressure	low	28.8%	28.6%	28.2%	-0.4%	-1.5%
likelihood of having high blood pressure	med	25.8%	25.8%	24.9%	-0.9%	-3.4%
likelihood of having high blood pressure	high	22.6%	22.6%	21.6%	-1.0%	-4.4%
likelihood of having heart disease	low	5.8%	5.7%	5.6%	-0.1%	-1.7%
likelihood of having heart disease	med	4.2%	4.1%	4.0%	-0.1%	-2.6%
likelihood of having heart disease	high	4.1%	4.0%	4.0%	-0.1%	-1.5%
likelihood of having type 2 diabetes	low	8.4%	8.2%	7.9%	-0.3%	-3.4%
likelihood of having type 2 diabetes	med	5.9%	5.9%	5.6%	-0.3%	-4.6%
likelihood of having type 2 diabetes	high	3.8%	3.7%	3.4%	-0.3%	-8.1%
likelihood of having poor self-reported health	low	32.5%	32.4%	32.4%	0.0%	-0.1%
likelihood of having poor self-reported health	med	12.4%	12.4%	11.9%	-0.5%	-3.8%
likelihood of having poor self-reported health	high	6.2%	6.2%	6.1%	-0.1%	-1.3%

Table 67: CHTS adult models by income cohort

outcome	income cohort	mean sample observed outcome	mean base predicted outcome	mean change predicted outcome	absolute change	% change
minutes of transportation walking (daily)	low	7.7	7.7	8.9	1.2	15.0%
minutes of transportation walking (daily)	med	3.6	3.7	4.5	0.9	24.4%
minutes of transportation walking (daily)	high	4.1	4.1	4.9	0.8	20.5%
minutes of transportation biking (daily)	low	1.0	0.9	1.1	0.1	12.7%
minutes of transportation biking (daily)	med	1.0	1.0	1.2	0.2	24.3%
minutes of transportation biking (daily)	high	1.6	1.4	1.7	0.2	15.9%
minutes of automobile transportation (daily)	low	63.0	63.2	62.5	-0.7	-1.1%
minutes of automobile transportation (daily)	med	76.9	77.2	76.6	-0.6	-0.7%
minutes of automobile transportation (daily)	high	82.4	82.6	81.4	-1.2	-1.5%
minutes of recreational PA (daily)	low	12.8	12.7	13.2	0.5	4.0%
minutes of recreational PA (daily)	med	17.7	17.7	18.0	0.4	2.0%
minutes of recreational PA (daily)	high	22.0	22.1	22.6	0.5	2.3%

Aggregation scale validation metrics

Validation metrics in this section are intended to demonstrate the improvement in predictive accuracy obtained by aggregating predictions to larger spatial units, where model errors increasingly cancel one another out. This test was accomplished by first applying the draft CHIS and CHTS models at the grid cell level (as will be done in UrbanFootprint) for every grid cell containing a CHIS or CHTS participant, then aggregating the resulting predictions to a variety of spatial scales (Census tracts, zip codes, and counties).

While this test demonstrates the relative gain in predictive accuracy by aggregating grid-level predictions to larger spatial units, these validation metrics are not an indicator of the expected predictive accuracy of the models when applied to actual data for these same spatial units. For example, typical census tracts contain about 4,000 people. In the CHTS data, the average Census tract contains only 9 people with a maximum of 109. Conducting validation tests on spatial units with such low population sizes is subject to a high degree of random variation. In contrast, in the CHTS data, the average county contains 1,791 people, with a maximum of 12,798. Based on the available sample sizes, our county-level predictions are most equivalent to the accuracy we might expect from applying the models at the Census block group level, which range in size from 600-3,000 people. Descriptive statistics for CHTS participants within a variety of spatial units is provided below.

Table 68: Descriptive statistics -- CHTS participants, by spatial units

Spatial unit	# of units in CHTS data	Average unit population
Grid cell	25,055	2.1
Census tract	5,705	9.4
Zip code	1,190	45.2
County	30	1791.1

The tables below contain three different metrics for each model, and were calculate at the UrbanFootprint grid cell, Census tract, zip code, and county levels:

- Predicted correlation: this was derived by calculating the Pearson's correlation between predicted and observed outcomes at the spatial unit level. Correlation ranges from -1 to 1, where -1 is perfect negative correlation, 0 is no correlation, and 1 is perfect correlation.
- Mean absolute error: for each spatial unit, the mean observed value was subtracted from the mean predicted value and converted to an absolute value, then the mean value of these absolute differences was calculated. This metric indicates the magnitude of the average error for each model, and is in the same units as the outcome value.
- Mean absolute error/mean outcome: to allow for easier comparison across different models, this metric divides the mean absolute error by the mean outcome value. This can also be thought of as the mean absolute error percentage for each outcome.

Note: Validation metrics were generated only for the earlier version of the models before applying the corrections for multicollinearity problems. A review of validation metrics for the final models (after implementing the approach for dealing with multicollinearity) indicated that they would not differ substantively from those included in this report.

CHIS models

Table 69: CHIS adult models

	Predicted correlation				Mean absolute error				Mean absolute error / mean outcome			
outcome	grid	tract	zip	county	grid	tract	zip	county	grid	tract	zip	county
minutes of transportation walking	0.15	0.21	0.29	0.51	62.2	37.9	19.9	4.7	1.15	0.70	0.37	0.09
minutes of recreational walking	0.19	0.21	0.28	0.85	80.4	43.4	21.9	4.7	0.96	0.52	0.26	0.06
minutes of moderate PA	0.16	0.20	0.23	0.20	116.2	65.5	33.4	5.9	1.06	0.60	0.30	0.05
minutes of vigorous PA	0.21	0.23	0.27	0.77	74.6	43.0	22.4	3.9	1.28	0.74	0.38	0.07
body mass index	0.37	0.40	0.53	0.94	3.9	2.1	1.0	0.3	0.15	0.08	0.04	0.01
likelihood of being overweight or obese	0.38	0.41	0.56	0.68	40.8%	18.3%	8.6%	2.0%	0.72	0.32	0.15	0.04
likelihood of being obese	0.29	0.32	0.45	0.92	31.9%	16.4%	7.6%	1.6%	1.36	0.70	0.33	0.07
likelihood of having high blood pressure	0.41	0.44	0.44	0.40	31.1%	15.4%	6.8%	1.5%	1.20	0.59	0.26	0.06
likelihood of having heart disease	0.24	0.27	0.33	0.36	8.4%	6.6%	3.5%	0.6%	1.77	1.38	0.73	0.12
likelihood of having type 2 diabetes	0.32	0.33	0.35	0.86	10.0%	7.4%	3.9%	0.8%	1.64	1.22	0.63	0.13
likelihood of having poor self-reported health	0.52	0.57	0.70	0.92	20.8%	12.3%	5.6%	1.3%	1.17	0.69	0.32	0.07

Table 70: CHIS senior models

	Predicted correlation				Mean absolute error				Mean absolute error / mean outcome			
outcome	grid	tract	zip	county	grid	tract	zip	county	grid	tract	zip	county
minutes of transportation walking	0.16	0.19	0.18	0.84	57.0	39.7	23.0	5.3	1.29	0.90	0.52	0.12
minutes of recreational walking	0.23	0.23	0.21	0.72	85.5	55.7	30.9	6.9	1.01	0.66	0.36	0.08
minutes of moderate PA	0.18	0.21	0.21	0.41	138.3	90.1	50.8	9.4	1.08	0.70	0.40	0.07
minutes of vigorous PA	0.17	0.16	0.08	0.78	48.4	34.1	21.3	3.3	1.53	1.07	0.67	0.11
body mass index	0.39	0.40	0.44	0.53	3.4	2.1	1.1	0.2	0.13	0.08	0.04	0.01
likelihood of being overweight or obese	0.33	0.34	0.38	0.62	42.5%	22.8%	10.9%	1.9%	0.79	0.42	0.20	0.04
likelihood of being obese	0.33	0.33	0.37	0.82	26.7%	17.5%	8.9%	1.6%	1.41	0.92	0.47	0.08
likelihood of having high blood pressure	0.27	0.29	0.31	0.64	43.1%	22.8%	10.6%	1.8%	0.73	0.39	0.18	0.03
likelihood of having heart disease	0.26	0.27	0.26	0.23	30.3%	18.3%	9.1%	1.5%	1.42	0.86	0.43	0.07
likelihood of having type 2 diabetes	0.27	0.28	0.32	0.54	23.2%	16.1%	8.1%	1.6%	1.54	1.07	0.53	0.10
likelihood of having poor self-reported health	0.48	0.52	0.55	0.91	25.9%	16.5%	8.0%	1.3%	1.15	0.73	0.36	0.06

Table 71: CHIS teen models

	Predicted correlation				Mean absolute error				Mean absolute error / mean outcome			
outcome	grid	tract	zip	county	grid	tract	zip	county	grid	tract	zip	county
days/week with at least 60 minutes PA	0.28	0.30	0.27	0.43	1.8	1.5	1.0	0.1	0.52	0.43	0.28	0.04
likelihood of walking/biking from school	0.31	0.32	0.38	0.74	44.9%	36.9%	23.4%	3.5%	0.97	0.80	0.51	0.08
body mass index percentile	0.27	0.28	0.35	0.75	23.7%	19.7%	12.7%	2.2%	0.40	0.34	0.22	0.04
likelihood of being overweight or obese	0.29	0.30	0.36	0.78	34.0%	29.5%	18.8%	3.7%	1.37	1.19	0.76	0.15
likelihood of being obese	0.23	0.25	0.29	0.67	17.7%	16.5%	12.5%	1.9%	1.69	1.58	1.19	0.19
likelihood of having poor self-reported health	0.29	0.30	0.37	0.59	14.9%	14.1%	11.0%	2.2%	1.66	1.57	1.22	0.24

Table 72: CHIS children models

	Predicted correlation				Mean absolute error				Mean absolute error / mean outcome			
outcome	grid	tract	zip	county	grid	tract	zip	county	grid	tract	zip	county
days/week with at least 60 minutes PA	0.33	0.35	0.41	0.47	1.9	1.6	0.9	0.1	0.52	0.42	0.25	0.04
likelihood of walking/biking from school	0.28	0.30	0.35	0.69	42.0%	33.8%	20.0%	4.4%	1.16	0.93	0.55	0.12
body mass index percentile	0.24	0.26	0.26	0.62	28.4%	23.6%	15.7%	3.5%	0.46	0.38	0.25	0.06
likelihood of being overweight or obese	0.30	0.32	0.33	0.73	42.6%	34.3%	21.9%	4.3%	1.13	0.91	0.58	0.11
likelihood of being obese	0.32	0.35	0.36	0.65	30.5%	26.0%	18.2%	3.3%	1.40	1.20	0.84	0.15
likelihood of having poor self-reported health	0.40	0.42	0.40	0.71	6.4%	6.2%	5.4%	1.2%	1.62	1.57	1.38	0.31

CHTS models

Table 73: CHTS adult models

	Predicted correlation				Mean absolute error				Mean absolute error / mean outcome			
Outcome	grid	tract	zip	county	grid	tract	zip	county	grid	tract	zip	county
minutes of transportation walking (daily)	0.39	0.48	0.65	0.97	6.5	4.2	2.2	0.8	1.29	0.85	0.45	0.17
minutes of transportation biking (daily)	0.14	0.16	0.22	0.80	2.2	1.8	1.2	0.3	1.76	1.46	1.00	0.24
minutes of automobile transportation (daily)	0.20	0.26	0.28	0.76	45.3	24.0	12.1	2.6	0.61	0.32	0.16	0.03
minutes of recreational PA (daily)	0.12	0.15	0.29	0.69	26.0	15.5	8.4	1.8	1.45	0.86	0.47	0.10

Table 74: CHTS senior models

	Predicted correlation				Mean absolute error				Mean absolute error / mean outcome			
outcome	grid	tract	zip	county	grid	tract	zip	county	grid	tract	zip	county
minutes of transportation walking (daily)	0.33	0.40	0.50	0.94	5.2	4.4	3.0	0.7	1.53	1.30	0.89	0.21
minutes of automobile transportation (daily)	0.26	0.28	0.37	0.59	46.7	34.0	19.8	3.8	0.77	0.56	0.33	0.06
minutes of recreational PA (daily)	0.16	0.15	0.24	0.29	27.3	21.6	14.3	3.6	1.53	1.21	0.80	0.20

Table 75: CHTS teen models

	Predicted correlation				Mean absolute error				Mean absolute error / mean outcome			
outcome	grid	tract	zip	county	grid	tract	zip	county	grid	tract	zip	county
minutes of transportation walking (daily)	0.27	0.29	0.36	0.57	9.7	8.3	5.7	1.4	1.34	1.14	0.79	0.19
minutes of automobile transportation (daily)	0.12	0.09	0.18	0.30	39.3	32.4	22.1	6.0	0.82	0.68	0.46	0.13
minutes of recreational PA (daily)	0.13	0.16	0.17	0.55	49.9	41.8	28.4	4.3	1.45	1.21	0.82	0.12

Table 76: CHTS children models

	Predicted correlation				Mean absolute error				Mean absolute error / mean outcome			
outcome	grid	tract	zip	county	grid	tract	zip	county	grid	tract	zip	county
minutes of transportation walking (daily)	0.34	0.35	0.44	0.60	7.0	6.1	4.3	1.0	1.35	1.18	0.84	0.20
minutes of automobile transportation (daily)	0.15	0.14	0.17	0.53	40.6	33.7	23.4	5.3	0.80	0.67	0.46	0.11
minutes of recreational PA (daily)	0.12	0.12	0.11	0.56	49.2	42.3	29.5	6.6	1.43	1.23	0.86	0.19

Adult models by income cohort

Table 77: CHIS adult models by income cohort

		Predicted correlation				Mean absolute error				Mean absolute error / mean outcome			
Outcome	income cohort	grid	tract	zip	county	grid	tract	zip	county	grid	tract	zip	county
minutes of transportation walking	low	0.16	0.17	0.20	0.18	72.3	53.1	29.6	6.9	1.09	0.80	0.44	0.10
minutes of transportation walking	med	0.10	0.12	0.13	0.82	61.4	48.9	31.4	5.6	1.24	0.98	0.63	0.11
minutes of transportation walking	high	0.12	0.15	0.14	0.34	52.5	37.1	23.2	6.6	1.20	0.85	0.53	0.15
minutes of recreational walking	low	0.15	0.15	0.16	0.26	77.6	54.9	30.6	6.4	1.05	0.74	0.41	0.09
minutes of recreational walking	med	0.21	0.21	0.28	0.41	82.4	61.4	36.8	5.7	0.98	0.73	0.44	0.07
minutes of recreational walking	high	0.21	0.21	0.18	0.61	85.2	55.6	31.6	4.4	0.90	0.59	0.34	0.05
minutes of moderate PA	low	0.17	0.21	0.31	0.65	119.0	87.5	50.7	8.2	1.15	0.85	0.49	0.08
minutes of moderate PA	med	0.17	0.18	0.23	0.14	114.9	86.5	53.3	8.5	1.05	0.79	0.49	0.08
minutes of moderate PA	high	0.18	0.20	0.17	0.32	117.6	77.8	46.2	8.4	1.00	0.66	0.39	0.07
minutes of vigorous PA	low	0.21	0.18	0.11	0.23	64.2	49.4	29.0	5.6	1.38	1.06	0.63	0.12
minutes of vigorous PA	med	0.21	0.22	0.28	0.44	76.5	59.3	36.5	8.3	1.34	1.04	0.64	0.14
minutes of vigorous PA	high	0.18	0.20	0.22	0.20	87.2	57.3	35.1	8.0	1.20	0.78	0.48	0.11
body mass index	low	0.35	0.37	0.42	0.86	4.5	3.1	1.5	0.3	0.16	0.11	0.06	0.01
body mass index	med	0.37	0.38	0.44	0.37	3.9	2.9	1.6	0.3	0.15	0.11	0.06	0.01
body mass index	high	0.39	0.41	0.46	0.95	3.4	2.2	1.3	0.3	0.13	0.09	0.05	0.01
likelihood of being overweight or obese	low	0.38	0.38	0.45	0.80	40.3%	25.2%	12.0%	2.0%	0.67	0.42	0.20	0.03
likelihood of being overweight or obese	med	0.37	0.37	0.45	0.21	42.2%	27.9%	14.6%	2.7%	0.75	0.50	0.26	0.05
likelihood of being overweight or obese	high	0.40	0.40	0.44	0.97	41.5%	23.7%	12.4%	2.2%	0.79	0.45	0.24	0.04
likelihood of being obese	low	0.29	0.30	0.34	0.85	36.1%	23.7%	11.6%	1.9%	1.29	0.84	0.41	0.07
likelihood of being obese	med	0.30	0.30	0.33	0.26	32.8%	24.1%	12.5%	2.6%	1.37	1.01	0.52	0.11
likelihood of being obese	high	0.27	0.29	0.34	0.87	27.2%	18.8%	10.1%	1.7%	1.51	1.04	0.56	0.09
likelihood of having high blood pressure	low	0.45	0.46	0.38	0.61	32.1%	21.6%	10.7%	2.6%	1.11	0.75	0.37	0.09
likelihood of having high blood pressure	med	0.40	0.41	0.47	0.67	32.0%	23.2%	12.3%	2.0%	1.24	0.90	0.48	0.08
likelihood of having high blood pressure	high	0.38	0.39	0.42	0.35	29.8%	19.3%	9.9%	1.8%	1.32	0.85	0.44	0.08

		Predicted correlation				Mean absolute error				Mean absolute error / mean outcome			
Outcome	income cohort	grid	tract	zip	county	grid	tract	zip	county	grid	tract	zip	county
likelihood of having heart disease	low	0.29	0.29	0.24	0.60	10.0%	8.7%	5.5%	0.9%	1.71	1.49	0.94	0.16
likelihood of having heart disease	med	0.22	0.21	0.23	0.19	7.6%	7.0%	5.2%	1.0%	1.81	1.68	1.26	0.23
likelihood of having heart disease	high	0.21	0.21	0.35	0.08	7.4%	6.6%	4.8%	0.7%	1.83	1.61	1.18	0.18
likelihood of having type 2 diabetes	low	0.34	0.32	0.33	0.11	13.2%	10.8%	6.4%	1.3%	1.57	1.28	0.76	0.15
likelihood of having type 2 diabetes	med	0.33	0.38	0.35	0.60	9.8%	8.8%	6.6%	1.4%	1.66	1.50	1.11	0.23
likelihood of having type 2 diabetes	high	0.25	0.27	0.40	0.96	6.7%	6.1%	4.6%	1.1%	1.78	1.62	1.22	0.29
likelihood of having poor self-reported health	low	0.48	0.50	0.54	0.70	33.0%	22.1%	11.0%	2.1%	1.02	0.68	0.34	0.07
likelihood of having poor self-reported health	med	0.41	0.43	0.53	0.64	17.9%	14.9%	8.7%	1.8%	1.44	1.20	0.70	0.15
likelihood of having poor self-reported health	high	0.33	0.37	0.37	0.72	10.2%	8.7%	5.8%	1.2%	1.65	1.41	0.93	0.20

Table 78: CHTS adult models by income cohort

		Predicted correlation				Mean absolute error				Mean absolute error / mean outcome			
outcome	income cohort	grid	tract	zip	county	grid	tract	zip	county	grid	tract	zip	county
minutes of transportation walking (daily)	low	0.42	0.44	0.54	0.94	9.0	7.0	4.2	0.9	1.17	0.90	0.54	0.12
minutes of transportation walking (daily)	med	0.36	0.40	0.55	0.97	5.1	4.4	3.0	1.0	1.44	1.22	0.84	0.27
minutes of transportation walking (daily)	high	0.36	0.41	0.51	0.95	5.5	4.2	2.6	1.1	1.37	1.03	0.64	0.26
minutes of transportation biking (daily)	low	0.13	0.13	0.14	0.64	1.8	1.7	1.4	0.3	1.81	1.68	1.36	0.30
minutes of transportation biking (daily)	med	0.16	0.19	0.12	0.79	1.9	1.7	1.4	0.4	1.77	1.65	1.35	0.34
minutes of transportation biking (daily)	high	0.14	0.14	0.11	0.54	2.7	2.4	1.8	0.5	1.75	1.54	1.16	0.30
minutes of automobile transportation (daily)	low	0.24	0.24	0.27	0.66	43.9	31.3	18.8	4.3	0.70	0.50	0.30	0.07
minutes of automobile transportation (daily)	med	0.15	0.16	0.15	0.74	46.7	33.7	20.2	3.3	0.61	0.44	0.26	0.04
minutes of automobile transportation (daily)	high	0.14	0.18	0.19	0.17	45.8	30.4	17.7	3.3	0.56	0.37	0.22	0.04
minutes of recreational PA (daily)	low	0.12	0.11	0.17	0.37	20.5	16.5	11.0	1.7	1.61	1.30	0.86	0.13
minutes of recreational PA (daily)	med	0.11	0.10	0.14	0.33	26.3	20.5	12.5	2.3	1.49	1.16	0.71	0.13
minutes of recreational PA (daily)	high	0.08	0.09	0.17	0.54	30.2	21.3	13.0	3.0	1.37	0.97	0.59	0.14

CHIS cross-year validation metrics

The draft CHIS models, which were developed using 2009 CHIS data, were applied to 2005 CHIS data to validate the results with a second CHIS data set that was not used for model fitting. In terms of the variables used in the draft CHIS models, the 2005 CHIS data set is nearly identical. One exception is that data were not collected in 2005 on children's physical activity participation. For all other CHIS model outcomes, cross-year validation metrics were calculated.

