Multimodal Transportation Choices and Health

Exploratory Analysis Using Data Fusion Techniques

Miguel Lugo and Sivaramakrishnan Srinivasan

This study demonstrates the feasibility of fusing large-scale travel and health surveys and uses the new comprehensive data set generated to model the relationship between health and multimodal (walking, biking, transit, and vehicle usage) long-term (weekly, monthly, and yearly) travel choices. Two measures of health, the body-mass index (BMI) and a self-assessed physical health score (SAPHS), were fused from a health survey onto a travel survey at the disaggregate (individual) level. The probabilistic record linkage software, Link Plus, was used for the data fusion purposes. The methodology was validated by using the eating and health module (EH) of the American Time Use Surveys (ATUS). Subsequently, the algorithm was used to match the health information from the ATUS to the National Household Travel Surveys (NHTS) of 2008 to 2009, and the resulting master data set was used to develop models for multimodal travel choices and health. The statistical analysis indicates that although increasing walking and transit use were associated with better health (relative to nonusers of the mode), those with the highest levels of walking and transit use were also found to be in poor health relative to moderate users of the mode. Similarly, those at the two ends of the vehicle miles traveled spectrum (first and fourth quartiles) had higher BMI compared with those in the middle of the spectrum. There were no statistically significant effects of weekly bike trips on health measures. Overall, this study is envisioned as a proof-of-concept of how data fusion techniques may be used to integrate multiple data sets to facilitate a comprehensive study of multimodal travel choices and health.

The U.S. transportation system is a complex multimodal array of highways, roads, transit routes, sidewalks, and bike paths. Nonetheless, the United States remains a car-centric nation: as many as 69% of short trips (which could be undertaken by walking or biking) are made by private motorized vehicles (*1*). Inactive lifestyles, often characterized by the predominant use of a private vehicle for most trips and long commutes, have been known to show health implications (*2*). A review by McCormack and Virk on time spent driving private motor vehicles finds longer times likely contributed to the obesity epidemic. Conversely, higher transit use has been linked with more physical activity (*3*). When it comes to nonmotorized modes, there is clear evidence of the benefits of physical activity through walking (*4*), and cycling (*5*). Properly implemented, the inclusion of health outcomes in transportation planning promotes environmental justice principles and practices. Thus, understanding the relationship between a person's well-being and available transportation infrastructure can further support and guide projects that impact the natural and human environments of all residents, including low-income and minority communities.

Although the overall volume of studies on time use, transportation choices, and health is extensive [see Lugo for an extensive discussion about this literature (*6*)], these studies often focus on a single mode or a specific aspect of time use. Studies also often rely on short-term or one-day travel information [e.g., Wojan and Hamrick, and Merom et al. (*8, 4*)]. Given the significant day-to-day variabilities in travel patterns, the travel pattern of a single day may not be representative enough to be a predictor of long-term health patterns. Travel surveys do collect multimodal and long-term travel information, and technological advances [see Wolf for a review (*9*)] have enhanced the ease of collecting multiday travel information. However, practically none of the travel surveys collect data on health while health surveys are generally limited in the travel data collected.

Therefore, developing models that can examine multimodal travel behavior and health requires new surveys that collect both travel and health data. However, such new surveys will be expensive. As an alternate, a comprehensive data set may also be synthesized by fusing data from existing travel and health surveys. Broadly, data fusion methods link (on an individual level) records from travel and health surveys so that the resulting data set has both travel and health information for a large sample of individuals.

The transportation literature offers examples of data fusion or record linkage. Fused data include smart card transit data stops and personal trip surveys [Kusakabe and Asakura (*10*)], police and hospital road crash records, Amorim et al. (*11*) and Rosman (*12*), work fatigue and crashes, Williamson and Boufous (*13*), and state and federal motor carrier safety databases (*14*). Pawlak et al. (*15*) used separate information and communication technology (ICT) data sets on digital lifestyle (ICT use) and physical mobility to match records as an alternative when suitable data could not be obtained in a single data set. Similarly, Kressner and Garrow (*16*) also found third-party data from targeted marketing to be useful in incorporating lifestyle variables into transportation survey.

This study demonstrates the feasibility of such a data fusion in the context of large-scale travel and health surveys, subsequently using the new comprehensive data set generated to model the relationship between health and multimodal (walking, biking, transit, and vehicle usage) long-term (weekly, monthly, and yearly) travel choices. Two measures of health are fused from a health survey onto a travel survey at the disaggregate level.