To do so, the 2005 CHIS data were first developed identically to the 2009 CHIS data. All variables were coded identically, exclusion criteria were applied identically, and 2005 CHIS participants were matched to UF grid cells to derive built environment characteristics for each participant. Note that built environment variables used for 2005 CHIS participants are from the same UF base year as used to develop the models - built environment data are not available for earlier years. This will likely lead to increased error for 2005 predictions.

The draft CHIS regression models were then applied to the 2005 data set to calculate predicted outcome values for every participant. The tables below indicate the mean observed outcome, mean predicted outcome, and percent error between the mean observed outcome and mean predicted outcome. Aggregation scale validation metrics were also calculated at the UrbanFootprint grid cell, Census tract, zip code, and county levels:

- Predicted correlation: this was derived by calculating the Pearson's correlation between predicted and observed outcomes at the spatial unit level. Correlation ranges from -1 to 1, where -1 is perfect negative correlation, 0 is no correlation, and 1 is perfect correlation.
- Mean absolute error: for each spatial unit, the mean observed value was subtracted from the mean predicted value and converted to an absolute value, then the mean value of these absolute differences was calculated. This metric indicates the magnitude of the average error for each model, and is in the same units as the outcome value.
- Mean absolute error/mean outcome: to allow for easier comparison across different models, this metric divides the mean absolute error by the mean outcome value. This can also be thought of as the mean absolute error percentage for each outcome.

Because the 2005 data set was not used for fitting the models, we should expect the prediction error to be higher when applying the models to the 2005 data set, in comparison to applying the models to the 2009 data set. Temporal mismatches (e.g. differences in external conditions/events between 2005 and 2009 that may affect CHIS covariate and outcome values; and between 2005 CHIS data collection and more recent UF built environment data collection) both should also contribute to increased error when applying the models to 2005 CHIS data.

Note: Validation metrics were generated only for the earlier version of the models before applying the corrections for multicollinearity problems. A review of validation metrics for the final models (after implementing the approach for dealing with multicollinearity) indicated that they would not differ substantively from those included in this report.

CHIS models

Results indicate more prediction error when applying the models (fit with 2009 data) to the 2005 data set than when applied to the 2009 data. The largest error was a 17% under-prediction for adult transportation walking, though in the majority of cases, error was less than 5%. For comparison, the largest error when applying the models to the 2009 data set was 2.6% for adult diabetes, and in the majority of cases, the error was less than 1%. Mean absolute error / mean outcome ratio was marginally higher when applying the models to the 2005 data set (~0.1 on average) than to the 2009 data set (~0.07 on average).

Table 79: CHIS adult models

Outcome	Mean 2005 sample observed outcome	Mean 2005 predicted outcome	% error	Predicted correlation				Mean absolute error				Mean absolute error / mean outcome			
				grid	tract	zip	county	grid	tract	zip	county	grid	tract	zip	county
minutes of transportation walking	72.8	60.7	-17%	0.11	0.15	0.15	0.41	79.7	50.7	27.8	12.2	1.10	0.70	0.38	0.17
minutes of recreational walking	79.5	81.4	2%	0.15	0.17	0.32	0.76	79.8	42.6	19.8	4.9	1.00	0.54	0.25	0.06
minutes of moderate PA	111.5	109.0	-2%	0.15	0.19	0.32	0.55	118.2	66.8	34.6	7.9	1.06	0.60	0.31	0.07
minutes of vigorous PA	55.4	59.5	7%	0.20	0.21	0.25	0.50	73.7	43.3	22.2	6.2	1.33	0.78	0.40	0.11
body mass index	26.6	26.8	1%	0.36	0.40	0.50	0.93	3.8	2.1	1.0	0.3	0.14	0.08	0.04	0.01
likelihood of being overweight or obese	54.9%	55.7%	1%	0.36	0.37	0.47	0.89	41.0%	19.0%	8.6%	2.4%	0.75	0.35	0.16	0.04
likelihood of being obese	22.0%	23.1%	5%	0.28	0.32	0.42	0.90	31.2%	16.1%	7.5%	2.3%	1.42	0.73	0.34	0.10
likelihood of having high blood pressure	23.0%	22.6%	-2%	0.33	0.34	0.32	0.03	8.3%	6.5%	3.3%	0.6%	1.65	1.30	0.65	0.13
likelihood of having heart disease	4.6%	4.0%	-14%	0.51	0.53	0.64	0.96	19.4%	11.8%	5.1%	1.2%	1.22	0.74	0.32	0.07
likelihood of having type 2 diabetes	5.0%	4.8%	-5%	0.23	0.22	0.22	0.08	7.7%	6.0%	3.3%	0.8%	1.67	1.32	0.72	0.17
likelihood of having poor self-report health	15.9%	15.9%	0%	0.42	0.44	0.43	0.61	28.6%	14.5%	6.3%	1.0%	1.24	0.63	0.27	0.04

Table 80: CHIS senior models

Outcome	mean sample observed outcome	mean predicted outcome	% error	Predicted correlation				Mean absolute error				Mean absolute error / mean outcome			
				grid	tract	zip	county	grid	tract	zip	county	grid	tract	zip	county
minutes of transportation walking	55.7	46.6	-16%	0.13	0.17	0.18	0.58	67.2	50.5	34.3	10.9	1.21	0.91	0.62	0.20
minutes of recreational walking	84.2	84.5	0%	0.21	0.19	0.18	0.18	88.8	61.3	36.1	8.4	1.06	0.73	0.43	0.10
minutes of moderate PA	148.1	133.7	-10%	0.16	0.20	0.23	0.87	158.1	115.9	75.5	20.8	1.07	0.78	0.51	0.14
minutes of vigorous PA	31.7	29.9	-6%	0.16	0.15	0.12	0.47	48.9	37.3	26.0	4.7	1.54	1.18	0.82	0.15
body mass index	26.1	26.3	1%	0.37	0.39	0.43	0.82	3.3	2.3	1.3	0.4	0.13	0.09	0.05	0.01
likelihood of being overweight or obese	53.5%	55.2%	3%	0.31	0.34	0.35	0.80	43.1%	26.6%	13.4%	3.0%	0.81	0.50	0.25	0.06
likelihood of being obese	17.5%	19.6%	12%	0.29	0.32	0.36	0.76	26.9%	19.0%	10.4%	2.7%	1.54	1.09	0.60	0.16
likelihood of having high blood pressure	58.2%	59.0%	1%	0.27	0.26	0.28	0.38	43.6%	26.0%	13.6%	3.0%	0.75	0.45	0.23	0.05
likelihood of having heart disease	22.0%	20.8%	-5%	0.22	0.22	0.15	0.46	30.9%	21.0%	11.6%	2.6%	1.41	0.96	0.53	0.12
likelihood of having type 2 diabetes	13.1%	14.8%	13%	0.27	0.27	0.29	0.59	22.1%	17.1%	9.8%	2.2%	1.68	1.30	0.75	0.17
likelihood of having poor self-report health	26.3%	23.0%	-13%	0.48	0.51	0.54	0.70	28.0%	20.1%	11.1%	3.6%	1.06	0.76	0.42	0.14

Table 81: CHIS teen models

outcome	mean sample observed outcome	mean predicted outcome	% error	Predicted correlation				Mean absolute error				Mean absolute error / mean outcome			
				grid	tract	zip	county	Grid	tract	zip	county	grid	tract	zip	county
days/week with at least 60 minutes PA	3.7	3.5	-6%	0.18	0.20	0.13	0.50	1.9	1.6	1.0	0.3	0.52	0.42	0.27	0.08
body mass index percentile	60.3%	59.5%	-1%	0.21	0.22	0.26	0.68	23.9%	19.6%	12.2%	1.8%	0.40	0.33	0.20	0.03
likelihood of being overweight or obese	26.2%	25.2%	-4%	0.23	0.23	0.25	0.72	35.5%	30.2%	18.9%	3.5%	1.36	1.16	0.72	0.13
likelihood of being obese	11.6%	10.3%	-11%	0.19	0.18	0.19	0.42	18.6%	17.3%	12.9%	3.4%	1.61	1.50	1.11	0.30
likelihood of having poor self-reported health	10.1%	8.4%	-16%	0.22	0.24	0.29	0.52	15.7%	14.5%	11.1%	2.7%	1.56	1.44	1.11	0.26

Table 82: CHIS child models

outcome	mean sample observed outcome	mean predicted outcome	% error	Predicted correlation				Mean absolute error				Mean absolute error / mean outcome			
				grid	tract	zip	county	Grid	tract	zip	county	grid	tract	zip	county
body mass index percentile	63.8%	63.2%	-1%	0.13	0.13	0.20	0.63	29.4%	23.3%	14.0%	2.9%	0.46	0.37	0.22	0.05
likelihood of being overweight or obese	model error – results could not be generated														
likelihood of being obese	model error – results could not be generated														
likelihood of having poor self-reported health	3.8%	3.5%	-8%	0.16	0.11	0.22	0.48	6.7%	6.4%	5.6%	1.2%	1.75	1.68	1.46	0.31

Adult models by income cohort

Table 83: CHIS adult models by income cohort

					Predicted correlation				Mean absolute error				Mean absolute error / mean outcome			
outcome	income cohort	mean sample observed outcome	mean base predicted outcome	% error	grid	tract	zip	county	grid	tract	zip	county	grid	tract	zip	county
minutes of transportation walking	low	89.8	75.2	-16%	0.15	0.17	0.26	0.52	123.5	82.9	51.7	12.1	1.01	0.68	0.42	0.10
	med	66.8	53.8	-19%	0.18	0.19	0.09	0.07	88.3	58.5	34.4	9.4	1.24	0.82	0.48	0.13
	high	56.3	48.2	-14%	0.16	0.13	0.09	0.43	85.2	54.4	31.7	6.7	0.92	0.59	0.34	0.07
minutes of recreational walking	low	70.0	74.5	6%	0.09	0.10	0.09	0.62	65.0	47.8	30.2	10.5	1.16	0.85	0.54	0.19
	med	79.2	80.7	2%	0.14	0.16	0.21	0.46	117.5	87.8	48.7	9.8	1.18	0.88	0.49	0.10
	high	92.8	92.0	-1%	0.19	0.19	0.19	0.25	65.3	50.5	31.3	9.1	1.49	1.15	0.71	0.21
minutes of moderate PA	low	100.0	102.0	2%	0.10	0.11	0.18	0.48	79.1	56.3	30.4	6.9	1.13	0.80	0.43	0.10
	med	115.8	107.9	-7%	0.09	0.12	0.10	0.70	95.2	72.4	43.5	15.5	1.06	0.81	0.48	0.17
	high	122.4	119.0	-3%	0.14	0.15	0.24	0.48	120.7	88.4	53.9	11.8	1.04	0.76	0.47	0.10
minutes of vigorous PA	low	44.0	49.4	12%	0.19	0.20	0.15	0.64	74.8	56.6	32.9	6.6	1.36	1.03	0.60	0.12
	med	55.0	57.6	5%	0.14	0.16	0.25	0.04	80.6	57.1	32.0	7.5	1.02	0.72	0.40	0.09
	high	71.4	76.3	7%	0.05	0.07	0.07	0.47	77.6	61.5	39.2	14.0	1.16	0.92	0.59	0.21
body mass index	low	27.4	27.5	0%	0.33	0.35	0.46	0.83	4.4	3.1	1.5	0.4	0.16	0.11	0.06	0.01
	med	26.7	26.7	0%	0.34	0.37	0.39	0.91	3.8	2.7	1.5	0.4	0.14	0.10	0.06	0.01
	high	25.6	25.9	1%	0.41	0.40	0.38	0.56	3.2	2.2	1.3	0.4	0.13	0.09	0.05	0.02
likelihood of being overweight or obese	low	58.8%	59.5%	1%	0.33	0.35	0.47	0.82	41.7%	26.8%	13.3%	2.6%	0.71	0.46	0.23	0.04
	med	56.1%	55.7%	-1%	0.34	0.34	0.38	0.88	42.6%	27.5%	13.7%	3.3%	0.76	0.49	0.24	0.06
	high	49.0%	51.1%	4%	0.41	0.41	0.39	0.31	40.8%	24.4%	12.6%	3.7%	0.83	0.50	0.26	0.08
likelihood of being obese	low	27.2%	27.7%	2%	0.27	0.28	0.36	0.69	36.1%	24.1%	11.5%	3.1%	1.33	0.89	0.42	0.12
	med	22.7%	23.5%	4%	0.26	0.26	0.26	0.89	32.5%	23.0%	12.0%	2.4%	1.44	1.02	0.53	0.11
	high	15.0%	17.6%	17%	0.26	0.29	0.24	0.56	25.2%	18.0%	10.2%	3.3%	1.68	1.20	0.68	0.22

					Predicted correlation				Mean absolute error				Mean absolute error / mean outcome			
outcome	income cohort	mean sample observed outcome	mean base predicted outcome	% error	grid	tract	zip	county	grid	tract	zip	county	grid	tract	zip	county
likelihood of having high blood pressure	low	25.0%	25.4%	2%	0.30	0.29	0.27	0.68	5.2%	4.7%	3.5%	0.8%	1.77	1.62	1.22	0.28
	med	23.5%	22.6%	-4%	0.33	0.32	0.28	0.31	9.5%	8.1%	5.4%	1.1%	1.64	1.40	0.93	0.20
	high	20.2%	19.6%	-3%	0.15	0.16	0.09	0.56	6.9%	6.0%	4.3%	0.9%	1.78	1.55	1.12	0.22
likelihood of having heart disease	low	5.2%	4.9%	-6%	0.38	0.40	0.42	0.71	27.2%	17.9%	9.8%	1.4%	1.35	0.89	0.48	0.07
	med	4.6%	3.4%	-26%	0.32	0.33	0.33	0.51	11.1%	9.6%	5.7%	1.1%	1.60	1.39	0.83	0.16
	high	3.9%	3.5%	-9%	0.49	0.49	0.53	0.49	30.7%	20.8%	9.9%	1.9%	1.07	0.73	0.35	0.07
likelihood of having type 2 diabetes	low	6.9%	6.6%	-5%	0.28	0.30	0.27	0.44	8.8%	7.8%	5.2%	0.9%	1.69	1.49	1.00	0.16
	med	4.8%	4.6%	-5%	0.43	0.45	0.41	0.22	30.1%	20.8%	10.5%	1.4%	1.21	0.83	0.42	0.06
	high	2.9%	2.8%	-2%	0.32	0.32	0.28	0.39	8.0%	7.2%	5.0%	1.0%	1.65	1.50	1.03	0.20
likelihood of having poor self-reported health	low	28.7%	29.2%	2%	0.38	0.37	0.42	0.99	16.6%	13.6%	7.7%	1.6%	1.49	1.22	0.69	0.14
	med	11.2%	11.2%	0%	0.24	0.24	0.24	0.21	7.2%	6.5%	4.9%	1.4%	1.59	1.42	1.06	0.32
	high	5.8%	5.5%	-4%	0.42	0.45	0.44	0.96	29.4%	20.8%	11.0%	2.2%	1.25	0.88	0.47	0.09

CHIS validation with Behavioral Risk Factor Surveillance Survey

The latest Behavioral Risk Factor Surveillance Survey (BRFSS) data (from 2012) were summarized and compared to UF model predictions to further validate the CHIS-based models. However, like the CHIS, BRFSS is also a self-reported survey representing only a small sample of California residents. Data collection methodologies, variable definitions, and sample composition differed to some extent between both surveys. For any differences identified between the BRFSS results and CHIS-based model predictions, there is no way to determine from these data which data set provides the more accurate measurement. Ideally, we should find similar results between the two data sets.

BRFSS data are available at the individual level but the smallest geographic identifier provided for each BRFSS participant is the county. Only adult and senior data are available through the BRFSS. Those currently pregnant were removed from the BRFSS data, to match the exclusion criteria for the CHIS models. None of the other CHIS model exclusion variables were available through the BRFSS. Using person weights provided by the BRFSS, relevant outcome variables (i.e. those identical or similar to those used for the CHIS models) were weighted at the individual level and aggregated to the county level. The tables below summarize the weighted mean outcome values at the county level for the 30 counties in the UF study area. The table includes the following outcome variables that differ from the CHIS outcome variables:

- Any exercise: percent of sample participating in any recreational physical activities over past 30 days
- CHD: percent of sample reporting that they have been diagnosed with angina or coronary heart disease
- Diabetes: percent of sample reporting that they have ever been diagnosed with diabetes (but participants did not indicate what type)

These tables are followed by summary tables for predicted CHIS outcomes when applied to all UF grid cells in the 30-county study area. Adult and senior age cohort-specific models were applied at the grid cell level using the base UF grid data. Grid cell-level predictions were then weighted by the age cohort population within each grid cell and aggregated to the county level. UF variable definitions identically matched the CHIS variable definitions, with the following exceptions. Because the mismatched UF variables likely have a negative impact on predictive accuracy, any discrepancies will be addressed as fully as possible in the final UF base data.

- Average age by age cohort was not available in the UF data. CHIS sample mean age by age cohort was assumed as a constant value for all grid cells.
- Percent of adults with children <18 years old was not available in the UF data. CHIS sample mean percent of adults with children <18 years old was assumed as a constant value for all grid cells.

- A factor needed to properly retransform any outcome variables that were log transformed in linear regression models (e.g. minutes of transportation walking), called the Duan Smearing Estimate, was not available at the time of running the BRFSS validation tests. Temporary values were assumed based on prior experience (1.5 for physical activity models, 1.04 for BMI). [ff]
- In general, UF variables were only available for the overall population, but not by age cohort. For some covariates, there are major differences across age cohort (e.g. employment status for adults v. seniors). As data allows, UF variables will be stratified by age cohort in the final UF base data. For this validation test, multipliers were developed based on differences between the UF data and CHIS age group-specific data to convert the general UF variables to age group-specific variables.

Finally, for any outcome variables that matched identically between the BRFSS and CHIS data sets, county-level outcomes were compared and the following summary statistics were calculated:

- Predicted correlation: this was derived by calculating the Pearson's correlation between predicted and observed outcomes at the spatial unit level. Correlation ranges from -1 to 1, where -1 is perfect negative correlation, 0 is no correlation, and 1 is perfect correlation.
- Mean absolute error: for each spatial unit, the mean observed value was subtracted from the mean predicted value and converted to an absolute value, and then the mean value of these absolute differences was calculated. This metric indicates the magnitude of the average error for each model, and is in the same units as the outcome value. This was only calculated for counties with a BRFSS sample size >= 30.
- Mean absolute error/mean outcome: to allow for easier comparison across different models, this metric divides the mean absolute error by the mean outcome value. This can also be thought of as the mean absolute error percentage for each outcome. This was only calculated for counties with a BRFSS sample size >= 30.

Note that due to an error with the preliminary UF base data, CHIS model predictions were not calculated for Imperial County.

Note: Validation metrics were generated only for the earlier version of the models before applying the corrections for multicollinearity problems. A review of validation metrics for the final models (after implementing the approach for dealing with multicollinearity) indicated that they would not differ substantively from those included in this report.

[ff] Duan N. Smearing Estimate: A Nonparametric Retransformation Method. Journal of the American Statistical Association. 1983; 78(3838): 605-610.

Table 84: BRFSS Adult Outcomes

County	Sample size	Any exercise	BMI	Overwt	Obese	CHD	Diabetes	Poor health
Alameda	398	86.3%	26.2	36.2%	19.3%	1.0%	7.2%	10.4%
Contra Costa	217	78.7%	27.6	30.9%	33.0%	0.9%	6.1%	12.0%
El Dorado	65	91.6%	26.2	37.4%	17.2%	1.8%	13.3%	10.4%
Fresno	206	84.4%	27.7	33.5%	28.3%	0.8%	7.3%	29.2%
Imperial	41	64.4%	27.3	44.7%	26.8%	0.0%	7.0%	13.9%
Kern	221	72.8%	28.6	32.7%	36.8%	2.6%	12.8%	20.6%
Kings	41	79.0%	27.9	45.0%	28.5%	3.4%	3.8%	30.0%
Los Angeles	2251	79.1%	27.3	34.4%	26.1%	1.6%	9.3%	20.6%
Madera	44	77.8%	26.9	34.1%	22.4%	1.7%	7.6%	19.0%
Marin	79	92.0%	26.1	26.9%	21.3%	3.1%	8.8%	12.7%
Merced	55	76.4%	29.1	45.4%	34.7%	0.2%	7.7%	16.8%
Napa	36	89.5%	28.1	35.3%	35.8%	0.0%	0.0%	20.0%
Orange	652	84.8%	27.0	38.1%	23.2%	1.3%	7.0%	13.6%
Placer	126	86.0%	26.8	35.1%	22.7%	3.0%	10.4%	4.9%
Riverside	513	79.0%	27.5	36.5%	28.5%	3.1%	10.7%	20.7%
Sacramento	408	85.4%	27.5	37.5%	29.4%	2.0%	8.0%	18.1%
San Bernardino	494	79.1%	28.5	33.4%	34.8%	2.3%	7.0%	14.4%
San Diego	676	83.0%	26.9	36.2%	23.7%	2.2%	7.8%	15.6%
San Francisco	178	87.6%	24.9	33.4%	10.9%	0.2%	4.4%	13.8%
San Joaquin	172	81.9%	28.4	29.6%	35.6%	2.5%	8.4%	19.1%
San Mateo	192	78.4%	26.1	27.9%	23.3%	0.8%	6.9%	10.1%
Santa Clara	427	84.5%	26.0	31.8%	16.0%	2.1%	4.5%	10.0%
Solano	91	85.4%	28.1	34.2%	27.2%	5.4%	4.5%	17.7%
Sonoma	169	86.7%	28.5	42.9%	28.4%	1.2%	8.2%	15.3%
Stanislaus	116	80.2%	29.7	23.0%	41.9%	1.0%	13.2%	20.0%
Sutter	28	47.0%	28.7	45.1%	40.4%	0.0%	6.5%	25.4%
Tulare	123	84.6%	29.1	27.6%	43.9%	1.5%	3.9%	19.8%
Ventura	190	80.8%	26.1	35.4%	19.8%	0.9%	5.6%	17.3%
Yolo	50	80.6%	26.1	28.1%	23.3%	0.0%	1.2%	6.4%
Yuba	24	77.2%	28.8	46.0%	29.3%	0.0%	1.8%	12.3%
Total	**8283**	**81.4%**	**27.3**	**34.7%**	**26.5%**	**1.8%**	**8.0%**	**17.1%**

Table 85: BRFSS Seniors Outcomes

County	Sample size	Any exercise	BMI	Overwt	Obese	CHD	Diabetes	Poor health
Alameda	180	83.0%	25.6	35.7%	13.0%	7.3%	20.2%	19.6%
Contra Costa	113	78.6%	26.5	39.2%	17.4%	9.4%	12.4%	10.9%
El Dorado	44	73.4%	26.8	34.4%	26.7%	2.5%	8.2%	23.2%
Fresno	104	81.1%	27.8	22.7%	35.8%	14.3%	30.8%	28.9%
Imperial	14	62.9%	28.4	69.6%	23.3%	14.9%	13.6%	70.5%
Kern	74	67.5%	26.0	29.8%	16.0%	12.5%	19.4%	33.8%
Kings	10	49.8%	27.8	76.4%	5.4%	3.8%	62.8%	46.4%
Los Angeles	835	74.6%	27.0	37.3%	22.4%	9.1%	21.1%	30.8%
Madera	17	45.6%	25.8	35.0%	27.5%	15.6%	1.7%	59.4%
Marin	50	90.0%	24.2	37.2%	8.9%	0.5%	17.6%	10.2%
Merced	21	56.0%	29.4	41.5%	38.4%	6.3%	44.8%	28.4%
Napa	28	77.3%	26.6	22.8%	29.9%	4.5%	23.9%	5.2%
Orange	315	80.4%	26.2	43.7%	14.6%	15.4%	22.3%	17.7%
Placer	58	69.7%	26.7	23.4%	29.2%	16.7%	24.4%	12.8%
Riverside	264	77.6%	27.1	38.0%	23.6%	18.1%	13.6%	24.4%
Sacramento	159	78.7%	26.1	38.0%	19.8%	14.2%	16.8%	20.2%
San Bernardino	169	73.4%	27.0	42.9%	17.9%	9.9%	34.7%	20.9%
San Diego	326	80.4%	26.4	41.7%	18.3%	10.6%	17.9%	20.2%
San Francisco	84	86.3%	27.0	30.4%	28.4%	5.7%	31.1%	18.7%
San Joaquin	64	73.5%	27.4	41.9%	26.3%	9.2%	22.8%	27.3%
San Mateo	86	79.9%	26.1	31.2%	21.7%	10.9%	22.2%	25.7%
Santa Clara	172	79.6%	26.2	35.5%	18.1%	8.0%	23.7%	16.0%
Solano	52	74.0%	27.7	46.5%	25.3%	11.7%	14.5%	16.3%
Sonoma	97	83.4%	27.0	31.0%	27.6%	21.1%	11.5%	17.1%
Stanislaus	61	71.1%	28.3	25.4%	35.5%	11.5%	19.5%	33.4%
Sutter	15	58.8%	27.6	46.2%	20.3%	19.1%	26.4%	9.2%
Tulare	43	46.5%	26.6	42.5%	20.3%	12.5%	23.6%	34.7%
Ventura	97	72.6%	27.8	47.2%	26.0%	9.1%	24.0%	34.8%
Yolo	16	76.2%	26.5	49.0%	14.3%	10.3%	30.5%	24.3%
Yuba	10	94.5%	27.7	68.7%	14.9%	29.6%	38.5%	32.9%
Total	**3578**	**76.5%**	**26.7**	**38.1%**	**21.1%**	**11.3%**	**21.0%**	**24.1%**

Table 86: CHIS Adult Outcomes

County	County pop.	Any trans walk	Any leis walk	Any mod PA	Any vig PA	Trans walk min.	Leis walk min.	Mod PA min.	Vig PA min.	BMI	Overwt	Obese	High BP	Heart disease	Type 2 diabetes	Poor health
Alameda	941578	53.9%	59.4%	54.5%	29.8%	56.9	76.9	88.9	50.1	27.0	34.6%	18.6%	22.3%	2.9%	3.7%	15.1%
Contra Costa	589750	47.0%	60.7%	57.9%	31.5%	47.3	78.5	101.0	54.4	27.4	36.4%	21.3%	23.1%	3.0%	3.5%	12.2%
El Dorado	96173	33.3%	60.9%	62.0%	32.3%	30.1	76.7	125.5	55.1	27.8	37.0%	24.6%	24.8%	3.4%	2.9%	9.2%
Fresno	463773	39.7%	56.3%	64.2%	27.2%	42.9	71.1	129.6	43.8	28.2	37.6%	25.7%	25.7%	3.6%	4.0%	20.4%
Kern	388212	38.2%	55.4%	64.1%	26.7%	41.0	69.3	130.5	42.9	28.6	37.9%	28.5%	26.5%	3.8%	4.2%	20.9%
Kings	82376	37.1%	54.5%	63.6%	26.0%	40.4	68.4	130.3	41.0	29.0	38.7%	31.1%	27.0%	3.5%	4.7%	19.6%
Los Angeles	5924689	54.9%	58.7%	53.3%	26.8%	59.4	74.2	82.5	45.5	27.6	37.3%	22.0%	24.3%	3.0%	4.5%	18.7%
Madera	73046	37.7%	56.3%	65.3%	27.4%	41.8	71.5	134.4	44.2	28.7	38.6%	29.3%	25.7%	3.6%	4.5%	19.3%
Marin	163672	48.6%	65.5%	60.0%	37.4%	45.3	87.2	108.7	65.5	26.8	33.5%	18.1%	19.0%	2.7%	2.1%	7.2%
Merced	117866	36.0%	54.7%	63.8%	26.0%	39.6	68.5	129.8	40.8	28.8	38.8%	29.3%	27.6%	3.9%	4.6%	22.3%
Napa	75195	42.1%	59.6%	57.2%	29.6%	42.7	75.7	101.9	49.8	28.1	38.0%	25.6%	25.2%	3.6%	4.0%	13.6%
Orange	1797107	50.7%	61.6%	56.5%	30.7%	51.4	78.4	91.1	52.7	26.9	35.8%	18.1%	21.3%	2.6%	3.5%	12.3%
Placer	150081	39.9%	61.4%	59.8%	32.0%	37.7	78.3	112.8	54.7	27.6	36.7%	22.8%	23.4%	3.1%	2.8%	9.0%
Riverside	880732	42.6%	57.6%	57.8%	27.4%	43.4	70.9	99.9	46.1	28.3	39.7%	26.5%	26.2%	3.3%	4.9%	17.0%
Sacramento	750022	48.1%	57.3%	54.4%	27.1%	49.6	71.7	91.7	45.3	27.8	36.9%	23.3%	25.6%	3.6%	4.0%	17.8%
San Bernardino	1010928	43.8%	56.8%	56.8%	26.4%	45.2	69.6	96.0	44.4	28.4	39.4%	27.2%	27.1%	3.4%	5.1%	18.6%
San Diego	1776422	49.9%	60.6%	56.9%	30.9%	52.3	79.0	97.5	52.4	27.6	36.8%	22.0%	22.4%	2.9%	3.7%	13.1%
San Francisco	557820	61.5%	60.5%	49.6%	28.8%	67.1	80.6	73.8	48.7	26.6	34.1%	15.8%	21.6%	2.9%	3.5%	14.7%
San Joaquin	329230	39.2%	56.0%	64.1%	27.3%	41.7	70.4	130.5	43.9	28.0	37.0%	24.8%	25.4%	3.6%	3.8%	19.4%
San Mateo	456976	51.0%	62.0%	57.2%	31.7%	53.2	81.9	95.8	54.4	26.8	34.7%	17.0%	21.1%	2.5%	3.1%	10.2%
Santa Clara	1105656	48.8%	66.0%	59.5%	38.2%	44.9	88.6	102.1	66.6	26.1	31.8%	14.1%	18.3%	2.3%	2.2%	6.7%
Solano	245264	42.2%	57.6%	56.0%	28.7%	42.4	72.0	97.8	48.6	27.9	37.6%	24.2%	26.2%	3.2%	4.0%	14.2%
Sonoma	288484	44.4%	60.8%	58.5%	31.5%	43.9	78.0	106.6	53.2	27.8	37.2%	23.6%	22.6%	3.1%	3.3%	11.2%
Stanislaus	261078	39.4%	56.4%	64.6%	27.1%	41.7	71.0	133.4	43.7	28.3	37.2%	26.7%	25.3%	3.7%	3.6%	17.7%
Sutter	46306	40.6%	54.6%	54.6%	24.4%	42.6	66.9	95.9	39.2	28.4	39.1%	26.9%	27.7%	3.9%	5.1%	20.4%
Tulare	207852	33.2%	56.1%	64.4%	27.4%	36.1	70.8	130.1	43.7	28.8	39.2%	29.9%	26.7%	3.4%	4.5%	20.5%
Ventura	462149	43.4%	60.3%	57.1%	30.5%	43.4	75.9	97.7	52.0	28.0	38.2%	24.8%	24.4%	2.9%	4.1%	12.4%
Yolo	110399	49.1%	60.0%	56.4%	30.6%	50.5	77.3	97.3	50.3	27.5	35.9%	21.4%	21.9%	3.0%	3.8%	14.6%
Yuba	35119	37.3%	53.0%	55.0%	23.5%	38.5	63.6	102.4	37.4	28.8	37.8%	31.0%	30.4%	5.0%	5.0%	25.4%
Total	**19387955**	**49.5%**	**59.4%**	**56.4%**	**29.0%**	**51.7**	**75.8**	**95.2**	**49.1**	**27.5**	**36.7%**	**21.9%**	**23.6%**	**3.0%**	**4.0%**	**15.7%**

Table 87: CHIS Senior Outcomes

County	County pop.	Any trans walk	Any leis walk	Any mod PA	Any vig PA	Trans walk min.	Leis walk min.	Mod PA min.	Vig PA min.	BMI	Overwt	Obese	High BP	Heart disease	Type 2 diabetes	Poor health
Alameda	147591	39.7%	57.6%	56.8%	15.1%	40.3	85.6	106.7	26.6	26.3	37.2%	12.1%	61.9%	18.7%	13.5%	19.5%
Contra Costa	107272	35.6%	56.7%	59.6%	16.4%	35.6	82.6	118.2	29.5	26.4	38.0%	13.2%	59.6%	19.8%	12.6%	13.9%
El Dorado	19334	26.1%	53.7%	62.7%	16.4%	26.7	75.1	139.7	30.9	26.6	39.7%	14.9%	59.2%	21.4%	12.0%	9.7%
Fresno	79209	26.3%	50.2%	60.6%	13.9%	24.4	63.1	131.1	25.8	27.2	43.7%	16.4%	64.2%	20.4%	13.0%	17.6%
Kern	62054	24.8%	49.2%	60.8%	13.6%	22.7	61.4	134.6	25.6	27.4	44.1%	17.9%	64.4%	21.1%	13.0%	16.4%
Kings	9557	24.3%	46.7%	59.7%	12.2%	23.7	57.4	129.2	22.1	27.7	46.2%	19.3%	65.6%	21.2%	14.4%	16.6%
Los Angeles	926673	40.5%	55.8%	54.3%	13.8%	41.8	80.3	98.7	23.1	26.7	39.7%	14.5%	62.3%	18.4%	16.0%	22.8%
Madera	13596	22.8%	50.7%	61.1%	13.2%	21.2	63.5	140.6	26.4	27.5	43.7%	19.2%	64.3%	22.4%	13.4%	16.5%
Marin	33432	37.8%	62.3%	63.4%	20.4%	36.9	95.5	129.4	37.4	26.0	36.2%	11.3%	54.7%	19.1%	9.4%	9.0%
Merced	20004	23.0%	47.5%	59.5%	12.0%	22.1	58.7	129.4	22.9	27.7	45.2%	19.3%	66.5%	21.9%	14.9%	18.7%
Napa	19086	30.7%	53.3%	58.7%	15.1%	32.2	75.6	118.9	27.2	26.8	40.1%	15.7%	60.3%	21.3%	13.5%	13.0%
Orange	280763	39.2%	58.9%	58.4%	16.5%	39.9	87.8	113.8	28.2	26.1	36.5%	11.5%	58.1%	18.0%	12.3%	15.4%
Placer	32560	30.5%	54.6%	61.2%	16.5%	31.1	78.2	128.6	29.7	26.5	38.8%	14.1%	59.0%	20.9%	12.0%	10.7%
Riverside	195964	31.0%	52.5%	57.9%	13.9%	31.1	72.4	121.1	24.5	27.0	40.7%	16.6%	61.5%	20.4%	14.8%	16.2%
Sacramento	135875	34.9%	53.7%	56.1%	14.1%	35.2	76.6	108.1	24.6	26.7	39.4%	14.9%	62.5%	21.4%	14.0%	18.1%
San Bernardino	146459	30.8%	50.5%	55.7%	12.5%	30.7	67.1	111.8	22.0	27.3	42.0%	18.5%	63.5%	20.2%	16.6%	19.4%
San Diego	313750	36.4%	56.0%	57.9%	15.5%	37.4	80.9	113.8	27.8	26.7	39.7%	14.4%	60.3%	20.0%	13.5%	15.4%
San Francisco	106111	48.3%	58.5%	53.1%	14.4%	56.0	94.7	89.2	22.9	25.8	35.1%	9.6%	61.1%	20.0%	15.6%	24.1%
San Joaquin	59799	25.5%	50.4%	60.4%	13.4%	23.2	63.3	133.2	25.2	27.1	43.4%	16.0%	64.9%	21.2%	12.0%	17.5%
San Mateo	88085	39.9%	57.6%	59.2%	16.9%	43.7	87.1	110.2	29.6	25.8	36.1%	9.9%	59.1%	17.8%	12.7%	14.0%
Santa Clara	160527	42.5%	60.6%	62.3%	20.2%	44.0	93.5	121.5	35.1	25.7	35.3%	9.4%	54.1%	17.6%	10.7%	10.2%
Solano	37426	30.8%	52.5%	57.7%	14.0%	31.9	73.9	112.0	25.5	26.8	40.3%	15.4%	63.8%	20.3%	15.1%	16.1%
Sonoma	57977	32.0%	56.4%	60.1%	16.3%	31.8	81.0	127.8	29.8	26.6	38.8%	14.7%	58.4%	21.0%	11.5%	11.8%
Stanislaus	46697	25.6%	48.5%	60.3%	13.4%	23.3	60.2	132.3	24.0	27.2	43.9%	16.5%	64.4%	22.4%	11.9%	16.7%
Sutter	9755	29.1%	51.4%	56.2%	12.5%	31.0	71.9	118.9	22.7	27.0	40.1%	16.8%	62.3%	22.4%	15.3%	17.1%
Tulare	35917	21.3%	49.6%	59.6%	12.5%	20.1	61.2	132.5	26.7	27.8	44.0%	20.8%	66.0%	21.1%	14.9%	16.3%
Ventura	76804	33.2%	53.6%	58.5%	15.6%	35.1	76.9	113.9	27.2	26.8	40.3%	15.4%	59.9%	19.6%	14.1%	13.3%
Yolo	15782	34.3%	55.3%	57.0%	14.4%	34.8	79.8	116.1	26.2	26.8	39.4%	15.7%	60.4%	21.4%	13.3%	16.0%
Yuba	6410	25.5%	50.0%	57.2%	12.4%	26.0	67.5	125.0	23.1	27.2	41.4%	18.5%	63.6%	23.8%	14.7%	16.4%
Total	**3244469**	**36.5%**	**55.3%**	**57.3%**	**14.9%**	**37.4**	**79.4**	**111.6**	**26.0**	**26.6**	**39.3%**	**14.2%**	**61.1%**	**19.4%**	**14.1%**	**17.9%**

Table 88: Adult Outcome Comparison

	UF prediction – BRFSS observed difference			
County	BMI	Overwt	Obese	Poor health
Alameda	0.8	-1.6%	-0.7%	4.7%
Contra Costa	-0.2	5.5%	-11.7%	0.2%
El Dorado	1.6	-0.4%	7.5%	-1.2%
Fresno	0.5	4.1%	-2.6%	-8.8%
Kern	0.0	5.1%	-8.3%	0.3%
Kings	1.1	-6.3%	2.6%	-10.4%
Los Angeles	0.3	2.9%	-4.2%	-1.9%
Madera	1.8	4.5%	6.9%	0.4%
Marin	0.7	6.6%	-3.2%	-5.5%
Merced	-0.4	-6.6%	-5.4%	5.5%
Napa	-0.1	2.7%	-10.3%	-6.4%
Orange	-0.1	-2.3%	-5.1%	-1.3%
Placer	0.7	1.6%	0.2%	4.1%
Riverside	0.8	3.1%	-1.9%	-3.7%
Sacramento	0.3	-0.6%	-6.1%	-0.3%
San Bernardino	-0.1	6.0%	-7.6%	4.2%
San Diego	0.6	0.6%	-1.7%	-2.5%
San Francisco	1.7	0.7%	4.9%	0.9%
San Joaquin	-0.4	7.5%	-10.8%	0.2%
San Mateo	0.7	6.8%	-6.3%	0.2%
Santa Clara	0.1	-0.1%	-1.9%	-3.3%
Solano	-0.1	3.4%	-3.0%	-3.5%
Sonoma	-0.7	-5.7%	-4.8%	-4.1%
Stanislaus	-1.4	14.2%	-15.2%	-2.3%
Sutter	-0.3	-6.0%	-13.5%	-5.0%
Tulare	-0.3	11.6%	-14.1%	0.7%
Ventura	1.9	2.8%	5.0%	-4.9%
Yolo	1.4	7.8%	-1.8%	8.2%
Yuba	0.1	-8.2%	1.6%	13.1%
Total	0.3	2.0%	-4.5%	-1.3%

Table 89: Senior Outcome Comparison

	UF prediction – BRFSS observed difference			
County	BMI	Overwt	Obese	Poor health
Alameda	0.6	1.5%	-1.0%	-0.1%
Contra Costa	-0.1	-1.3%	-4.2%	3.0%
El Dorado	-0.2	5.3%	-11.8%	-13.5%
Fresno	-0.6	21.0%	-19.3%	-11.3%
Kern	1.4	14.3%	1.9%	-17.4%
Kings	-0.1	-30.2%	13.9%	-29.8%
Los Angeles	-0.2	2.4%	-8.0%	-8.0%
Madera	1.7	8.8%	-8.3%	-42.9%
Marin	1.8	-1.0%	2.5%	-1.3%
Merced	-1.8	3.8%	-19.1%	-9.7%
Napa	0.2	17.3%	-14.2%	7.8%
Orange	-0.1	-7.2%	-3.1%	-2.3%
Placer	-0.2	15.4%	-15.1%	-2.1%
Riverside	-0.1	2.7%	-7.0%	-8.2%
Sacramento	0.6	1.3%	-4.9%	-2.1%
San Bernardino	0.3	-0.8%	0.6%	-1.4%
San Diego	0.3	-2.1%	-3.9%	-4.8%
San Francisco	-1.2	4.7%	-18.7%	5.4%
San Joaquin	-0.3	1.6%	-10.3%	-9.7%
San Mateo	-0.3	4.9%	-11.8%	-11.7%
Santa Clara	-0.6	-0.2%	-8.6%	-5.8%
Solano	-0.9	-6.2%	-9.9%	-0.2%
Sonoma	-0.5	7.8%	-12.9%	-5.2%
Stanislaus	-1.1	18.5%	-19.0%	-16.7%
Sutter	-0.6	-6.1%	-3.5%	7.8%
Tulare	1.2	1.4%	0.5%	-18.4%
Ventura	-1.1	-6.9%	-10.6%	-21.5%
Yolo	0.3	-9.7%	1.5%	-8.3%
Yuba	-0.5	-27.3%	3.6%	-16.5%
Total	-0.1	1.5%	-7.0%	-6.0%

Table 90: Adult UF prediction – BRFSS observed difference

	UF prediction – BRFSS observed difference			
Metric	BMI	Overwt	Obese	Poor health
Predicted correlation	0.75	0.37	0.66	0.59
Mean absolute error	0.70	4.5%	5.7%	3.3%
Mean absolute error / mean outcome ratio	0.03	0.14	0.20	0.25

Table 91: Senior UF prediction – BRFSS observed difference

	UF prediction – BRFSS observed difference			
Metric	BMI	Overwt	Obese	Poor health
Predicted correlation	0.51	0.26	0.14	0.28
Mean absolute error	0.62	5.8%	8.4%	7.7%
Mean absolute error / mean outcome ratio	0.02	0.20	0.34	0.30

CHTS validation with National Household Transportation Survey

The latest National Household Travel Survey (NHTS) data (from 2009) were summarized and compared to UF model predictions to further validate the CHTS models. However, like the CHTS, NHTS is a self-reported survey representing only a small sample of California residents. Data collection methodologies, variable definitions, and sample composition differed to some extent between both surveys. For any differences identified between the NHTS results and CHTS-based model predictions, there is no way to determine from these data which data set provides the more accurate measurement. Ideally, we should find similar results between the two data sets.