Engineering School of Sustainable Infrastructure and Engineering, University of Florida, 365 Weil Hall, P.O. Box 116580, Gainesville, FL 32611. Corresponding author: S. Srinivasan, siva@ce.ufl.edu.

Transportation Research Record: Journal of the Transportation Research Board, No. 2598, Transportation Research Board, Washington, D.C., 2016, pp. 37–45.
DOI: 10.3141/2598-05

The rest of this paper comprises three major sections. The next section of the paper presents an overview of the data fusion methodology and the empirical validation. Subsequently, the methodology is applied to fuse a national-level travel survey with a health survey, and the resulting data are used to develop models for the impacts of multimodal travel choices on health. The paper ends with a summary, conclusions, and directions for future research.

DATA FUSION: METHODOLOGY AND VALIDATION

The data fusion approach involves comparing the records of a receiver data set with records of a donor data set to identify the record from the donor that best matches each record in the receiver on a set of predefined socioeconomic attributes. Desired attributes (that are unknown in the receiver record) of the matched donor record are then set as attributes of the receiver record. The unknown attributes in the receiver that are matched from the donor in this study are the health measures.

This study uses Link Plus, a probabilistic record linkage software program developed at the U.S. Centers for Disease Control (CDC), based on the theoretical framework of Fellegi and Sunter (*17*) and Dempster et al. (*18*). The process of linking records spans from the assumption that there is a potential matching pair within the two sources. Record-pair comparisons will lead to one of three outcomes: (*a*) a match, (*b*) a possible match, and (*c*) a nonmatch. The degree of separation between the two data sets can be an indication of the difficulty level of the linkage and amount of Types 1 and 2 errors. The standard algorithm requires a method that minimizes the probability of possible matches, as opposed to true matches and nonmatches. This is done by estimating the ratio between the conditioned probabilities of observed identifying fields, given that the record pair is a true match and a true nonmatch. In many instances, the expectation–maximization algorithm (EM), as described by Jaro (*19*), is used to estimate the weights or threshold values of this association.

Disaggregate level linking is performed on the basis of several variables. The software allows these variables of interest to be specified as either blocking or matching variables. In this study, gender and the census region of the household (northeast, midwest, south, and west) were used as simultaneous blocking variables. This means that a matching record for a male from the northeast region (in the receiver data set) will necessarily be obtained from males from the northeast in the donor data set. As such, the matching on these attributes is deterministic (the software will report that a matching record could not be found if such a deterministic match is not possible). Variables that are not blocked are matched probabilistically. Three matching methods are employed: (*a*) exact matching (binary all-or-nothing comparison), (*b*) generic string (comparison of characters and the needed number of operations needed to transform one string into the other), and value-specific (sets weights matching values based on the frequencies of values in the file being compared). See Lugo's dissertation (6) for further details about the matching procedure.

Data from the 2006, 2007, and 2008 ATUS and the corresponding eating and health (EH) modules are used first for validating the data fusion methodology. The ATUS collected detailed socioeconomic, demographic, and one-day activity-travel information for a large sample of persons (one adult ≥ 15 years old per household surveyed). The EH module of the ATUS collected additional eating, meal preparation, and health information for a random selection of household members participating in the ATUS. There are two measures of health available from the well-being module of the ATUS. These are the Body Mass Index (BMI) and a Self-Assessed Physical Health Score (SAPHS). BMI is calculated from the self-reported weight (How much do you weigh without shoes?) and height (How tall are you without shoes?). The SAPHS was obtained as the response to the question: In general, would you say that your physical health was excellent (4), very good (3), good (2), fair (1), or poor (0)? Overall, both time-use and health data are available for about 36,000 individuals.

The health of a person can be described using a multitude of attributes, including fitness, diseases, and quality of life (see Lugo (*6*) for further discussion). Each of these may be assessed subjectively and objectively. BMI is largely an objective indicator of fitness even though it has been correlated to several diseases. Further, as measured, BMI uses excess weight as an indicator of excess fat. Even though other methods exist for calculating excess fat, the BMI measure is perhaps easier to determine in large-scale surveys. The SAPHS used in this survey is a subjective (self-assessed) measure of overall well-being and could be related to the general quality of life. Unlike BMI, which describes only excess weight, the SAPHS describes health more broadly. The objectivity involved in BMI calculation (standardized procedure) could lead to consistency in its assessment across people; something that could be compromised in subjective measures such as the SAPHS. In summary, while both BMI and SAPHS have been used in several studies in the past, using both measures (and possibly more) would provide a better picture of the respondents' health.