NHTS data are available at the individual level and a variety of geographic identifiers are provided for each NHTS participant. Because of the small sample sizes available at smaller geographic levels, NHTS data were used at the county level. Using person weights provided by the NHTS, relevant outcome variables (i.e. those identical or similar to those used for the CHIS models) were weighted at the individual level and aggregated to the county level. The tables below summarize the weighted mean outcome values at the county level for the 30 counties in the UF study area. Data was not available in the NHTS on recreational physical activity engagement.

These tables are followed by summary tables for predicted CHTS outcomes when applied to all UF grid cells in the 30-county study area. Adult and senior age cohort-specific models were applied at the grid cell level using the base UF grid data. Grid cell-level predictions were then weighted by the age cohort population within each grid cell and aggregated to the county level. UF variable definitions identically matched the CHTS variable definitions, with the following exceptions. Note that any discrepancies will be resolved in the final UF base data.

- Average age by age cohort was not available in the UF data. CHIS sample mean age by age cohort was assumed as a constant value for all grid cells.
- Percent of adults with children <18 years old was not available in the UF data. CHIS sample mean percent of adults with children <18 years old was assumed as a constant value for all grid cells.
- In general, UF variables were only available for the overall population, but not by age cohort. For some covariates, there are major differences across age cohort (e.g. employment status for adults v. seniors). As data allows, UF variables will be stratified by age cohort in the final UF base data. For this validation test, multipliers were developed based on differences between the UF data and CHIS age group-specific data to convert the general UF variables to age group-specific variables.

Finally, for any outcome variables that matched identically between the NHTS and CHTS data sets, county-level outcomes were compared and the following summary statistics were calculated:

- Predicted correlation: this was derived by calculating the Pearson's correlation between predicted and observed outcomes at the spatial unit level. Correlation ranges from -1 to 1, where -1 is perfect negative correlation, 0 is no correlation, and 1 is perfect correlation.
- Mean absolute error: for each spatial unit, the mean observed value was subtracted from the mean predicted value and converted to an absolute value, and then the mean value of these absolute differences was calculated. This metric indicates the magnitude of the average error for each model, and is in the same units as the outcome value. This was only calculated for counties with an NHTS sample size >= 30.
- Mean absolute error/mean outcome: to allow for easier comparison across different models, this metric divides the mean absolute error by the mean outcome value. This can also be thought of as the mean absolute error percentage for each outcome. This was only calculated for counties with an NHTS sample size >= 30.

Note that due to errors with the preliminary UF base data, CHTS model predictions were not calculated for Imperial County or for teens and children in Los Angeles County.

Note: Validation metrics were generated only for the earlier version of the models before applying the corrections for multicollinearity problems. A review of validation metrics for the final models (after implementing the approach for dealing with multicollinearity) indicated that they would not differ substantively from those included in this report.

Table 92: NHTS adult outcomes

County	Sample size	Walk minutes	Bike minutes	Auto minutes
Alameda	547	8.6	0.9	63.5
Contra Costa	406	8.1	2.9	84.0
El Dorado	78	1.6	0.1	81.3
Fresno	217	3.8	1.9	63.4
Imperial	38	0.5	0.0	66.5
Kern	204	7.0	0.7	72.1
Kings	47	4.2	2.3	75.9
Los Angeles	2,242	10.0	1.3	72.5
Madera	41	2.8	0.0	76.4
Marin	119	9.0	1.9	73.9
Merced	61	6.8	0.0	53.0
Napa	50	2.4	1.2	62.8
Orange	943	6.6	1.0	75.0
Placer	176	3.4	0.1	73.6
Riverside	521	5.0	0.1	80.6
Sacramento	501	5.8	1.9	73.5
San Bernardino	526	3.7	0.5	74.8
San Diego	4,108	5.8	0.9	69.2
San Francisco	234	25.8	2.9	42.8
San Joaquin	188	4.1	0.0	64.5
San Mateo	245	7.8	2.1	68.0
Santa Clara	611	5.5	2.0	63.6
Solano	147	5.2	0.0	69.3
Sonoma	263	3.2	1.2	68.9
Stanislaus	175	6.0	0.0	83.2
Sutter	23	0.1	1.7	70.3
Tulare	134	4.5	0.3	62.2
Ventura	282	2.7	1.9	68.6
Yolo	89	6.5	4.9	55.9
Yuba	15	0.0	0.0	107.3
Total	13,231	7.5	1.2	70.7

Table 93: Senior NHTS outcomes

County	Sample size	Walk minutes	Bike minutes	Auto minutes
Alameda	169	6.5	0.8	53.5
Contra Costa	198	3.1	0.6	56.5
El Dorado	25	0.1	10.6	57.4
Fresno	106	1.3	0.0	66.1
Imperial	7	0.0	0.0	37.6
Kern	88	1.5	0.4	58.6
Kings	11	0.0	0.0	39.8
Los Angeles	759	6.9	0.6	56.7
Madera	23	0.4	0.0	50.5
Marin	51	7.3	0.4	58.9
Merced	20	0.0	2.5	38.4
Napa	21	12.5	0.0	62.6
Orange	319	4.0	0.5	57.2
Placer	82	1.5	0.0	57.7
Riverside	233	0.9	0.4	59.8
Sacramento	171	3.1	0.1	68.2
San Bernardino	152	2.8	0.2	57.6
San Diego	1,552	4.6	0.3	55.5
San Francisco	71	26.0	0.0	46.6
San Joaquin	69	3.9	0.0	56.9
San Mateo	100	3.9	0.0	55.4
Santa Clara	218	6.7	0.6	61.7
Solano	56	3.7	0.8	51.8
Sonoma	90	3.8	0.3	52.2
Stanislaus	62	1.7	0.0	76.1
Sutter	6	0.0	0.0	59.0
Tulare	22	1.9	3.2	42.2
Ventura	97	4.2	0.0	59.5
Yolo	22	3.5	4.2	42.4
Yuba	10	0.0	2.8	86.4
Total	4,810	5.1	0.5	57.4

Table 94: Teen NHTS outcomes

County	Sample size	Walk minutes	Bike minutes	Auto minutes
Alameda	53	17.9	0.3	28.3
Contra Costa	43	16.6	0.9	37.3
El Dorado	9	0.0	0.0	38.1
Fresno	34	11.5	1.9	23.1
Imperial	6	2.7	0.0	46.7
Kern	25	4.7	4.0	21.1
Kings	8	7.1	0.0	112.0
Los Angeles	253	18.3	0.7	30.8
Madera	1	0.0	0.0	0.0
Marin	17	10.1	0.0	50.5
Merced	10	13.5	0.0	14.9
Napa	4	4.7	0.0	17.5
Orange	144	13.2	0.7	39.2
Placer	23	2.9	3.4	37.4
Riverside	55	12.0	1.1	38.6
Sacramento	46	15.2	1.7	32.8
San Bernardino	64	5.7	2.4	28.1
San Diego	451	12.6	1.4	33.5
San Francisco	10	9.3	0.0	21.8
San Joaquin	20	1.9	0.2	20.7
San Mateo	29	8.5	7.3	22.5
Santa Clara	67	4.7	1.3	34.2
Solano	11	0.0	0.0	34.0
Sonoma	25	2.7	0.6	39.8
Stanislaus	22	13.0	0.0	37.3
Sutter	5	0.0	0.0	11.3
Tulare	14	5.4	7.4	17.0
Ventura	32	4.5	4.6	35.3
Yolo	7	20.3	8.8	22.2
Yuba	5	4.9	0.0	54.2
Total	1,493	12.8	1.3	32.5

Table 95: Child NHTS outcomes

County	Sample size	Walk minutes	Bike minutes	Auto minutes
Alameda	78	4.5	0.7	28.9
Contra Costa	43	4.0	0.8	25.6
El Dorado	10	1.2	0.0	23.4
Fresno	37	7.4	3.6	25.5
Imperial	9	2.1	0.0	23.1
Kern	42	3.5	0.0	32.7
Kings	5	4.0	0.0	7.6
Los Angeles	273	16.2	0.1	29.4
Madera	3	5.7	0.0	28.8
Marin	18	5.4	2.0	30.0
Merced	8	20.9	0.0	19.3
Napa	2	0.0	0.0	31.2
Orange	118	12.0	0.1	34.1
Placer	28	4.7	1.6	32.9
Riverside	46	3.7	0.0	39.0
Sacramento	40	3.6	0.3	30.0
San Bernardino	52	2.4	0.3	28.6
San Diego	487	7.2	0.7	34.5
San Francisco	21	11.3	0.0	28.3
San Joaquin	17	8.5	0.0	32.3
San Mateo	33	0.6	0.0	29.1
Santa Clara	72	8.7	0.4	31.4
Solano	15	22.9	0.0	33.4
Sonoma	23	7.1	0.5	32.5
Stanislaus	38	8.9	0.0	28.4
Sutter	2	0.0	0.0	25.6
Tulare	17	9.2	0.0	31.5
Ventura	31	7.3	0.0	22.8
Yolo	8	6.1	0.0	34.1
Yuba	3	0.0	0.0	17.9
Total	1,579	9.9	0.4	30.4

Table 96: CHTS adult outcomes

County	County pop.	Any trans walk	Any trans bike	Any auto trans	Any rec PA	Trans walk min.	Trans bike min.	Auto trans min.	Rec PA min.
Alameda	941578	25.8%	2.8%	77.9%	15.9%	10.2	1.6	74.1	17.2
Contra Costa	589750	16.1%	1.9%	82.6%	15.7%	5.2	1.1	80.9	17.4
El Dorado	96173	6.1%	1.2%	87.7%	15.6%	1.5	0.6	88.1	17.5
Fresno	463773	9.4%	1.1%	80.2%	13.0%	2.5	0.7	85.0	15.2
Kern	388212	7.6%	1.0%	81.8%	12.7%	1.8	0.6	85.1	14.9
Kings	82376	7.7%	1.0%	81.1%	12.3%	1.8	0.6	81.5	14.9
Los Angeles	5924689	8.6%	0.7%	87.3%	12.8%	5.9	0.8	83.1	14.3
Madera	73046	8.1%	1.1%	80.9%	12.7%	2.0	0.6	81.6	15.0
Marin	163672	18.2%	2.8%	84.9%	20.0%	5.6	1.7	83.9	21.2
Merced	117866	7.1%	0.9%	81.9%	12.3%	1.7	0.5	83.4	14.6
Napa	75195	13.1%	1.8%	82.3%	14.3%	4.4	0.8	79.0	16.5
Orange	1797107	13.8%	1.5%	84.1%	14.4%	4.5	0.9	85.1	16.1
Placer	150081	10.2%	1.6%	85.8%	15.6%	2.9	0.8	84.4	17.4
Riverside	880732	7.6%	0.9%	85.0%	11.7%	2.2	0.5	85.7	13.7
Sacramento	750022	16.7%	1.9%	80.3%	14.1%	5.6	1.0	76.2	15.4
San Bernardino	1010928	8.1%	0.8%	84.6%	11.6%	2.4	0.4	85.8	13.5
San Diego	1776422	21.0%	2.5%	80.1%	15.7%	7.3	1.3	75.8	17.2
San Francisco	557820	42.5%	5.2%	71.7%	18.8%	19.5	2.8	64.4	18.8
San Joaquin	329230	8.4%	1.1%	81.1%	13.1%	2.1	0.7	84.5	15.2
San Mateo	456976	20.0%	2.1%	81.4%	16.9%	7.0	1.3	80.3	18.6
Santa Clara	1105656	20.9%	2.7%	84.5%	20.0%	7.0	1.8	86.6	21.6
Solano	245264	11.9%	1.4%	84.0%	13.9%	3.6	0.7	81.5	15.9
Sonoma	288484	13.4%	2.0%	84.0%	15.3%	4.0	1.0	80.0	16.9
Stanislaus	261078	8.0%	1.1%	81.2%	13.4%	1.9	0.7	84.1	15.5
Sutter	46306	10.1%	1.3%	82.2%	11.9%	3.0	0.5	75.0	13.7
Tulare	207852	5.9%	0.7%	85.3%	12.1%	1.3	0.4	87.9	14.2
Ventura	462149	10.6%	1.3%	85.7%	14.6%	3.2	0.7	84.8	16.8
Yolo	110399	19.2%	2.6%	80.8%	16.0%	6.8	1.3	74.6	17.8
Yuba	35119	9.6%	1.3%	82.1%	11.9%	2.8	0.5	76.5	13.5
Total	**19387955**	**13.7%**	**1.5%**	**83.7%**	**14.3%**	**5.6**	**1.0**	**81.8**	**15.9**

Table 97: CHTS senior outcomes

County	County pop.	Any trans walk	Any auto trans	Any rec PA	Trans walk min.	Auto trans min.	Rec PA min.
Alameda	147591	12.9%	71.9%	16.9%	4.8	49.5	18.9
Contra Costa	107272	8.2%	74.3%	16.4%	2.7	53.1	17.0
El Dorado	19334	2.8%	77.5%	14.4%	0.8	58.2	14.0
Fresno	79209	2.6%	72.8%	12.4%	0.9	67.1	15.2
Kern	62054	2.1%	73.7%	12.4%	0.7	66.5	15.0
Kings	9557	2.2%	69.5%	10.1%	0.8	61.2	12.8
Los Angeles	926673	5.2%	82.1%	14.9%	2.6	52.9	15.9
Madera	13596	1.9%	72.8%	13.5%	0.6	63.8	15.5
Marin	33432	8.5%	78.1%	20.5%	2.8	56.7	21.3
Merced	20004	2.1%	71.5%	11.0%	0.7	62.8	13.1
Napa	19086	6.0%	73.4%	13.7%	1.8	52.1	13.6
Orange	280763	7.5%	76.9%	16.9%	2.1	53.7	17.9
Placer	32560	4.9%	76.2%	15.3%	1.4	54.8	15.0
Riverside	195964	3.5%	75.3%	13.5%	0.9	53.4	13.7
Sacramento	135875	8.8%	70.2%	14.8%	3.1	47.6	15.7
San Bernardino	146459	3.6%	75.7%	12.2%	1.0	54.8	12.5
San Diego	313750	9.9%	72.1%	15.4%	3.5	50.3	16.0
San Francisco	106111	32.6%	61.4%	18.4%	12.6	37.2	18.8
San Joaquin	59799	2.3%	71.8%	12.9%	0.8	64.4	16.5
San Mateo	88085	12.4%	73.9%	15.8%	3.8	53.1	16.5
Santa Clara	160527	11.9%	77.0%	18.1%	3.6	55.3	18.5
Solano	37426	6.4%	73.4%	13.2%	1.9	51.0	13.3
Sonoma	57977	5.5%	75.2%	16.2%	1.8	53.7	17.0
Stanislaus	46697	2.2%	71.6%	12.5%	0.7	64.7	16.1
Sutter	9755	4.7%	70.2%	12.4%	1.4	47.1	13.3
Tulare	35917	1.9%	77.0%	12.0%	0.6	66.2	12.5
Ventura	76804	6.1%	76.1%	13.9%	1.7	53.3	13.7
Yolo	15782	7.6%	72.0%	15.7%	2.6	48.6	17.2
Yuba	6410	4.4%	70.4%	12.7%	1.5	49.6	13.4
Total	**3244469**	**7.5%**	**76.0%**	**15.1%**	**2.7**	**53.4**	**16.1**

Table 98: CHTS teen outcomes

County	County pop.	Any trans walk	Any auto trans	Any rec PA	Trans walk min.	Auto trans min.	Rec PA min.
Alameda	113530	34.5%	67.9%	18.9%	12.9	34.4	34.1
Contra Costa	81085	27.9%	70.4%	18.9%	9.4	36.9	34.6
El Dorado	14552	15.8%	77.3%	19.7%	4.0	43.1	36.2
Fresno	84897	18.9%	56.6%	15.1%	4.0	39.2	30.2
Kern	68861	17.8%	58.8%	16.2%	3.7	40.1	32.3
Kings	12137	17.8%	58.8%	16.6%	3.8	39.0	33.1
Madera	12185	19.9%	56.4%	16.5%	4.2	37.8	33.3
Marin	15860	26.4%	75.4%	23.3%	8.1	41.2	42.8
Merced	23895	17.3%	58.5%	15.9%	3.5	39.3	32.0
Napa	10343	26.1%	72.0%	18.1%	8.5	36.4	33.3
Orange	239197	30.8%	71.5%	16.6%	10.9	34.3	29.4
Placer	22193	21.6%	74.9%	19.4%	6.4	39.9	35.7
Riverside	152167	21.9%	72.3%	16.2%	6.8	36.6	29.1
Sacramento	109927	30.6%	67.5%	17.6%	10.7	34.0	32.2
San Bernardino	179599	23.6%	71.4%	15.6%	7.7	36.0	27.9
San Diego	239753	33.7%	66.7%	18.7%	12.1	34.1	34.3
San Francisco	38488	47.4%	64.8%	17.8%	21.1	28.8	31.4
San Joaquin	58687	17.9%	58.4%	16.4%	3.7	40.1	32.5
San Mateo	50989	32.4%	69.0%	17.4%	11.2	35.7	32.0
Santa Clara	131681	27.8%	77.9%	20.8%	8.4	40.4	36.8
Solano	37052	27.4%	71.0%	18.1%	9.4	36.2	32.9
Sonoma	39235	25.8%	72.2%	20.5%	8.3	37.5	37.5
Stanislaus	46388	16.7%	60.0%	16.6%	3.4	41.3	33.0
Sutter	7635	24.7%	69.6%	17.5%	7.9	33.7	32.5
Tulare	40928	18.6%	57.9%	14.8%	3.9	38.9	29.8
Ventura	69482	27.8%	74.3%	18.0%	9.8	35.9	32.2
Yolo	17638	29.6%	69.9%	20.8%	9.8	35.5	37.9
Yuba	6191	23.3%	69.4%	17.5%	7.1	36.1	32.4
Total	**1924575**	**27.1%**	**68.5%**	**17.6%**	**8.9**	**36.4**	**32.2**

Table 99. CHTS child outcomes

County	County pop.	Any trans walk	Any auto trans	Any rec PA	Trans walk min.	Auto trans min.	Rec PA min.
Alameda	143356	24.7%	81.1%	20.1%	6.0	37.0	32.8
Contra Costa	101838	20.1%	83.2%	22.1%	4.4	39.2	36.6
El Dorado	16660	9.8%	87.0%	24.0%	1.9	42.0	40.8
Fresno	103320	14.7%	76.3%	19.7%	2.4	39.5	33.6
Kern	85477	12.8%	77.5%	19.4%	2.0	39.4	32.8
Kings	15008	12.5%	75.6%	20.0%	2.1	37.5	34.3
Madera	14573	14.3%	74.1%	18.8%	2.5	37.0	31.9
Marin	20080	20.3%	87.2%	23.4%	4.0	43.7	38.9
Merced	29488	11.9%	77.7%	19.9%	1.9	38.9	33.7
Napa	11972	17.0%	82.1%	23.4%	4.0	38.0	39.3
Orange	313576	21.4%	81.3%	20.4%	5.4	38.1	33.4
Placer	26728	14.6%	85.5%	23.4%	3.0	40.3	39.7
Riverside	193022	13.6%	82.0%	18.9%	3.2	38.7	31.2
Sacramento	137154	21.4%	80.2%	20.1%	5.0	36.1	33.4
San Bernardino	227141	15.1%	80.9%	18.9%	3.7	38.1	31.2
San Diego	292697	25.6%	79.1%	20.9%	6.2	36.5	35.0
San Francisco	43889	40.1%	74.4%	25.8%	11.4	32.4	41.8
San Joaquin	70469	12.6%	79.1%	18.5%	2.0	39.8	31.2
San Mateo	64602	25.2%	81.7%	25.4%	5.5	38.3	41.4
Santa Clara	165302	21.8%	86.0%	27.5%	4.9	43.5	45.1
Solano	45076	16.6%	82.7%	21.0%	3.9	37.2	35.2
Sonoma	45059	17.3%	83.9%	20.0%	3.9	39.0	33.7
Stanislaus	56319	12.6%	79.3%	19.8%	1.8	40.2	33.4
Sutter	9262	15.5%	80.2%	19.3%	3.8	34.8	32.3
Tulare	49784	12.4%	79.6%	19.4%	2.0	40.2	33.0
Ventura	87400	17.3%	80.8%	23.0%	4.5	37.6	38.1
Yolo	17135	21.2%	80.7%	19.6%	5.2	37.6	32.5
Yuba	7445	14.3%	80.3%	18.8%	3.4	35.5	31.9
Total	**2393831**	**19.3%**	**80.8%**	**21.0%**	**4.4**	**38.4**	**34.8**

Table 100: Adult outcome comparison.

	UF prediction – BRFSS observed difference		
County	Trans walk	Trans bike	Trans auto
Alameda	1.6	0.7	10.6
Contra Costa	-2.9	-1.8	-3.2
El Dorado	-0.1	0.5	6.8
Fresno	-1.4	-1.1	21.7
Kern	-5.2	-0.1	13.0
Kings	-2.4	-1.8	5.6
Los Angeles	-4.0	-0.4	10.6
Madera	-0.7	0.6	5.2
Marin	-3.4	-0.1	10.0
Merced	-5.2	0.5	30.4
Napa	2.0	-0.3	16.2
Orange	-2.1	-0.1	10.1
Placer	-0.5	0.7	10.8
Riverside	-2.8	0.4	5.1
Sacramento	-0.2	-0.9	2.8
San Bernardino	-1.3	0.0	11.0
San Diego	1.6	0.5	6.6
San Francisco	-6.3	-0.1	21.5
San Joaquin	-2.0	0.7	19.9
San Mateo	-0.9	-0.8	12.3
Santa Clara	1.5	-0.3	22.9
Solano	-1.6	0.6	12.2
Sonoma	0.8	-0.2	11.1
Stanislaus	-4.1	0.7	0.9
Sutter	2.9	-1.2	4.7
Tulare	-3.2	0.2	25.7
Ventura	0.5	-1.2	16.2
Yolo	0.3	-3.6	18.7
Yuba	2.8	0.5	-30.8
Total	-1.8	-0.2	11.1

Table 101: Senior outcome comparison.

	UF prediction – BRFSS observed difference	
County	Trans walk	Trans bike
Alameda	-1.7	-4.0
Contra Costa	-0.4	-3.4
El Dorado	0.7	0.7
Fresno	-0.4	1.1
Kern	-0.8	7.8
Kings	0.8	21.4
Los Angeles	-4.3	-3.8
Madera	0.3	13.3
Marin	-4.5	-2.2
Merced	0.7	24.3
Napa	-10.7	-10.5
Orange	-1.8	-3.5
Placer	-0.1	-2.9
Riverside	0.0	-6.4
Sacramento	-0.1	-20.6
San Bernardino	-1.9	-2.8
San Diego	-1.1	-5.2
San Francisco	-13.4	-9.3
San Joaquin	-3.1	7.5
San Mateo	0.0	-2.3
Santa Clara	-3.1	-6.4
Solano	-1.9	-0.8
Sonoma	-2.0	1.5
Stanislaus	-1.0	-11.5
Sutter	1.4	-11.9
Tulare	-1.4	24.0
Ventura	-2.5	-6.2
Yolo	-0.9	6.2
Yuba	1.5	-36.8
Total	-2.4	-4.0

Table 102: Teen outcome comparison.

	UF prediction – BRFSS observed difference	
County	Trans walk	Trans bike
Alameda	-5.0	6.0
Contra Costa	-7.2	-0.4
El Dorado	4.0	5.1
Fresno	-7.5	16.1
Kern	-1.0	18.9
Kings	-3.3	-72.9
Madera	4.2	37.8
Marin	-2.0	-9.3
Merced	-10.0	24.4
Napa	3.8	18.9
Orange	-2.3	-4.9
Placer	3.4	2.5
Riverside	-5.3	-2.0
Sacramento	-4.4	1.2
San Bernardino	2.0	7.9
San Diego	-0.5	0.7
San Francisco	11.9	7.0
San Joaquin	1.8	19.4
San Mateo	2.7	13.3
Santa Clara	3.8	6.2
Solano	9.4	2.2
Sonoma	5.6	-2.2
Stanislaus	-9.6	4.0
Sutter	7.9	22.5
Tulare	-1.5	21.9
Ventura	5.3	0.6
Yolo	-10.5	13.3
Yuba	2.1	-18.0
Total	-3.9	3.9

Table 103: Child outcome comparison.

	UF prediction – BRFSS observed difference	
County	Trans walk	Trans bike
Alameda	1.5	8.1
Contra Costa	0.4	13.6
El Dorado	0.7	18.6
Fresno	-5.0	14.0
Kern	-1.5	6.7
Kings	-1.9	29.9
Madera	-3.2	8.2
Marin	-1.4	13.7
Merced	-19.0	19.6
Napa	4.0	6.8
Orange	-6.7	4.0
Placer	-1.8	7.4
Riverside	-0.5	-0.4
Sacramento	1.4	6.1
San Bernardino	1.2	9.6
San Diego	-1.0	2.0
San Francisco	0.1	4.2
San Joaquin	-6.5	7.4
San Mateo	4.9	9.2
Santa Clara	-3.8	12.1
Solano	-19.0	3.8
Sonoma	-3.2	6.5
Stanislaus	-7.0	11.8
Sutter	3.8	9.2
Tulare	-7.2	8.6
Ventura	-2.8	14.9
Yolo	-0.9	3.5
Yuba	3.4	17.6
Total	-5.4	8.0

Table 104: Adult UF prediction – NHTS observed difference

	UF prediction – NHTS observed difference		
Metric	Trans walk	Trans bike	Trans auto
Predicted correlation	0.85	0.55	0.29
Mean absolute error	2.17	0.70	12.64
Mean absolute error / mean outcome ratio	0.37	1.81	0.20

Table 105: Senior UF prediction – NHTS observed difference

	UF prediction – NHTS observed difference	
Metric	Trans walk	Trans auto
Predicted correlation	0.88	-0.05
Mean absolute error	2.13	5.50
Mean absolute error / mean outcome ratio	0.39	0.09

Table 106: Teen UF prediction – NHTS observed difference

	UF prediction – NHTS observed difference	
Metric	Trans walk	Trans auto
Predicted correlation	0.32	0.18
Mean absolute error	4.32	4.61
Mean absolute error / mean outcome ratio	0.46	0.16

Table 107: Child UF prediction – NHTS observed difference

	UF prediction – NHTS observed difference	
Metric	Trans walk	Trans auto
Predicted correlation	0.05	0.15
Mean absolute error	2.90	8.13
Mean absolute error / mean outcome ratio	1.03	0.26

Notes

This memo was prepared for TAC review in advance of TAC meeting #4, and finalized based on feedback from the TAC during meeting #4 and subsequent follow-up discussions. The final models described in this report were programmed into UrbanFootprint and applied to a pilot case study for further testing. Appendix G contains details used to calculate calibration multipliers for every grid cell within the 30-county study area, in order to adjust the UF predictions to equal expected control totals

Appendix A: Responses to TAC Recommendations

TAC recommendations from the July 7, 2014 meeting are summarized below in italics, followed by brief UD4H responses. We also refer you to the location of further details related to each recommendation. In the responses below, "Report" refers to the Urban Footprint Activity-Based Public Health Module: 2nd Draft Interim Data Development and Preliminary Model Results Update, submitted on October 31, 2014.

UF Public Health Technical Advisory Committee Meeting
Summary of TAC Recommendations

July 7, 2014

TAC Recommendations on CHIS Data Development

- *Explore the effect of the disabled population data on the model.*
    - We have included the disabled population in the sample and are now adjusting for disability status in models.
    - See Report, Appendix B, p.143 for further details.
- *Make recommendations for future CHIS data collection be consistent with the census data in regard to people with disabilities.*
    - We sent a recommendation to David Grant at CHIS on 10/27/14, noting some differences in question phrasing and the discrepancy between CHIS and American Community Survey reported disability status (30% for CHIS adults and 13% for ACS in the most recent surveys).
- *Don't include people who have not lived in a location for at least a year for all variables that are not behaviorally-related.*
    - We are now including all participants in physical activity models but only participants who reported living at their address for at least one year in the body weight and health outcome models.
- *Compare samples sizes to general people in same geographies*
    - General population count, CHIS sample size, and CHTS sample size are provided at the county level in Appendix D on page 173.

TAC Recommendations on CHTS Data Development

- *UD4H to include the 7k in duplicated activities (with assumptions of what that activity was equally divided among the categories) in model runs and will keep as long as it does not introduce significant error.*
    - We have added these participants back into the sample and divided their activity time equally between all reported categories with identical start/end times.
- *Will separate bike and walk trips in the model*.

- We have separated walk and bike trips. Small numbers of reported bike trips allowed us to develop a bike model for adults only. Walk models were developed for all age groups.
- See Report, Appendix B, p.145 for further details.

TAC Recommendations on UF Data

- *To explore the barrier effect and determine feasibility of adding variables.*
  - Modifying buffers to account for barriers can be explored in the future, but is a major undertaking that is beyond the scope of this project, and may not be computationally feasible given the presence of millions of buffers in California.
  - The current UF data includes a major road (arterial or highway) density variable and a variable to indicate if a major road is located within 500 meters of a grid cell, and these variables are used in the CHIS/CHTS models.
  - The current UF data also includes intersection density and minor road (local or collector) density variables, which help adjust for overall accessibility of the 1km buffer area. These variables are used in the CHIS/CHTS models.
- *For future development, recommended incorporating sidewalk and bike lane data into the model. Will not be a part of this iteration of model development.*
  - This will definitely be revisited in the future. Limited data availability and consistency are major challenges for incorporating these variables statewide, but hopefully those data sets can be improved and standardized over time.

TAC Recommendations on UF variable multicollinearity

- *Keep park acreage and park distance separate variables.*
  - These two variables are too highly correlated to use separately in the models, so the park access index variable will continue to be used.
  - See Report, Appendix B, p.153 for further details.
- *UD4H will consider exploring the addition of open space as a land use category in the entropy measure as time and resources allow.*
  - Including park area in land use mix led to increasing multicollinearity between the walk index and the park access index, without providing any clear benefits for predictive accuracy. We have continued to include park area in the park access index but not in land use mix.
  - See Report, Appendix B, p.155 for further details.
- *Model components of the walk index separately.*
  - These variables are too highly correlated to use separately in the models, so the walkability index variable will continue to be used.
  - See Report, Appendix B, p.151 for further details.

TAC Recommendations on CHIS and CHTS Descriptive Statistics

- *Need to make sure the ethnicity categories are consistent across UD4H and UF data sets*
    - We are using a five-category race/ethnicity variable, including white/non-Hispanic, black/non-Hispanic, Hispanic, Asian, and all others. These categories are consistently available across CHIS, CHTS, and UF data sets.
- *Compare demographic and health data to county data sets (UF data does not cover the entire state)*
    - Validation metrics are provided comparing county-level CHIS model predictions to BRFSS data and CHTS model predictions to NHTS data.
        - See Report, p.113 for further details.
    - A comparison of county-level UF demographic/socioeconomic data to American Community Survey 5-year estimates is provided in Appendix F, p. 175.
- *Add disabled population to CHIS population characteristics table*
    - Descriptive statistics for the disabled population are now included in both the CHIS and CHTS tables.
        - See Report, p.34 for further details.
- *Add presence of children in the household to covariate list*
    - A binary variable "any children under 18 years old in household" was added to the CHIS and CHTS models and to the UF dataset.

Guidance Needed from TAC:

- *Follow-up email on how many cohorts for model runs. Larry will send email to group.*
    - Models have been developed for four age groups (adults, seniors, teens, and children). Additional models have been developed for adults after stratifying the sample by income group (low, medium, high). Sample sizes are too small to develop income sub-group models for seniors, teens, and children.
    - See Report, Appendix B, p.162 for further details.

Appendix B: Model development recommendations

Date: September 17, 2014

To: Gordon Garry
Director of Research and Analysis
Sacramento Area Council of Governments

From: Larry Frank & Jared Ulmer, Urban Design 4 Health

Re: UrbanFootprint Health Module -- Model development recommendations

Urban Design 4 Health (UD4H) has explored 11 analysis-related issues raised and discussed at the July 2014 TAC meeting. Based on the results of our latest draft model results, internal validation metrics, and additional exploratory analyses, we have developed the following responses and used them to guide the model development that has occurred since that July meeting. This memo is provided for your review. Please let us know if you would like to discuss it or have comments.

Note -- Models based on the approaches described below will be reported on in Draft #2 of the Model Development Validation memo (Draft #2). This memo will also be incorporated into Draft #2.

After the summary list below, commentary and justification for each recommendation are provided in the following pages.

Summary list of issues and recommendations:

- **Issue 1: Can we include the disabled population in the analysis? If so, how?**
  - **Response: Include disabled population in sample and adjust for disability status in models.**
- **Issue 2: Should we keep CHTS walk and bike models as two separate models, or combine them into a single active transportation model?**
  - **Response: For adults, fit separate walk and bike models. For all other age cohorts, due to low sample sizes for participants with any bike trips, fit only a walk model.**
- **Issue 3: Should we keep CHIS leisure walk, moderate PA, and vigorous PA models as three separate models, or combine them into a single total recreational PA model?**
  - **Response: Keep the models separate, as sample sizes are adequate and the coefficients vary widely between models.**
- **Issue 4: Should we use minutes or distance for the CHTS walk and bike outcome variable?**
  - **Response: Use minutes, as distance models have worse predictive accuracy and have generally lower sensitivity to BE changes.**
- **Issue 5: Should we separate walk index components into individual predictors?**
  - **Response: Do not separate walk index components, as multicollinearity is a problem.**
- **Issue 6: Should we include park access index components as individual predictors?**

    - **Response: Do not separate park access index components, as multicollinearity is a problem.**
- **Issue 7: Should we include park land area in the land use mix variable?**
    - **Response: Do not include park land area in land use mix, as this increases multicollinearity with the park access index and has no impact on predictive accuracy.**
- **Issue 8: Should we adjust for survey weights?**
    - **Response: Do not adjust for survey weights, as weighting factors are already included as independent variables in the models, and the resulting predictive accuracy is worse.**
- **Issue 9: Should we include random effects to adjust for household nesting and spatial autocorrelation?**
    - **Response: Do not include random effects, as the coefficient impact is minimal and the resulting predictive accuracy is worse.**
- **Issue 10: How many age and income cohorts should we develop?**
    - **Response: Include all four age cohorts (adults, seniors, teens, and children). For adults, because the sample size is uniformly adequate, develop income group-specific models (low, medium, high). For seniors, teens, and children, because the sample size is too small for some outcomes, develop only a single (pooled) income cohort.**
- **Issue 11: For any income-specific (low, medium, high) models, what model fitting approach should we use?**
    - **Response 1: For Draft #2 of the Model Development and Validation memo, adult income-specific models have been fit using same set of variables selected for pooled income model.**
    - **Response 2: For the final set of models, uniquely select variables specific to each income cohort.**

Issue 1: Can we include the disabled population in the analysis? If so, how?

Response: Include disabled population in sample and adjust for disability status in models.

Commentary:

- Including the disabled population increases the CHTS sample by about 4,000 participants and the CHIS sample by about 7,400 participants.
- The disability status variable is categorical with 3 categories:
    - no disability
    - non-ambulatory disability
    - ambulatory disability
- Including the disabled population and adjusting for disability status results in very minimal change in the built environment variable coefficients:

Cohort:	Adults			
Outcome:	Likelihood of any transportation walking			
Model:	Binary logistic regression			
Version:	Includes disabled population and disability status variable		Excludes disabled population	
	coefficient	p-value	coefficient	p-value
walkability index	0.07	7.E-37	0.07	5.E-34
transit access	0.03	2.E-07	0.03	0.0001
rail access	0.31	3.E-11	0.32	1.E-10
major road index				
regional access	0.04	0.0001	0.05	1.E-05
school distance	-0.06	9.E-22	-0.07	3.E-19
park access	0.09	1.E-16	0.09	2.E-15

- The disability status variable was significantly associated with the outcome in the majority of models.

- Predictive accuracy is very similar whether including the disabled population or not:
    - Higher ***predicted r^2*** is better, where 0 = no correlation, and 1 = perfect correlation.
    - Lower values for the ***mean absolute error : mean outcome ratio*** are better, where a value approaching zero indicates nearly perfect mean predictive accuracy.

		zip code level[gg] accuracy metrics	
Model	Age cohort	predicted r^2	mean absolute error : mean outcome ratio
Any transportation walking	Including disabled	0.47	0.37
Any transportation walking	Excluding disabled	0.45	0.38
Any recreational PA	Including disabled	0.18	0.33
Any recreational PA	Excluding disabled	0.17	0.34

[gg] Aggregating predictions to the zip code level was mainly based on the need to aggregate to a scale with a reasonable CHTS sample size per spatial unit. Zip code level is not intended to be a recommendation for the reporting scale in Urban Footprint.

Issue 2: Should we keep CHTS walk and bike models as two separate models, or combine them into a single active transportation model, or model walking only?

Response: For adults, fit separate walk and bike models. For all other age cohorts, due to low sample sizes for participants with any bike trips, fit only a walk model.

Commentary:

- Sample size of participants with >0 minutes of biking is relatively very small:
  - Adults, n = 801 out of 35,695
  - Seniors, n = 90 out of 8,475
  - Teens, n = 207 out of 4,734
  - Children, n = 118 out of 4,829
- Sample size seems adequate for adults, but for all other age groups, there are not participants in every covariate category:
  - For example, for children, including "race/ethnicity" in the model improves model fit, but no bike trips were made by "American Indian/Alaska native", "Native Hawaiian/Pacific islander", or "Other" groups. The result is that the "race/ethnicity" coefficient is illogical for several of the groups. This could be resolved by further collapsing "race/ethnicity" categories or by removing "race/ethnicity" from the model. Alternatively, we could combine walk and bike models or remove the bike model entirely.
  - These types of customizations can be done for each of the bike models, but this would not resolve the fact that the remaining coefficient values will have been calibrated based on small sample sizes, resulting in associations with low statistical significance. Also such customization requires substantial manual effort above and beyond the general model fitting procedures.
  - Teen bike model cannot currently be run without customizations, as some of the categorical variables include participants in only category (which results in an error when trying to fit the model).

- With its reasonably large sample size, the adult bike model coefficients are very similar to walk model or walk + bike model coefficients. We will fit a separate bike model for adults only, though, as shown below, the outcome will be very similar to the separate walk model and a single combined active transportation model:

Cohort:	Adults					
Outcome:	Likelihood of active transportation					
Model:	Binary logistic regression					
Version:	Walk only		Bike only		Walk + bike	
	coefficient	p-value	coefficient	p-value	coefficient	p-value
walkability index	0.07	7.E-35	0.05	0.0001	0.07	6.E-37
transit access	0.03	1.E-05	0.03	0.02	0.03	2.E-07
rail access	0.31	1.E-10	0.29	0.01	0.32	2.E-11
major road index			-0.03	0.14	-0.01	0.40
regional access	0.05	1.E-05			0.04	4.E-05
school distance	-0.06	2.E-18	-0.05	0.0005	-0.07	7.E-22
park access	0.09	8.E-15	0.10	0.0001	0.09	1.E-16

- But for children, very few built environment variables are significant, and one of the coefficients (regional access) is positive in the walk or walk + bike model but negative in the bike only model:

Cohort:	Children					
Outcome:	Likelihood of active transportation					
Model:	Binary logistic regression					
Version:	Walk only		Bike only		Walk + bike	
	coefficient	p-value	coefficient	p-value	coefficient	p-value
walkability index	0.06	1.E-06	0.06	0.03	0.06	2.E-06
transit access					0.02	0.31
rail access	0.16	0.20			0.14	0.27
major road index	-0.04	4.E-02			-0.05	0.01
regional access	0.08	0.0003	-0.08	0.16	0.06	0.01
school distance	-0.08	1.E-06			-0.07	9.E-06
park access	0.05	0.05	0.15	0.02	0.06	0.01

- Predictive accuracy is relatively poor for bike models as compared to walk or walk + bike models, especially for children & seniors.
    - Higher ***predicted r^2*** is better; lower values for the ***mean absolute error : mean outcome ratio*** are better

		zip code-level accuracy metrics	
Model	Age cohort	predicted r^2	mean absolute error : mean outcome ratio
walk minutes	Adult	0.42	0.45
bike minutes		0.05	1.00
walk + bike minutes		0.41	0.45
walk minutes	seniors	0.26	0.88
bike minutes		0.03	1.62
walk + bike minutes		0.25	0.85
walk minutes	Teens	0.13	0.79
bike minutes		error	error
walk + bike minutes		0.11	0.74
walk minutes	children	0.20	0.84
bike minutes		0.02	1.71
walk + bike minutes		0.14	0.84

Issue 3: Should we keep CHIS leisure walk, moderate PA, and vigorous PA models as three separate models, or combine them into a single total recreational PA model?