To perform validation of the data fusion method, it is necessary to know the health measures for the receiver sample as well so that these may be compared with the imputed measures from the matched donor sample. Therefore, a random subset of 10,000 records were extracted and defined as the receiver sample, and the remaining 26,000 records were set as the donor sample. Both the BMI and the SAPHS measures were matched from the donor to the receiver. As already indicated, gender and census region were used as blocking variables, and education, metropolitan location, housing tenure, student status, age, race, employment status, income, and household size were used as other matching variables (see Lugo (*6*) for further details).

The donor data set was matched to the receiver data set with Link Plus software. Five runs of the matching were performed by resorting the donor data set before the run. As a consequence, the same record from the donor may not be matched to a receiver in all the runs (this is especially true for the more typical households, which may have multiple possible matches in the donor).

Figure 1 presents a plot of the observed BMI against the BMI of the matched record from the donor for each of the five runs. Figure 2 presents the distribution of the observed and matched BMI across the 10,000 records from the receiver sample. Overall, Figure 2 indicates that the observed aggregate distribution is replicated fairly well. However, Figure 1 indicates that the possibility of extreme mismatches (points in the bottom right or top left) are more likely from a single run. Therefore, the BMI values are averaged across the five runs and assigned as the matched BMI of the record in the donor. Figure 2 shows that the distribution of the average BMI is still consistent with the observed distribution.

Figure 3 presents a plot of the observed SAPHS against the SAPHS of the matched record from the donor for each of the five runs. Figure 4 presents the distribution of the observed and matched SAPHS across the 10,000 records from the receiver sample. Figure 4 indicates that the observed aggregate SAPHS distribution is fairly well replicated by the distribution of the matched SAPHS. As in the case of BMI, the average (rounded to the nearest integer) SAPHS is used across the five runs as the assigned value to the receiver data set (other ways of aggregating across the five runs are identified as an area of future study).

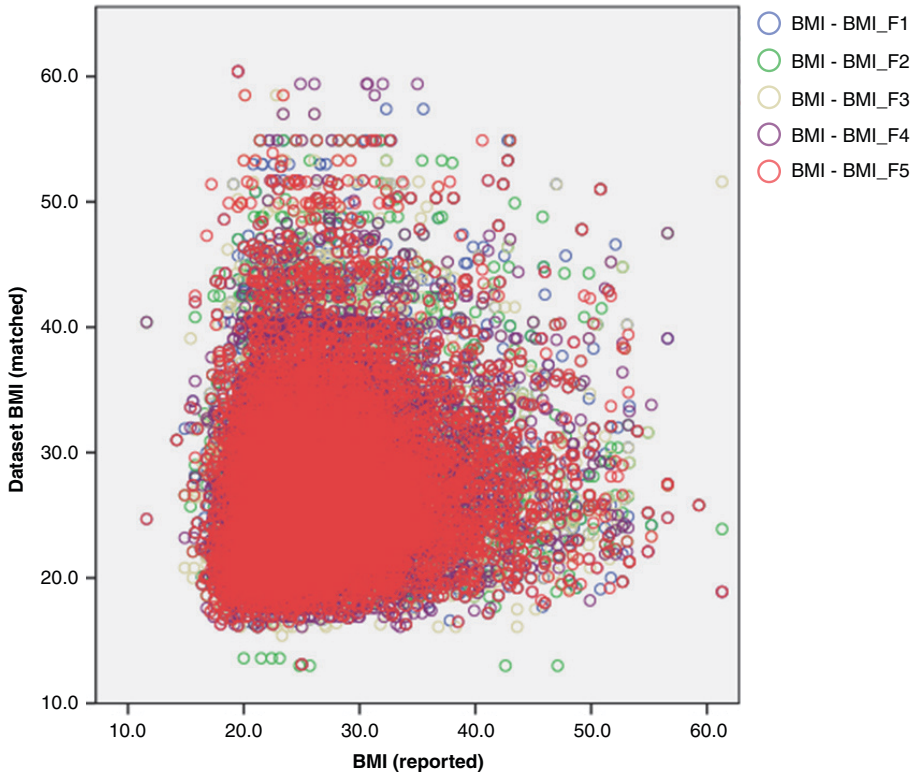

FIGURE 1 Validation of data fusion: matched versus reported BMI values.

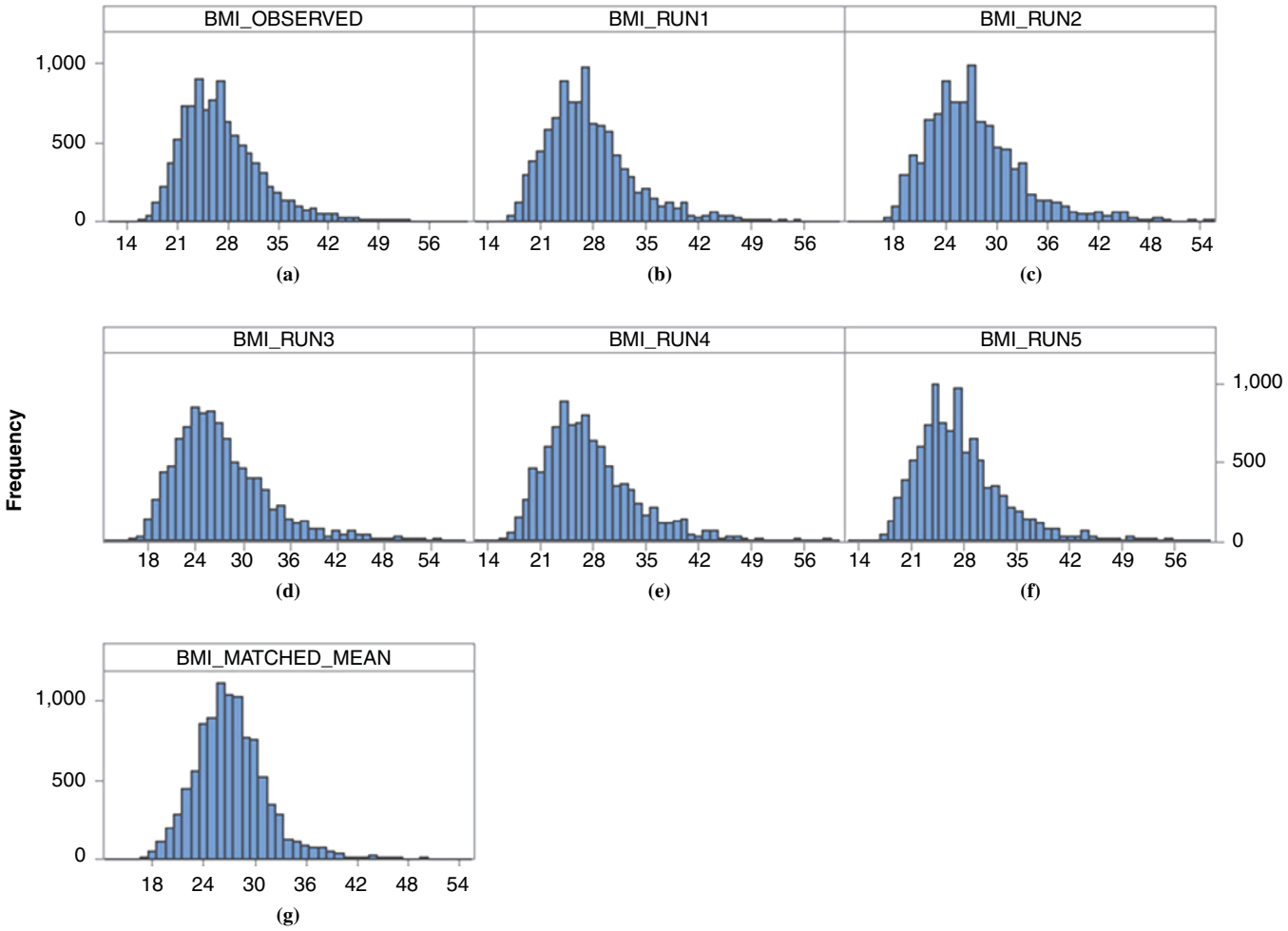

FIGURE 2 Frequency distribution of observed BMI, matched BMIs for five runs, and five-sample mean.