Response: Keep the models separate, as sample sizes are adequate and the coefficients vary widely between models.

Commentary:

- This decision is only applicable to CHIS adult and senior models.
- Sample size is adequate for all model options:
    - Smallest sample size is for senior vigorous PA model, where n = 1,985 participants with >0 minutes of vigorous PA.
- Built environment variable coefficients differ somewhat across models, though are generally similar:

Cohort:	Adults							
Outcome:	Likelihood of recreational PA							
Model:	Binary logistic regression							
Version:	Leisure walk		Moderate PA		Vigorous PA		All recreational PA	
	coefficient	p-value	coefficient	p-value	coefficient	p-value	coefficient	p-value
walkability index	-0.01	0.06	-0.03	3.E-16	-0.01	0.00	-0.02	3.E-05
transit access					0.01	0.23		
rail access	-0.05	0.31			-0.09	0.11	-0.07	0.28
major road index	-0.01	0.26					-0.02	0.05
regional access	0.02	0.05	0.01	0.29			0.02	0.05
school distance			-0.01	0.22				
park access	0.02	0.04			0.03	0.00	0.03	0.02

- Predictive accuracy does not vary much by recreational PA model, with the exception of the senior vigorous PA model having relatively worse predictive accuracy:
    - Higher ***predicted r^2*** is better; lower values for the ***mean absolute error : mean outcome ratio*** are better

		zip code level accuracy metrics	
Model	Age cohort	predicted r^2	mean absolute error : mean outcome ratio
Leisure walking minutes	adult	0.08	0.26
Moderate PA minutes		0.05	0.31
Vigorous PA minutes		0.07	0.38
Recreational PA minutes		0.06	0.21
Leisure walking minutes	seniors	0.05	0.36
Moderate PA minutes		0.04	0.40
Vigorous PA minutes		0.01	0.67
Recreational PA minutes		0.05	0.29

- One benefit of keeping these models separate is that we can convert each type of physical activity to METs, allowing us to more precisely calculate total METs, which is then used as a predictor variable in body weight and health outcome models.

Issue 4: Should we use minutes or distance for the CHTS walk and bike outcome variable?

Response: Use minutes, as distance models have worse predictive accuracy and have generally lower sensitivity to BE changes.

Commentary:

- Minutes were self-reported by CHTS participants.
- Distance was estimated by CALTRANS using Google Maps, based on the self-reported origin and destination of each walk or bike trip.
- Coefficients are similar whether using minutes or distance:

Cohort:	Adults			
Outcome:	Minutes/distance of transportation walking			
Model:	Linear regression			
Version:	Minutes		Distance	
	coefficient	p-value	coefficient	p-value
walkability index	0.02	2.E-10	0.01	0.00
transit access				
rail access	0.08	0.01	0.09	0.04
major road index				
regional access	0.02	0.05	0.02	0.03
school distance				
park access	0.02	0.08	0.01	0.20

- But predictive accuracy is much better for minutes than for distance:
  - Higher ***predicted r^2*** is better; lower values for the ***mean absolute error : mean outcome ratio*** are better

		zip code level accuracy metrics	
Model	Age cohort	predicted r^2	mean absolute error : mean outcome ratio
walk minutes	adult	0.42	0.45
walk miles	adult	0.01	0.95
walk minutes	seniors	0.26	0.88
walk miles	seniors	0.12	0.95
walk minutes	teens	0.13	0.79
walk miles	teens	0.00	1.15
walk minutes	children	0.20	0.84
walk miles	children	0.05	0.94

Issue 5: Should we separate walk index components into individual predictors?

Response: Do not separate walk index components, as multicollinearity is a problem.

Commentary:

- Walkability index consists of:
  - Street connectivity index
  - Residential access index
  - Commercial access index
  - Land use mix
- Each walkability index component further consists of several individual correlated variables:
  - street connectivity index consists of intersection density and local street length
  - residential access index consists of dwelling unit count and residential density
  - commercial access index consists of retail floor area, non-residential floor-area ratio, distance to nearest retail, and distance to nearest restaurant
  - land use mix is the entropy of five land use categories: residential, retail & services, restaurant & entertainment, office, and public administration
- This decision has no impact on the UrbanFootprint user. In either case, the user will only modify the values of the most basic component parts (e.g. adding multi-family residential units to a grid cell). The user will never directly modify a "walkability index" value. Behind the scenes, UrbanFootprint will do all of the calculations needed to derive the walkability index from its component parts.
  - As such, we can evaluate the health impact in UrbanFootprint of increasing residential density (or any of the component variables) alone, rather than only being able to evaluate the impact of increasing the walkability index.
  - But if a walkability index is used, we cannot evaluate the health impact of increasing residential density (or any of the component variables) alone directly from a model coefficient. We will only have a single coefficient for the walkability index.
- The walkability index components are highly correlated:

	connectivity	residential	commercial	mix
Street connectivity index	1			
Residential access index	0.79	1		
Commercial access index	0.67	0.78	1	
Land use mix	0.42	0.48	0.72	1

- Multicollinearity leads to counterintuitive model results. For example, in the binary regression model predicting any adult transportation walking (versus none):

- All walkability index components are strongly and positively associated with walking when entered one-at-a-time.
- When entered simultaneously, the sign on the commercial access index coefficient flips to negative, and the association with land use mix is no longer significant.
- Following variable selection, the commercial access index coefficient remains negative, while the land use mix variable is no longer included.

Cohort:	Adults					
Outcome:	Likelihood of any transportation walking					
Model:	Binary logistic regression					
Version:	BE variables entered one-at-a-time		All BE variables entered simultaneously		Final model following variable selection	
	coefficient	p-value	coefficient	p-value	coefficient	p-value
Street connectivity index	0.22	1.E-119	0.06	3.E-05	0.06	3.E-05
Residential access index	0.27	7.E-153	0.14	3.E-14	0.14	4.E-14
Commercial access index	0.11	2.E-81	-0.02	0.07	-0.01	0.09
Land use mix	1.15	5.E-35	0.10	0.46		

- The variance inflation fractions (which indicate multicollinearity) for the binary regression model predicting any adult transportation walking (versus none) are much higher for the walkability index component variables than for the walkability index itself:

	Variance Inflation Fractions (which indicate multicollinearity)	
	Model with walkability index	Model with separate walkability index components
Walkability index	1.9	
Street connectivity index		2.2
Residential access index		3.9
Commercial access index		3.9
Land use mix		2.1
transit access	1.6	1.6
rail access	1.3	1.3
major road index	1.1	1.1
regional access	1.5	1.6
school distance	1.2	1.2
park access	1.1	1.1

Issue 6: Should we include park access index components as individual predictors?

Response: Do not separate park access index components, as multicollinearity is a problem.

Commentary:

- Park access index consists of:
  - Park acres
  - Distance to nearest park
- This decision has no impact on the UrbanFootprint user. In either case, the user will only modify the values of the most basic component parts (e.g. adding park acreage to a grid cell). The user will never directly modify a "park access index" value. Behind the scenes, UrbanFootprint will do all of the calculations needed to derive the park access index from its component parts.
  - As such, we can evaluate the health impact in UrbanFootprint of increasing park acreage (or decreasing park distance) alone, rather than only being able to evaluate the impact of increasing the park access index.
  - But if a park access index is used, we cannot evaluate the health impact of increasing park acreage (or decreasing park distance) alone directly from a model coefficient. We will only have a single coefficient for the park access index.
- The park access index components are highly correlated:

	Park acres	Park distance
Park acres	1	
Park distance	-0.59	1

- Multicollinearity leads to counterintuitive model results. For example, in the binary regression model predicting any adult transportation walking (versus none):
    - Both park index components are strongly associated with walking when entered one-at-a-time. Park acres is positively associated and park distance is negatively associated.
    - When entered simultaneously, the sign on the park distance coefficient flips to positive and the association is no longer significant.
    - Following variable selection, the park distance variable is no longer included.

Cohort:	Adults					
Outcome:	Likelihood of any transportation walking					
Model:	Binary logistic regression					
Version:	BE variables entered one-at-a-time		All BE variables entered simultaneously		Final model following variable selection	
	coefficient	p-value	coefficient	p-value	coefficient	p-value
Park acres	0.21	6.E-41	0.15	8.E-14	0.15	2.E-18
Park distance	-0.02	1.E-31	0.00	0.78		

- The variance inflation fractions (which indicate multicollinearity) for the binary regression model predicting any adult transportation walking (versus none) are higher for the park access index component variables than for the park access index itself:

	Variance Inflation Fractions (which indicate multicollinearity)	
	Model with park access index	Model with separate park access index components
walkability index	1.9	1.9
transit access	1.6	1.6
rail access	1.3	1.3
major road index	1.1	1.1
regional access	1.5	1.5
school distance	1.2	1.2
park access	1.1	NA
park acres	NA	1.5
park distance	NA	1.4

Issue 7: Should we include park land area in the land use mix variable?

Response: Do not include park land area in land use mix, as this increases multicollinearity with the park access index and has no impact on predictive accuracy.

Commentary:

- Currently, the land use mix index is based on the building floor area of five land uses types:
  - Residential
  - Retail & services
  - Restaurant & entertainment
  - Office
  - Public administration
- Land use mix is then included in the walkability index.
- A second version of the walkability index was created, which includes a new land use mix variable that adds park land area.
- Park land area is already included in the park access index variable.
- Adding park land area to the mixed use index increased the correlation between the walkability index and the park access index from 0.30 to 0.35.
- Adding park land area to the mixed use index would be desirable if these two variables were highly correlated, but they are not (r=0.07).
- In a model like the binary regression model predicting any adult transportation walking (versus none), where the walkability index and park access index both have a positive association with the outcome, adding park area to the land use mix index results in a marginal increase in the walkability index coefficient, but this is offset by a decrease in the park access index coefficient.
  - There is essentially no difference between these two models in terms of predictive accuracy or to the sensitivity of the outcome in response to built environment changes.

Cohort:	Adults			
Outcome:	Likelihood of any transportation walking			
Model:	Binary logistic regression			
Version:	Original land use mix		Land use mix with park area	
	coefficient	p-value	coefficient	p-value
walkability index	0.07	7.E-37	0.07	6.E-36
transit access	0.03	2.E-07	0.03	2.E-05
rail access	0.31	3.E-11	0.31	2.E-10
major road index				
regional access	0.04	0.0001	0.05	8.E-06
school distance	-0.06	9.E-22	-0.06	8.E-18
park access	0.09	1.E-16	0.07	2.E-11

- In a model where the walkability index and park access index have opposite effects (e.g. walkability index is negatively associated and park access is positively associated with recreational PA), adding park area to the land use mix index will result in a weaker walkability index component and a stronger park access index component.
- The use of separate land use mix and park access variables already effectively measures the impact of both of these variables without needing them to be included in a single variable. To maximize the likelihood of transportation walking, a neighborhood needs both high land use mix and high park access.
- The only impact of adding park area to the mixes use index is to shift the coefficient strength between walkability index and park access index variables, but does not change the actual effect of a change in park area.

Issue 8: Should we adjust for survey weights?

Response: Do not adjust for survey weights, as weighting factors are already included as independent variables in the models, and the resulting predictive accuracy is worse.

Commentary:

- Survey weights are used in regression modeling to generate unbiased estimates of the model coefficients.
- Survey weights for CHIS and CHTS are based on:
    - Sampling probability
    - Age
    - Sex
    - Race/ethnicity
    - Educational attainment
    - Household size
    - Household income
    - Workers per household
    - Vehicles per household
    - County of residence
- The CHTS documentation cautions against using weights at a sub-state level.
- Published guidance suggests that weights may not be needed for regression analyses, especially when the weights are a function of the independent variables.[hh]
    - With the exception of county of residence, all of the factors used for calculating the weights are included in our models as independent variables.
    - County is the only weight factor not included as an independent variable. (see further commentary on including county at the end of this section)

[hh] Winship C, Radbill L. Sampling Weights and Regression Analysis. Sociological Methods Research. 1994; 23(2): 230-257.

- Adjusting for weights does result in some difference to the CHTS/CHIS model coefficients. For example, in the binary regression model predicting any adult transportation walking (versus none):
    - The coefficient for rail access dropped by almost 1/3rd
    - Major road index was included in the weighted model but not the unweighted model
    - All other coefficient values were similar

Cohort:	Adults			
Outcome:	Likelihood of any transportation walking			
Model:	Binary logistic regression			
Version:	No weights		With weights	
	coefficient	p-value	coefficient	p-value
walkability index	0.07	7.E-37	0.08	5.E-43
transit access	0.03	2.E-07	0.02	0.00
rail access	0.31	3.E-11	0.20	1.E-05
major road index			0.01	0.10
regional access	0.04	0.0001	0.06	1.E-09
school distance	-0.06	9.E-22	-0.05	8.E-13
park access	0.09	1.E-16	0.07	8.E-11

- Predictive accuracy appears to be worse for models including weights. For example, in the binary regression model predicting any adult transportation walking (versus none):
    - Using the model without weights, the predicted likelihood of walking is 0.6% lower than the observed prevalence of walking.
    - Using the model with weights, the predicted likelihood of walking is 4.4% lower than the observed prevalence of walking.
    - Other predictive accuracy metrics are only marginally worse for the model with weights as compare to the model without weights.

- Because county is correlated with built environment variables, including county as an independent variable would result in reducing the effect of the built environment variables (i.e. the effect would shift somewhat from the built environment variables to the county variable). Using the same example of the binary regression model predicting any adult transportation walking (versus none):
    - The value of all coefficients except for transit access decreased.

Cohort:	Adults			
Outcome:	Likelihood of any transportation walking			
Model:	Binary logistic regression			
Version:	Excluding county variable		Including county variable	
	coefficient	p-value	coefficient	p-value
walkability index	0.07	7.E-37	0.05	9.E-15
transit access	0.03	2.E-07	0.05	2.E-08
rail access	0.31	3.E-11	0.21	3.E-05
major road index				
regional access	0.04	0.0001	0.03	0.03
school distance	-0.06	9.E-22	-0.01	0.21
park access	0.09	1.E-16	0.04	0.00

Issue 9: Should we include random effects to adjust for household nesting and spatial autocorrelation?

Response: Do not include random effects, as the coefficient impact is minimal and the resulting predictive accuracy is worse.

Commentary:

- Several types of nesting or spatial autocorrelation are present in the CHTS & CHIS data:
    - CHTS persons are nested within CHTS households
    - CHTS/CHIS participants are nested within UrbanFootprint grid cells.
    - People and built environments in nearby grid cells tend to be more similar than people and built environments in distant grid cells.
- Nesting is often dealt with by fitting random intercepts to each group level. For example, an intercept can be calculated for each household, as we expect persons within a household to be more similar than persons across households.
- Spatial autocorrelation can also be dealt with by fitting random intercepts to a spatial level. For example, because persons are nested within Census tracts, we could fit an intercept for each tract to model our expectation that persons within a household should be more similar than persons across tracts.
    - This suggestion for dealing with spatial autocorrelation was proposed during the first TAC meeting, due to its relative simplicity to implement (whereas other options for dealing with spatial autocorrelation tend to be much more complex).
    - It was also noted by Sue Babey that CHIS researchers have explored the impact of spatial autocorrelation but haven't found much effect.
- Typically, the use of random effects should have minimal impact on coefficient values, but will increase standard errors, and thus higher p-values, resulting in more conservative conclusions to be drawn from model results.
    - Because coefficient values are more relevant for this analysis than are p-values, spatial autocorrelation should not be a major issue.

- Using random intercepts to adjust for nesting and spatial autocorrelation does not seem to make much difference to the CHTS/CHIS models. For example, in the binary regression model predicting any adult transportation walking (versus none):
    - Coefficient values are very similar whether including random effects or not.
    - P-values are lower in the model with random effects, which is expected.

Cohort:	Adults			
Outcome:	Likelihood of any transportation walking			
Model:	Binary logistic regression			
	No random effects		With random effects	
	coefficient	p-value	coefficient	p-value
walkability index	0.07	7.E-37	0.07	4.E-25
transit access	0.03	2.E-07	0.03	0.0001
rail access	0.31	3.E-11	0.31	5.E-08
major road index				
regional access	0.04	0.0001	0.05	0.0002
school distance	-0.06	9.E-22	-0.06	2.E-13
park access	0.09	1.E-16	0.08	4.E-11

- One problem related to the use of random effects is that we can only fit random intercepts for the households and spatial units where we have CHIS or CHTS participants. This creates problems in the UrbanFootprint application:
    - Households do not exist as discrete units in UrbanFootprint, so there is no way to apply the random intercepts for households.
    - Below the county level, we do not have CHIS or CHTS participants for every possible spatial unit. For example, if tract 101 in San Diego County does not include any CHIS/CHTS participants, we can't fit an intercept for this tract. We can only apply random intercepts to tracts that we have CHIS or CHTS data for.
    - We can fit the models with random effects, then apply them in UrbanFootprint using only the fixed effects, but that approach results in worse predictive accuracy. For example, in the binary regression model predicting any adult transportation walking (versus none):
        - Using the model without random effects, the predicted likelihood of walking is 0.6% lower than the observed prevalence of walking.
        - Using the model with random effects but applying only the fixed effects portion, the predicted likelihood of walking is 4.3% lower than the observed prevalence of walking.
        - Other predictive accuracy metrics are only marginally worse for the model with random effects as compare to the model without random effects.

Issue 10: How many age and income cohorts should we include for each final model in UrbanFootprint?

Response: Include all four age cohorts (adults, seniors, teens, and children). For adults, because the sample size is uniformly adequate, develop income group-specific models (low, medium, high). For seniors, teens, and children, because the sample size is too small for some outcomes, develop only a single (pooled) income cohort.

Commentary:

- Currently, we are developing 66 unique models by age cohort:
  - Adults = 23 models/outcome
  - Seniors = 23 models/outcome
  - Teens = 10 models/outcome
  - Children = 10 models/outcome
- For each of these age cohorts, we have also investigated four models per outcome:
  - All incomes
  - Low
  - Medium
  - High
- Generating models by income cohort results in a total of 198 unique models (66 models x 3 income groups = 198).
  - For example, for the outcome of BMI, there could be up to 12 unique models (4 age groups x 3 income groups), which means that each built environment variable could have 12 slightly different interpretations depending on the age/income cohort.
  - Reviewing, cleaning, and programming this many models into UrbanFootprint will be challenging and time consuming.
  - Consideration should also be given to application of the tool, and the number/type of desired future scenario outcomes.
  -
- The maximum sample sizes by age/income cohort are as follows:

CHTS sample sizes	income			
Age	all	low	med	high
Adult	35695	10593	11283	13819
Seniors	8475	3191	3137	2147
Teens	4734	1479	1270	1985
Children	4829	1670	1378	1781

CHIS sample sizes	income			
	all	low	med	high
Adult	23515	9188	6537	7790
Seniors	11618	6448	3296	1874
Teens	2367	874	577	915
Children	3117	1159	757	1201

  - Developing models by age cohort is not a problem, since the sample sizes are reasonably large. For CHIS, developing models by age cohort is somewhat necessary, as the surveys differ by age cohort.
  - Sample sizes for age/income cohorts are adequate for adults, less so for seniors, and start getting very small for teens and children.
  - While the lowest sample size shown above is n=577 (CHIS Teens Med income) the available sample sizes are much smaller for certain models, such as "minutes of recreational PA for those with >0 minutes of recreational PA." For this model for teens, the sample sizes are relatively small:
    - Pooled incomes: 1,039
    - Low income: 232
    - Medium income: 264
    - High income: 543
  - Similarly, binary outcomes require an adequate sample size in both outcome categories, which can be challenging for "rare" outcomes. For example, we have data on self-reported health status for 2,885 children, but the sample size for those reporting poor health is very low for the income groups:
    - 157 out of 2,885 for pooled incomes
    - 116 out of 1029 for low incomes
    - 20 out of 702 for medium incomes
    - 21 out of 1,154 for high incomes
  - Body weight and health outcome models also have smaller sample sizes, as these models exclude participants who have been living at their address for less than 1 year.
  - Small sample sizes are exacerbated by the presence of categorical variables, such as race/ethnicity, because each model needs to have a reasonable sample within each variable category. This is even more challenging for binary outcomes, as we need a reasonable sample within each variable category for both outcome categories.
- Small sample sizes can result in similar problems to those discussed for the bike only models earlier. Participants may not be included in every covariate category, leading to some illogical or missing coefficients, while other categorical coefficients may be based on very small sample sizes.

- To ensure the validity of the income-specific models, we will need to review and customize every model by collapsing covariate categories, removing binary covariates where the sample size for one of the categories is very small, etc. That approach will require significant effort for every model.
- That approach doesn't necessarily resolve the fact that the remaining coefficient values may have been calibrated based on small sample sizes, resulting in associations with low statistical significance.
- Using the standard variable selection and model fitting approach, several senior, teen, and child models currently generate errors that are related to small sample sizes.