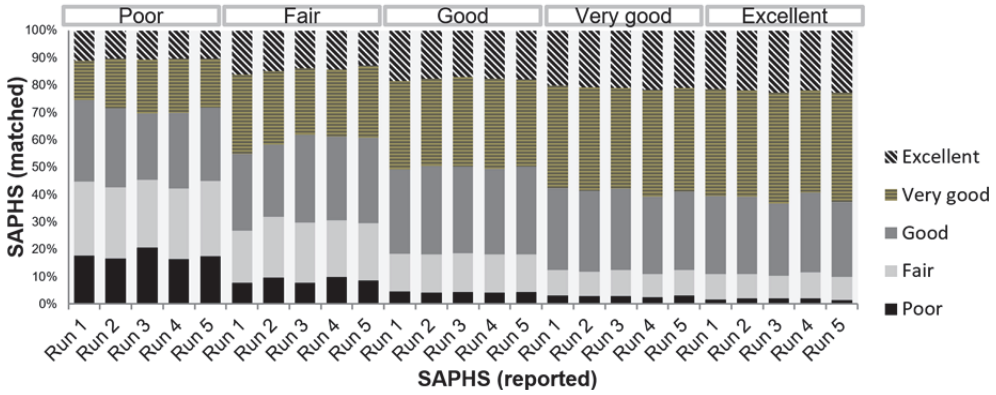

FIGURE 3 Validation of data fusion: count of matched SAPHS values by reported value.