Issue 11: For any income-specific models, what model fitting approach should we use?

Response 1: For Draft #2 of the Model Development and Validation memo, adult income-specific models have been fit using same set of variables selected for pooled income model.

Response 2: For the final set of models, uniquely select variables specific to each income cohort.

Commentary:

- For the age/income models completed to date (and which will be reported on in Draft #2 of the Model Development and Validation memo), variable selection has been done for only the pooled income cohort, then coefficients are refit for the same variable selection for each of three income cohorts (low, medium, and high).
- Alternatively, we could run variable selection uniquely for each income cohort.
    - The benefit of this approach is:
        - We may end up with a different and more appropriate set of variables in the model.
            - For example, transit access may be associated with walking only for those with low income but not for those with medium or high income. Running variable selection for only the pooled income cohort may result in the transit access variable being excluded from the model, whereas running variable selection for each cohort may result in transit access being included in the low income model but not in the medium or high income model.
    - The downsides are:
        - this will require almost 4 times as long to process the models
        - low sample sizes for income group-specific models may result in relatively lower statistical significance of built environment variables as compared to the pooled income model, meaning that fewer variables could end up being selected for the final model (solely based on low statistical power and not necessarily because there was any actual difference in the coefficient).
- In some cases, it appears that running variable selection for each income cohort will result in a better model:

Cohort:	Adults, all incomes					
Outcome:	Likelihood of any recreational PA					
Model:	Binary logistic regression					
	Pooled incomes		Medium income		Medium income	
	Pooled variable selection		Pooled variable selection		Unique variable selection	
	coefficient	p-value	coefficient	p-value	coefficient	p-value
walkability index	-0.01	0.00	-0.02	0.01	-0.02	0.02
transit access					-0.02	0.07
rail access					0.13	0.15
major road index						
regional access					0.03	0.05
school distance						
park access	0.06	2.E-11	0.05	0.0003	0.06	0.0002

- When using the variables selected for the pooled income model in the medium income model, only two built environment variables are selected. But running variable selection uniquely for the medium income model would result in three additional variables being selected.

- In other cases, it's less clear which is the "better" model:

Cohort:	Children					
Outcome:	Likelihood of any transportation walking					
Model:	Binary logistic regression					
	Pooled incomes		Medium income		Medium income	
	Pooled variable selection		Pooled variable selection		Unique variable selection	
	coefficient	p-value	coefficient	p-value	coefficient	p-value
walkability index	0.06	9.E-07	0.08	0.00	0.05	0.08
transit access					0.11	0.00
rail access	0.17	0.19	0.33	0.25		
major road index	-0.04	0.04	-0.06	0.21	-0.07	0.16
regional access	0.09	0.0002	0.01	0.85		
school distance	-0.08	1.E-06	-0.10	0.00	-0.09	0.01
park access	0.05	0.04	-0.05	0.32		

- When using the variables selected for the pooled income model in the medium income model, three variables remain which would otherwise be excluded if conducting variable selection uniquely for the medium income cohort. The coefficient for rail access is much larger for the medium income model, but the p-value is worse (in part due to a smaller sample size). When conducting variable selection uniquely for the

medium income model, rail access is dropped, so the rail access coefficient no longer has any effect. On the flip side, transit access is now included and is statistically significant.

- For either of the examples above, the variable selection method had a negligible impact on predictive accuracy.

Appendix C: Recommended approach to dealing with multicollinearity

After applying the previously agreed-upon model development methodology, several models exhibited evidence that multicollinearity between the walkability index, transit access, rail access, and regional access variables were leading to counterintuitive results. In several models, one of these variable coefficients had a sign in the opposite direction as found in the bivariate correlation and in the preliminary version of the model where each built environment variable was entered one-at-a-time. This memo summarizes our findings after exploring potential causes of these counterintuitive results and testing one potential solution, combining the walkability index and transit access variables into a single variable.

Exploring potential data problems

We reviewed the distributions and potential outliers for the predictor and outcome variables but did not uncover any previously unidentified concerns.

As an additional test to identify if UrbanFootprint data for any specific region was causing any problems, we also ran all of the CHIS and CHTS adult and senior models after stratifying the samples by region. This test did not provide any clear evidence of data problems within any specific region. In general, the model results were very consistent across regions. For example, the walkability index was not associated with adult overweight in any region except for the Bay Area. In the Bay Area, the walkability index association is negative with a p-value of 0.03 when entered alone, and negative with a p-value of 0.21 in the final model. In the final pooled (all-region) model, the walkability index was dropped from the adult overweight model, which seems consistent with the region-specific results.

A second example is that the walk index was significantly associated with less moderate physical activity (PA) in every region except for the San Joaquin Valley, where the association was negative but the p-value was only 0.18. In the final pooled (all-region) model, the walkability index was negatively and significantly associated with adult moderate PA, which seems consistent with the region-specific results.

Exploring potential multicollinearity problems

The correlation between the walkability index and transit access variables was 0.58, and this strong association was flagged as early as the June 30th preliminary model development memo as a potential problem to revisit. Transit access was originally left out of the walkability index, as it (barely) did not meet the criteria for inclusion, and every effort was being made to retain built environment variables outside of the walkability index whenever possible.

However, in the latest draft models, we noted five cases in the age-group specific models where either the walkability index or the transit access variables demonstrated an opposite sign when comparing the bivariate correlation and the preliminary version of the model where each built environment variable was entered one-at-a-time to the final model where both variables were entered simultaneously. In most other final models, both variables were included and have the same sign or only one of the two variables was included. In 10 cases (mostly commonly with senior & teen models), both variables were

included in the final model and had opposite signs, but those signs matched the sign of the bivariate correlations and the preliminary version of the model where each built environment variable was entered one-at-a-time. We also speculated that in some models, this multicollinearity was causing the walkability index to be dropped from the final model due to low coefficient strength, despite a strong association ($p<0.001$) in the preliminary version of the model where each built environment variable was entered one-at-a-time.

To explore the impact of combining the walkability index and transit access variables, we re-ran the CHTS and CHIS adult/senior models after adding the transit access variable into the walkability index. Summary results from these models are presented in Appendix A. Aside from some moderate changes in walkability index coefficient strength, the following major changes to walkability index coefficients were noted:

- The walkability index was dropped from the adult final BMI, overweight, or obesity models
    - Originally, the walkability index was in the final models and positively related in each, while the transit access variable was negatively related in each.
    - Using the new walkability index variable, the walkability index is negatively & significantly associated with each of the three weight outcomes in the preliminary version of the model where each built environment variable was entered one-at-a-time.
        - It appears that multicollinearity continues to affect the walkability index, which keeps it from being included in the final model. The most likely culprit is the "regional accessibility" variable, which is strongly and negatively associated with each of the three weight outcomes.
- The walkability index is now included in the final adult high blood pressure model, where it is negatively but weakly associated ($p=0.26$)
- The walkability index was dropped from the senior leisure walking model (none v. any)
    - Originally, the walkability index was in the final model and negatively but weakly associated, while the transit access variable was positively and significantly associated.
- The walkability index is now included in the senior overweight model, where it is negatively but weakly associated ($p=0.22$)

Even after combining the walkability index and transit access variables, there continues to be evidence of multicollinearity in a small number of models, mostly with regional accessibility but to a lesser extent with rail access. The walkability index has a correlation of 0.48 with regional accessibility and 0.33 with rail access. In only a few cases, this results in a variable being included in a final model with sign that was opposite from that found for the bivariate correlation and the preliminary version of the model where each built environment variable was entered one-at-a-time. More often, the result is a reduction in coefficient strength, sometimes leading to one or more of these variables being dropped from the final model, despite a strong association ($p<0.001$) in the one-at-a-time model.

Potential solutions to these multicollinearity issues include:

- **Recommended approach:** We can selectively combine built environment variables into the walkability index ONLY in cases where there is evidence that multicollinearity is a problem. To do so, we would compare the preliminary version of each model where each built environment variable was entered one-at-a-time to the final version. For only the walkability index, transit access, rail access, and regional access variables, we will note if any were significantly associated ($p<0.05$) with the outcome in the preliminary model, but either dropped from the final model or included in the final model but with an opposite sign from the preliminary model. Any variables meeting this criterion will be combined into the walkability index for the final model.
- We can continue to combine built environment variables into the walkability index for use in all models. While this will reduce multicollinearity, there are at least two major drawbacks:
    - Doing so would predetermine that all component variables in the walkability index MUST have the same directional relationship with the outcome (i.e. they MUST all be positive or all be negative). Because some of these variables legitimately appear to have opposite signs in some models, this change would at best result in the walkability index being dropped from the model, and at worst may result in including component variables with an opposite directional relationship (when compared to the one-at-a-time model) to be included in the final model.
    - Fewer unique built environment variables will likely have a negative impact on predictive accuracy and will further complicate the ability to interpret the coefficients.
- We can selectively drop variables from final models where the coefficient in the final model had an opposite sign from that found for the bivariate correlation and the preliminary version of the model where each built environment variable was entered one-at-a-time.

Applied approach

In response to the multicollinearity problems described above, we have developed and implemented the following approach. In cases where there was evidence that multicollinearity was a problem we have selectively added problematic variables (transit access, rail access or regional access) into the walkability index.

The process we applied was as follows:

1. We compared the coefficients and p-values for the four potentially multicollinear built environment variables (walk index, walkability, transit access, rail access, regional access) from the preliminary version of each model (where each built environment variable was entered one-at-a-time) to the final version.
2. For any of the four potentially multicollinear built environment variables, if the sign of the variable's coefficient in the preliminary model was opposite of the sign in the final model, we applied one of two remedies:
    a. If the variable coefficient's p-value in the preliminary model was < 0.05, we added the variable to the walk index with a sign that matches that found in the preliminary model.

We weighted the components of the revised walkability index according to the relative t-values for each component variable in the preliminary model.

   i. For example, in the preliminary binary logistic model of any senior transportation walking, the walkability index had a strong positive association ($p<0.001$) as did transit access ($p<0.001$). However, when they are added simultaneously, the walkability index continues to have a strong positive association ($p<0.001$), but the sign on transit access became negative with a non-significant association ($p=0.21$). Following the variable selection methodology described on page 46, both variables were retained in the final model. As the only reason for the change in sign for the transit access coefficient appears to be due to multicollinearity, we propose to create a new walkability index variable that includes the original walkability index and transit access variables. Based on the higher t-value for the original walkability index ($t=12.7$) in the preliminary model as compared to transit access ($t=6.6$), we would assign a weight of $12.7/6.6=1.92$ to the walkability index and a weight of 1.0 to transit access. The new walkability index is thus equal to 1.92 * z(original walkability index) + 1 * z(transit access). The new walkability index had a positive association in the revised final model ($p<0.001$).

b. If the variable coefficient's p-value in the preliminary model was > 0.05, we dropped the variable from the final model.

   i. For example, in the preliminary binary logistic model of any senior recreational physical activity, the walkability index had a positive association but was not statistically significant ($p=0.53$). However, when added simultaneously with other correlated variables (transit access, rail access, and regional access), the sign on the walkability index became negative and the association strengthened ($p=0.1$). Following the variable selection methodology described on page 46, the walkability index was retained in the final model. As the only reason for the change in sign for the transit access coefficient appears to be due to multicollinearity, As the only reason for including the walkability index in the final model appears to be due to multicollinearity, we propose instead to remove the walkability index from the final model, which had a positive association but was non-significant ($p=0.53$) in the preliminary model.

3. For any of the four potentially multicollinear built environment variables, if the variable coefficient's p-value in the preliminary model was < 0.05 but the variable was dropped from the final model, we added the variable to the walk index with a sign that matches that found in the preliminary model. We weighted the components of the revised walkability index according to the relative t-values for each component variable in the preliminary model.

   a. For example, in the preliminary linear model of senior transportation walking minutes, the walkability index had a positive association ($p=0.016$) as did transit access

(p=0.019). However, when they are added simultaneously, they both continue to have a positive association, but the p-value for walkability dropped to 0.53 and for transit access dropped to 0.13. Following the variable selection methodology described on page 46, only the transit access variable was retained in the final model, and the p-value dropped again to 0.019 after removing the walkability index. As the only reason for dropping the walkability index from the final model appears to be due to multicollinearity, we propose instead to create a new walkability index variable that includes both the original walkability index and transit access variables. Based on the slightly higher t-value for the original walkability index (t=2.41) in the preliminary model as compared to transit access (t=2.35), we would assign a weight of 2.41/2.35=1.026 to the walkability index and a weight of 1.0 to transit access. The new walkability index is thus equal to 1.026 * z(original walkability index) + 1 * z(transit access). The new walkability index had a positive association in the revised final model (p=0.007).

Counterintuitive findings not apparently affected by multicollinearity

Additional model results were noted in this memo as being counterintuitive to our prior expectations (based on published literature and previous experience). For example, the walkability index is associated with greater likelihood of type 2 diabetes and greater likelihood of self-reported poor general health for both adults and seniors. These associations do not appear to be affected by multicollinearity or other data problems. Other factors common to all of the models may be affecting these results, such as unadjusted confounding, reverse causation, and self-selection. As there is little that can be done about these issues, short of long-term strategies such as new data collection, these results are unlikely to significantly change with continuing model revision. In many of these cases, the associations are relatively weak, and many have plausible explanations. For each of these findings, we will need to make decisions about whether to accept these results as-is, selectively drop counterintuitive variables from models, or drop models entirely.

Appendix D: County-level Census population and CHIS/CHTS sample size

	2010 Census		2009 CHIS*		2010 CHTS	
County	population	% of total	sample size	% of total	sample size	% of total
Alameda	1,510,271	4.4%			2,616	4.9%
Contra Costa	1,049,025	3.0%			2,137	4.0%
El Dorado	181,058	0.5%			308	0.6%
Fresno	930,450	2.7%			1,825	3.4%
Imperial	174,528	0.5%			10	0.0%
Kern	839,631	2.4%			2,254	4.2%
Kings	152,982	0.4%			470	0.9%
Los Angeles	9,818,605	28.4%			12,798	23.8%
Madera	150,865	0.4%			464	0.9%
Marin	252,409	0.7%			615	1.1%
Merced	255,793	0.7%			730	1.4%
Napa	136,484	0.4%			468	0.9%
Orange	3,010,232	8.7%			3,816	7.1%
Placer	348,432	1.0%			623	1.2%
Riverside	2,189,641	6.3%			2,847	5.3%
Sacramento	1,418,788	4.1%			1,268	2.4%
San Bernardino	2,035,210	5.9%			2,712	5.0%
San Diego	3,095,313	8.9%			2,777	5.2%
San Francisco	805,235	2.3%			1,360	2.5%
San Joaquin	685,306	2.0%			990	1.8%
San Mateo	718,451	2.1%			1,741	3.2%
Santa Clara	1,781,642	5.2%			3,607	6.7%
Solano	413,344	1.2%			952	1.8%
Sonoma	483,878	1.4%			1,243	2.3%
Stanislaus	514,453	1.5%			880	1.6%
Sutter	94,737	0.3%			250	0.5%
Tulare	442,179	1.3%			1,277	2.4%
Ventura	823,318	2.4%			2,010	3.7%
Yolo	200,849	0.6%			398	0.7%
Yuba	72,155	0.2%			287	0.5%
Total	34,585,264	100%			53,733	100%

* adult/senior sample only

Appendix E: Model results symbol key

Table 108: Definitions of all symbols used in results tables.

+++	Positive association, p < 0.001 (strong statistically significant)
++	Positive association, p < 0.05 (statistically significant)
+	Positive association, p > 0.05 (not statistically significant)
-	Negative association, p > 0.05 (not statistically significant)
--	Negative association, p < 0.05 (statistically significant)
---	Negative association, p < 0.001 (strong statistically significant)
	Variable was tested but not selected for inclusion in the model
NA	Variable was not tested for the model
--?	Superscript "?" indicates that a prior expectation was not established for this coefficient.
--[b]	Red text indicates counterintuitive association, and footnote provides potential explanation for the counterintuitive association.
--[b]	Counterintuitive association was caused by multicollinearity problem.
	Variable was dropped from the final model due to multicollinearity problem.
--[b]	Cumulative impact of two-part model was in expected direction.
any biking for transportation	The purple shading indicates a model that was revised to deal with multicollinearity problems. All other model results are identical between the two versions.
WI	Variable was added to the walkability index to address multicollinearity problem.
dropped	Variable was dropped from the final model to address multicollinearity problem.
--	Indicates a difference in sign direction across income cohorts.
+	
--	A variable that was added to the model after conducting variable selection uniquely for each income group.
--	A variable that was dropped from the model after conducting variable selection uniquely for each income group, with the symbol indicating the earlier result.
--	A variable that was dropped from the model after conducting variable selection uniquely for each income group, with the symbol indicating the earlier result. The variable was significantly associated (p<0.05) with the outcome in the preliminary model, suggesting that it was dropped due to multicollinearity.

Table 109: Expected direction for the association of each built environment variable with various outcomes.

	BE variable with presumed association for outcomes						
Outcomes	Walkability index	Transit access	Rail access %	Major road exposure	Regional accessibility	School distance	Park access
Active transport	+	+	+	-	+	-	+
Recreational PA	?	?	?	-	?	-	+
Auto transport.	-	-	-	+	+	+	?
Body weight & health	-	-	-	+	?	+	-

Appendix F: County-level comparison between UF and ACS data

Table 110: Bay Area.

	Alameda		Contra Costa		Marin		Napa		San Francisco		San Mateo		Santa Clara		Solano		Sonoma	
demographic field	**UF**	**ACS**	**UF**	**ACS**	**UF**	**ACS**	**UF**	**ACS**	**UF**	**ACS**	**UF**	**ACS**	**UF**	**ACS**	**UF**	**ACS**	**UF**	**ACS**
pop_male	728,771	725,077	508,489	500,249	130,443	122,533	63,822	66,748	396,113	399,683	347,954	345,989	876,467	873,844	195,332	205,213	233,668	233,187
pop_female	739,496	752,903	529,914	524,560	110,390	126,068	67,830	67,303	385,160	389,489	359,345	358,338	871,520	865,552	204,754	204,829	240,176	240,860
pop_age65_up	162,857	159,693	141,981	123,512	26,214	39,383	21,325	20,021	105,207	108,173	94,891	93,047	211,138	186,574	50,735	44,310	70,174	63,395
pop_age20_64	947,898	938,335	631,142	616,577	165,791	153,717	78,562	79,218	561,749	558,015	445,392	439,307	1,089,925	1,087,527	242,875	250,501	294,743	290,563
pop_hs_not_comp	213,201	140,667	126,911	78,876	7,514	14,974	23,532	16,156	113,999	86,953	83,663	56,380	246,932	158,989	54,803	37,895	64,855	44,513
pop_hs_diploma	299,327	201,888	207,635	134,182	47,089	23,401	26,497	18,652	115,991	87,640	128,798	87,874	293,097	191,342	96,579	65,311	100,961	67,873
pop_some_college	371,522	251,528	318,974	207,824	68,571	45,412	42,034	28,788	161,645	122,257	185,772	131,283	435,054	285,681	150,231	99,177	160,784	108,657
pop_college_degree	349,201	239,192	248,419	167,077	71,565	57,501	26,516	17,963	239,725	191,549	189,438	132,940	441,425	298,284	68,182	44,521	95,333	65,842
pop_graduate_degree	235,016	161,685	136,465	93,203	46,094	41,389	13,073	9,235	149,914	120,164	119,628	83,897	331,481	227,554	30,291	19,440	51,912	35,673
pop_hispanic	314,071	339,880	243,929	255,569	9,927	39,069	37,867	44,015	120,743	121,774	176,185	182,503	444,538	479,209	92,861	99,352	107,286	120,425
pop_white	513,766	514,465	517,073	501,010	217,810	183,834	80,835	76,988	333,177	337,451	301,241	303,626	642,110	626,901	171,274	168,641	324,011	320,002
pop_black	180,405	184,123	84,000	93,607	107	6,621	1,703	2,441	45,340	46,781	18,256	18,763	38,903	42,331	51,160	58,743	5,974	6,768
pop_asian	380,055	390,512	141,907	148,893	4,989	13,577	7,586	8,986	249,285	265,700	172,604	175,936	557,713	565,465	57,909	59,027	16,689	17,777
pop_american_indian	4,342	4,189	3,090	2,984	167	531	435	544	1,804	1,828	1,236	1,125	4,186	4,042	1,913	1,864	3,401	3,584
pop_hawaiian_islander	12,889	11,931	4,316	4,382	7,156	436	303	313	2,738	3,128	10,103	9,884	6,743	6,252	3,243	3,243	1,718	1,434
pop_other_ethnicity	62,740	65,050	44,089	42,694	678	8,345	2,923	3,224	28,187	28,573	27,673	26,634	53,794	57,432	21,728	22,489	14,766	13,857
hh_avg_size	2.73	2.78	2.77	2.78	2.72	2.42	2.61	2.73	2.28	2.35	2.74	2.75	2.87	2.91	2.81	2.95	2.52	2.58
hh_avg_inc	92,529	91,054	105,753	103,224	102,739	128,544	93,589	92,400	101,269	102,261	121,530	118,774	101,424	113,158	22,041	82,464	33,288	82,330
hh_avg_vehicles	2.00	1.52	2.08	1.69	1.88	1.70	2.11	1.72	1.28	1.04	2.11	1.70	2.20	1.73	2.25	1.87	2.12	1.77
hh_owner_occ	303,335	293,277	268,926	255,807	34,302	65,720	32,039	31,996	113,179	126,028	158,354	156,149	377,150	353,399	97,866	91,447	118,741	114,787
hh_rental_occ	234,741	238,749	105,659	112,280	54,336	37,007	18,439	17,183	228,786	209,928	99,403	99,609	231,202	243,348	44,668	47,564	69,505	69,246

Table 111: SACOG.

	El Dorado		Placer		Sacramento		Sutter		Yolo		Yuba	
demographic field	**UF**	**ACS**	**UF**	**ACS**	**UF**	**ACS**	**UF**	**ACS**	**UF**	**ACS**	**UF**	**ACS**
pop_male	74,507	89,764	165,789	164,341	683,852	684,547	46,789	46,498	98,390	95,020	36,130	35,887
pop_female	76,866	89,289	176,380	172,136	728,199	712,371	47,800	46,922	103,488	99,624	36,351	35,273
pop_age65_up	25,062	24,499	55,983	49,765	169,820	152,766	12,859	11,399	21,079	18,506	8,117	6,928
pop_age20_64	88,680	108,015	196,167	194,894	848,890	841,143	53,722	53,296	127,127	118,358	42,199	41,183
pop_hs_not_comp	8,843	9,009	21,905	15,909	217,765	133,404	20,436	12,638	29,789	17,493	17,394	9,880
pop_hs_diploma	34,706	28,918	69,553	47,371	317,819	200,786	23,430	14,474	38,514	22,261	19,492	11,501
pop_some_college	58,625	47,474	131,029	86,269	491,486	311,188	33,099	20,352	51,067	29,421	26,625	16,015
pop_college_degree	32,702	25,947	81,739	52,571	264,496	169,717	12,376	7,687	43,249	22,824	6,725	3,999
pop_graduate_degree	16,498	12,514	37,944	24,739	120,486	79,086	5,248	3,249	39,260	19,569	2,245	1,320
pop_hispanic	13,983	21,961	40,785	44,710	297,961	306,095	26,155	27,241	58,055	60,970	17,666	18,049
pop_white	125,633	145,170	262,806	265,294	695,883	686,617	49,282	47,737	102,873	100,340	43,748	42,355
pop_black	1,038	1,306	4,402	4,427	135,588	139,938	1,645	1,713	4,722	4,753	1,985	2,116
pop_asian	4,647	6,481	20,299	19,963	195,970	198,600	12,988	13,438	26,928	25,650	4,278	4,710
pop_american_indian	1,302	1,557	2,027	2,080	7,842	7,870	924	925	1,054	1,099	1,266	1,260
pop_hawaiian_islander	223	261	697	697	13,518	13,099	244	256	792	817	233	270
pop_other_ethnicity	4,547	5,278	11,153	11,261	65,291	65,517	3,350	3,367	7,454	7,348	3,304	3,320
hh_avg_size	2.65	2.62	2.64	2.61	2.74	2.74	3.00	2.98	2.85	2.83	2.94	3.00
hh_avg_inc	91,812	88,811	93,410	92,100	71,896	72,374	65,358	64,910	75,882	75,697	60,438	58,506
hh_avg_vehicles	1.62	1.73	1.54	1.76	1.54	1.65	1.01	1.80	1.04	1.73	1.31	1.84
hh_owner_occ	44,984	52,353	94,997	94,206	296,697	302,730	19,501	19,380	38,337	37,247	15,119	14,214
hh_rental_occ	12,224	16,041	34,699	34,947	218,441	206,363	12,077	11,993	32,493	31,545	9,527	9,536

Table 112: San Joaquin Valley.

	Fresno		Kern		Kings		Madera		Merced		San Joaquin		Stanislaus		Tulare	
demographic field	**UF**	**ACS**	**UF**	**ACS**	**UF**	**ACS**	**UF**	**ACS**	**UF**	**ACS**	**UF**	**ACS**	**UF**	**ACS**	**UF**	**ACS**
pop_male	472,272	455,119	387,341	420,962	76,126	85,889	73,411	71,232	128,059	125,185	307,406	335,907	242,714	252,215	216,079	215,004
pop_female	486,901	455,608	387,984	394,731	72,797	65,233	71,860	76,506	128,479	123,617	314,630	337,706	248,802	257,467	215,625	214,400
pop_age65_up	96,149	89,917	73,704	72,402	13,668	11,647	16,860	16,489	26,061	22,364	71,549	68,012	51,889	52,736	41,925	40,073
pop_age20_64	533,581	514,222	437,160	466,998	84,954	93,126	80,306	84,120	142,056	136,990	354,486	382,776	283,836	292,080	236,890	234,300
pop_hs_not_comp	269,753	142,945	238,262	137,473	43,834	27,473	52,488	28,938	85,673	46,127	158,642	94,076	128,692	76,127	151,675	79,494
pop_hs_diploma	218,898	123,477	202,486	127,759	40,618	25,624	35,922	22,979	65,253	35,803	171,701	109,735	136,083	85,277	105,291	59,550
pop_some_college	299,211	159,649	222,818	141,035	46,847	27,365	39,591	26,082	73,769	40,409	192,986	128,962	151,013	96,575	123,079	72,308
pop_college_degree	117,682	71,063	75,251	46,822	13,120	7,788	12,106	8,367	21,529	11,835	69,271	49,328	52,044	34,281	35,134	21,317
pop_graduate_degree	53,630	33,189	36,508	23,079	4,505	2,974	5,164	3,838	10,314	5,741	29,437	21,241	23,685	15,812	16,525	10,144
pop_hispanic	471,117	468,314	386,342	413,028	83,064	77,866	80,285	80,972	139,622	140,261	244,179	266,338	205,872	215,673	261,920	268,065
pop_white	317,258	304,691	298,342	323,783	51,470	53,879	55,090	57,372	84,668	81,438	234,091	245,898	230,446	240,451	140,690	143,935
pop_black	51,431	45,011	36,438	45,376	5,402	10,314	2,985	5,009	8,020	8,779	40,458	48,540	12,873	13,065	5,065	5,497
pop_asian	92,016	86,859	32,126	33,100	4,853	5,339	2,529	2,533	17,597	18,180	76,813	94,547	22,321	24,712	14,607	14,204
pop_american_indian	5,936	5,985	5,264	5,893	868	1,297	1,552	1,790	1,089	1,120	3,143	3,179	2,943	2,870	2,494	3,323
pop_hawaiian_islander	1,024	1,066	898	995	174	228	84	107	383	476	2,766	3,248	2,948	3,016	360	370
pop_other_ethnicity	20,392	18,954	15,916	17,439	3,092	4,059	2,747	3,054	5,159	5,137	20,585	23,529	14,115	14,713	6,568	6,785
hh_avg_size	3.34	3.20	3.27	3.29	3.47	3.72	3.48	3.51	3.41	3.41	3.27	1,705.35	3.13	3.11	3.43	3.39
hh_avg_inc	63,409	62,655	63,263	63,003	59,690	60,784	61,007	62,094	60,118	59,066	69,297	70,726	65,595	66,366	58,247	58,593
hh_avg_vehicles	1.32	1.65	1.38	1.78	1.52	1.87	1.46	1.95	1.24	1.83	1.39	1.88	1.75	1.86	1.27	1.80
hh_owner_occ	159,788	156,356	147,239	152,284	23,534	22,738	26,950	26,535	44,370	40,911	114,996	131,357	97,959	101,826	76,167	75,081
hh_rental_occ	127,597	128,087	89,815	95,773	19,421	17,868	14,759	15,554	30,949	32,068	77,051	81,548	59,122	62,015	49,823	51,583

Table 113: SANDAG.

	San Diego	
demographic field	**UF**	**ACS**
pop_male	1,495,903	1,516,787
pop_female	1,529,036	1,505,681
pop_age65_up	374,537	337,724
pop_age20_64	1,871,679	1,871,431
pop_hs_not_comp	467,746	286,835
pop_hs_diploma	604,131	384,822
pop_some_college	953,971	611,814
pop_college_degree	628,034	415,538
pop_graduate_degree	371,058	247,516
pop_hispanic	988,546	991,348
pop_white	1,546,995	1,500,047
pop_black	142,598	146,600
pop_asian	333,871	328,058
pop_american_indian	0	14,098
pop_hawaiian_islander	13,697	13,504
pop_other_ethnicity	104,238	101,658
hh_avg_size	2.93	2.85
hh_avg_inc	87,477	83,934
hh_avg_vehicles	2.22	1.74
hh_owner_occ	585,300	593,945
hh_rental_occ	486,429	467,844

Table 114: SCAG.

	Imperial		Los Angeles		Orange		Riverside		San Bernardino		Ventura	
demographic field	**UF**	**ACS**	**UF**	**ACS**	**UF**	**ACS**	**UF**	**ACS**	**UF**	**ACS**	**UF**	**ACS**
pop_male	80,795	86,759	4,752,832	4,811,964	1,494,648	1,467,799	1,075,352	1,050,921	963,228	998,752	391,192	402,061
pop_female	85,361	81,293	4,912,691	4,946,292	1,528,652	1,497,726	1,105,399	1,058,506	990,060	1,006,535	398,303	407,019
pop_age65_up	18,333	17,293	1,045,500	1,026,041	348,039	331,871	280,626	244,902	174,749	171,357	92,313	91,302
pop_age20_64	90,893	95,216	5,947,014	5,973,554	1,844,137	1,806,430	1,225,508	1,184,554	1,134,489	1,165,433	470,452	480,554
pop_hs_not_comp	62,237	37,487	2,508,298	1,511,061	544,092	321,534	472,413	267,304	450,045	265,306	149,207	91,553
pop_hs_diploma	36,537	22,837	2,090,738	1,332,186	561,388	358,035	568,246	336,404	526,076	316,870	157,487	102,318
pop_some_college	45,710	26,996	2,454,361	1,608,268	855,233	556,542	702,659	417,454	627,506	382,175	248,868	163,695
pop_college_degree	14,492	8,142	1,722,519	1,192,974	698,321	454,402	287,506	171,369	232,419	142,586	149,636	101,315
pop_graduate_degree	7,181	3,962	889,609	623,632	364,267	239,481	149,928	91,846	117,251	74,476	84,301	57,858
pop_hispanic	135,666	140,261	4,639,688	4,687,872	1,013,467	1,013,000	967,335	995,254	966,509	1,001,132	316,921	331,568
pop_white	23,389	23,882	2,675,501	2,728,298	1,334,764	1,328,569	894,357	869,016	643,406	677,407	385,919	400,870
pop_black	1,954	5,114	788,577	815,070	43,460	44,017	127,351	130,823	160,053	170,700	12,454	13,082
pop_asian	2,070	2,201	1,305,329	1,325,509	538,874	532,639	123,387	125,921	122,995	123,978	51,329	54,099
pop_american_indian	0	1,641	0	18,886	0	6,216	0	10,931	0	8,518	0	2,389
pop_hawaiian_islander	83	87	22,209	22,464	8,308	8,357	5,780	5,849	5,592	5,845	1,274	1,353
pop_other_ethnicity	1,235	1,286	215,530	220,281	78,229	77,717	51,628	51,791	45,628	47,411	19,328	19,960
hh_avg_size	3.49	3.55	2.98	3.03	3.11	3.01	3.22	3.16	3.27	3.36	3.01	3.06
hh_avg_inc	57,578	54,947	79,378	79,655	100,841	99,712	77,293	75,078	71,696	70,297	96,860	96,331
hh_avg_vehicles	1.71	1.73	1.58	1.57	1.99	1.71	2.04	1.83	1.80	1.78	1.84	1.84
hh_owner_occ	27,954	26,779	1,549,145	1,552,091	592,943	599,032	469,436	467,058	387,611	388,167	174,136	175,452
hh_rental_occ	10,522	20,525	1,549,145	1,665,798	181,323	385,471	107,265	199,820	108,961	207,958	43,212	88,853

Appendix G: Calibration Multipliers Documentation

Urban Footprint (UF) public health models were applied to grid-level UF predictors for every grid cell within the 30-county study area. The grid-level outcomes were then aggregated and compared to weighted outcomes derived from the 2009 CHIS and 2010-2012 CHTS surveys for the same 30 counties. In general, the UF models predicted lower outcome values than the CHIS/CHTS survey data, particularly for less common outcomes (e.g. adult type 2 diabetes). However, when the UF models were applied to individual-level 2005 CHIS predictors and aggregated for the 30-county study area, the predicted outcomes were generally similar to the weighted outcomes derived from the 2009 CHIS survey. Two potential reasons for this discrepancy have been identified:

1. The UF models were fit with individual-level CHIS/CHTS data, whereas the UF application data is composed of group-level (e.g. groups of 1-1000 people) data derived from Census block group (or similar) aggregate estimates.
    a. This results in very different predictor distributions; for example:
        i. In the CHIS data, gender is 0 or 1 for each individual
        ii. In the UF data, gender is a continuous decimal value ranging from 0-1 for each grid cell, with a normal distribution with a tall peak. About 90% of the values range from 0.4-0.6.
2. Nearly every model uses either a logit or (binary logistic regression) log link (poisson), is a linear model with a log-transformed outcome (e.g. minutes of physical activity), or uses log-transformed predictor values.
    a. These log transformations further exacerbate differences between the CHIS and UF data distributions.

Table 1 provides and extreme example to illustrate this problem using the type 2 diabetes model for adults, which is a binary logistic regression model. We populated hypothetical predictor values for 5 individuals (which is how CHIS represents the data) and then calculated average predictor values for the group of 5 individuals (which is how UF represents the data). When we calculated the individual level predictions and then averaged those predictions over 5 people, we calculated a likelihood of 7.0%. But when we calculated the group-level prediction based on the average predictor values, we calculated a much lower likelihood of 1.2%!

Table 115. Example of averaging individual-level predictions versus predicting based on average characteristics.

	Estimate	Std. Error	z value	Pr(>|z|)		person1	person2	person3	person4	person5	person1-5
(Intercept)	-17.9867	0.610798	-29.4479	1.34E-190		1	1	1	1	1	1
mvpa_METS_log	-0.03559	0.011712	-3.03858	0.002377		0	0	5.824524	7.421178	10.34663	4.718467
BMI_log	3.559946	0.149824	23.76083	8.49E-125		3.912	3.388	3.249	3.125	2.741	3.038333
gender2	-0.40701	0.061648	-6.60214	4.05E-11		0	0	1	1	1	0.6
age	0.079108	0.004051	19.53039	6.06E-85		64	57	49	39	18	45.4
racehisp1	-0.70721	0.081855	-8.63977	5.63E-18		0	0	1	1	1	0.6
racehisp2	-0.31405	0.134059	-2.34261	0.019149		0	0	0	0	0	0
racehisp4	0.086058	0.107556	0.800124	0.423639		1	0	0	0	0	0.2
racehisp97	-0.19751	0.173232	-1.14017	0.254217		0	0	0	0	0	0
emply2	0.268858	0.066109	4.06689	4.76E-05		1	1	0	0	0	0.4
hhsize	0.074674	0.027816	2.684543	0.007263		5	4	3	2	1	3
incom2	-0.13173	0.129515	-1.01711	0.309099		1	0	0	0	0	0.2
incom3	-0.26325	0.148781	-1.76938	0.07683		0	0	0	0	0	0
incom4	-0.21437	0.142521	-1.50411	0.132553		0	1	0	0	0	0.2
incom5	-0.29711	0.153294	-1.93819	0.0526		0	0	1	0	0	0.2
incom6	-0.30065	0.150396	-1.99903	0.045605		0	0	0	1	0	0.2
incom7	-0.72542	0.163806	-4.42854	9.49E-06		0	0	0	0	1	0.2
child_any1	-0.19135	0.094627	-2.02214	0.043162		0	0	0	1	1	0.4
disabled1	0.177018	0.094418	1.874845	0.060814		1	0	0	0	0	0.2
disabled2	0.436867	0.074849	5.836636	5.33E-09		1	0	0	0	0	0.2
walk_index	0.021266	0.008835	2.406989	0.016085		12.23	2.067	-0.1807	-2.249	-14.74	-0.57454
regional_access	-0.03538	0.01677	-2.10983	0.034873		-5.165	-1.367	-0.2465	0.5438	2.475	-0.75194
park_access	-0.04742	0.018169	-2.60998	0.009055		-3.863	-0.01125	0.9506	1.973	4.096	0.62907
Predicted likelihood of type 2 diabetes for each of 5 individuals:						**20.9%**	**11.9%**	**1.8%**	**0.3%**	**0.0%**	
			Predicted likelihood of type 2 diabetes, averaged over 5 people:							**7.0%**	
		Predicted likelihood of type 2 diabetes for group of 5 people with average characteristics:									**1.2%**

After consulting with Dr. Sanjay Basu at Stanford University, it was decided address this discrepancy by calculating calibration multipliers that would help to inflate the UF predictions to the expected control totals. To do so, we applied every model to grid-level UF predictors for every grid cell within the 30-county study area, then aggregated the grid-level predictions to an overall predicted outcome value for each model. We then calculated weighted outcome values (control totals) from the 2009 CHIS and 2010-2012 CHTS surveys for the same model outcomes and for the same 30-county study area. Finally, we calculated the calibration multipliers by dividing the CHIS/CHTS control total by the UF predicted value for each model. The full list of calibration multipliers, along with the CHTS/CHIS control totals and the UF unadjusted predictions for the 30-county study area are shown in Table 2. The calibration multipliers are applied in UF to adjust the UF model prediction to match the CHIS and CHTS control totals.

Table 116. Model calibration multipliers calculated from CHTS/CHIS validation control totals and UF unadjusted model output for 30 county study area.

source	age group	Outcome	CHTS/CHIS validation	UF unadjusted	Calibration multiplier
CHTS	adults	transportation walking (minutes)	5.69	3.46	1.64
CHTS	adults	transportation biking (minutes)	1.14	0.73	1.58
CHTS	adults	auto transport (minutes)	71.87	78.92	0.91
CHTS	adults	recreational PA (minutes)	17.42	16.01	1.09
CHTS	seniors	transportation walking (minutes)	4.03	2.80	1.44
CHTS	seniors	auto transport (minutes)	55.81	62.18	0.90
CHTS	seniors	recreational PA (minutes)	16.38	15.79	1.04
CHTS	teens	transportation walking (minutes)	8.57	7.68	1.12
CHTS	teens	auto transport (minutes)	45.75	47.04	0.97
CHTS	teens	recreational PA (minutes)	30.98	30.35	1.02
CHTS	children	transportation walking (minutes)	6.09	4.19	1.45
CHTS	children	auto transport (minutes)	48.53	49.72	0.98
CHTS	children	recreational PA (minutes)	32.13	28.33	1.13
CHIS	adults	transportation walking (minutes)	61.36	54.66	1.12
CHIS	adults	leisure walking (minutes)	81.07	71.22	1.14
CHIS	adults	moderate PA (minutes)	117.58	106.18	1.11
CHIS	adults	vigorous PA (minutes)	68.72	60.34	1.14
CHIS	adults	BMI	26.86	26.45	1.02
CHIS	adults	overweight or obese	0.58	0.54	1.08
CHIS	adults	obese	0.23	0.19	1.18
CHIS	adults	high blood pressure	0.21	0.17	1.28
CHIS	adults	heart disease	0.04	0.02	2.33
CHIS	adults	type 2 diabetes	0.05	0.02	2.35
CHIS	adults	poor general health	0.18	0.13	1.41

CHIS	seniors	transportation walking (minutes)	44.57	42.82	1.04
CHIS	seniors	leisure walking (minutes)	84.60	80.95	1.05
CHIS	seniors	moderate PA (minutes)	127.31	126.17	1.01
CHIS	seniors	vigorous PA (minutes)	32.96	26.90	1.23
CHIS	seniors	BMI	26.43	26.42	1.00
CHIS	seniors	overweight or obese	0.56	0.58	0.98
CHIS	seniors	obese	0.19	0.18	1.06
CHIS	seniors	high blood pressure	0.58	0.58	1.00
CHIS	seniors	heart disease	0.20	0.16	1.25
CHIS	seniors	type 2 diabetes	0.17	0.13	1.29
CHIS	seniors	poor general health	0.24	0.17	1.39
CHIS	teens	days/week >60 minutes PA	3.47	3.45	1.01
CHIS	teens	likelihood of walking to school	0.49	0.47	1.04
CHIS	teens	overweight or obese	0.33	0.28	1.16
CHIS	teens	obese	0.16	0.11	1.48
CHIS	teens	poor general health	0.08	0.08	0.96
CHIS	children	days/week >60 minutes PA	4.33	3.37	1.29
CHIS	children	likelihood of walking to school	0.38	0.38	0.99
CHIS	children	poor general health	0.05	0.04	1.34