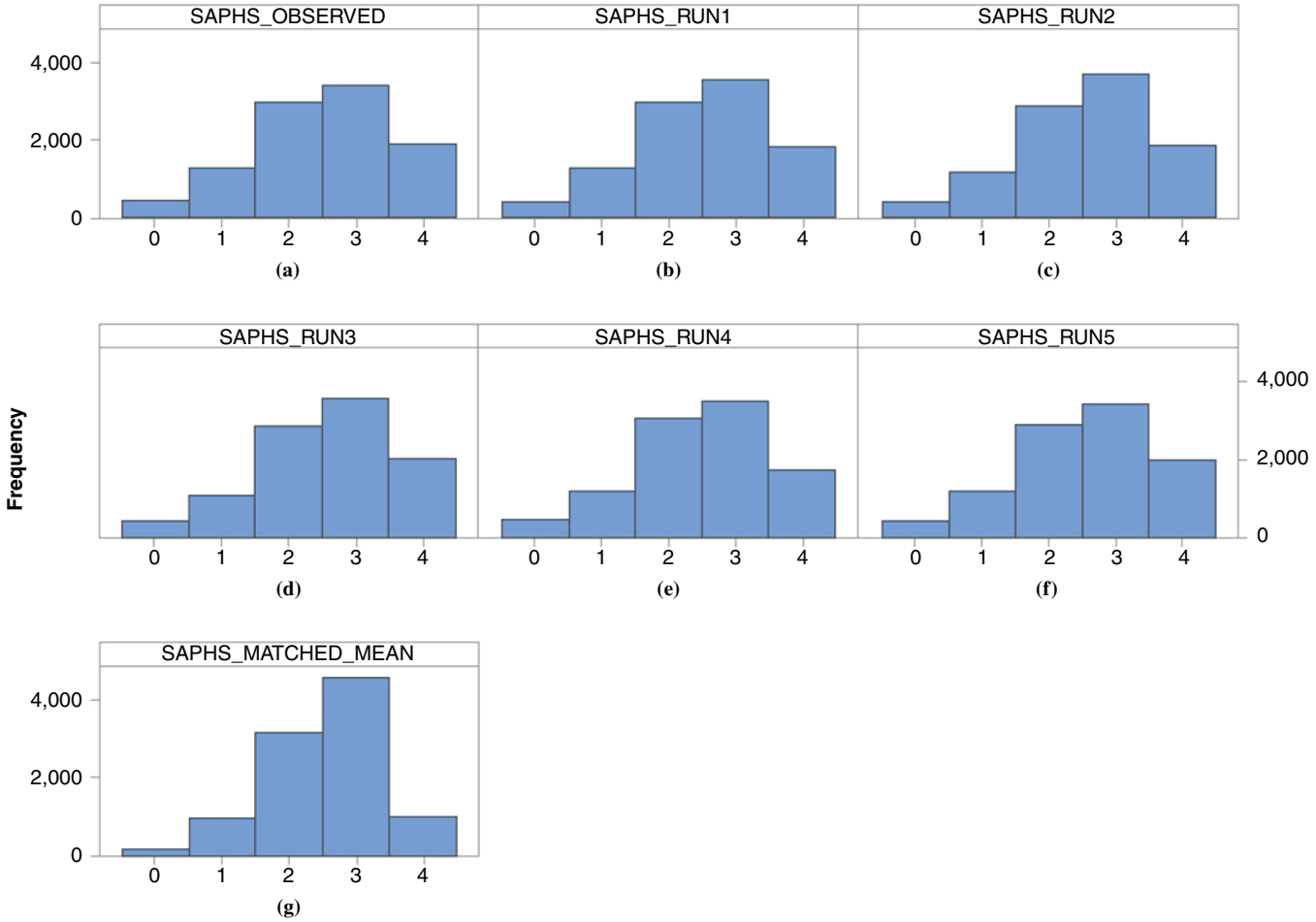

FIGURE 4 Frequency distribution of observed SAPHS, matched SAPHS for five runs, and five-sample mean (poor = 0; fair = 1; good = 2; very good = 3; excellent = 4).

FUSION OF TRAVEL AND HEALTH SURVEYS AND APPLICATION FOR MODELING

The previous section of the paper demonstrated the feasibility of data fusion and presented the results of empirical validation. In this section, the methodology outlined in the previous section is applied to fuse health data from the EH module of the ATUS (donor) with the travel data from the National Household Travel Surveys of 2008–2009 (receiver).

The ATUS data have been described previously. The NHTS data were processed to ensure consistency with the ATUS. Specifically, data records representing the add-on samples were excluded. Next, only one person aged 15 years or older was retained from each household (ATUS has data for only one person per household; NHTS surveyed all household members). This selection was accomplished by randomly choosing one person from households with multiple members equal to or older than 15 years of age.

Because the ultimate intent of this study is to use the fused data set to understand the impacts of long-term multimodal travel choices on health, the sample was further reduced to those households that participated in the extended interview. For these households, data are available for YEARMILE (vehicle-miles driven over the year), NWALKTRP (number of walk trips last week), NBIKETRP (number of bicycle trips last week), and PTUSEDDP (number of transit trips last month). Public transit availability was self-reported and those reporting no access were excluded. Similarly, individuals without a driver's license were excluded as well. Bicycle ownership was not assessed by the survey. Overall, the final receiver travel survey includes 11,362 individual respondents who had access to transit, a household vehicle, and a driver's license. The participation rates of these individuals in the four modes are presented in Table 1, divided into four quartiles (the rates are conditional on participation; the data also include those who did not use each of the modes). The donor includes the entire ATUS-EH module (about 36,000 individuals).

TABLE 1 Mode Participation Rates by Quartiles

Term	Biking Trips, Past Week	Walking Trips, Past Week	Public Transit Trips, Past Month	VMT, Past Year
Q1	1	1–3	1	1–5,000
Q2	2	4–5	2–3	5,001–10,000
Q3	3–4	6–7	4–10	10,001–15,000
Q4	5+	8+	11+	15,001+

NOTE: VMT = vehicle miles traveled.

The data fusion of the NHTS and ATUS-EH modules was performed using the Link Plus software. Five runs were performed with re-sorting of the ATUS donor data set after each run. The variables used in the blocking and matching are identified in Table 2. This table also presents the aggregate distribution of these variables in the NHTS data set and the corresponding distributions in the matched ATUS records in each run. Overall, Table 2 indicates that the NHTS records are generally matched to fairly identical households in the ATUS, leading to overall similar aggregate distributions.

TABLE 2 Comparison of ATUS Sample Matched to NHTS

Variable	Matched ATUS Donor Attributes					NHTS Receiver
	Run 1	Run 2	Run 3	Run 4	Run 5	
Data set size	35,599	35,599	35,599	35,599	35,599	11,362
Gender						
Male	48.8%	48.8%	48.8%	48.8%	48.8%	48.8%
Female	51.2%	51.2%	51.2%	51.2%	51.2%	51.2%
Age (average)	55.8	55.4	55.3	54.9	55.3	55.9
Race						
Asian	1.4%	1.4%	1.4%	1.3%	1.4%	2.5%
African American	4.5%	4.5%	4.5%	4.3%	4.5%	4.7%
Caucasian	91.2%	91.2%	91.2%	91.2%	91.1%	88.1%
Hispanic	2.0%	2.2%	2.1%	2.2%	2.1%	1.4%
Education: college (some or complete)	71.4%	71.4%	71.4%	71.4%	71.4%	71.1%
Employment						
Partial–multiple jobs	22.6%	22.1%	21.2%	22.2%	19.9%	16.1%
Full-time job	34.2%	35.3%	34.9%	33.8%	37.1%	40.2%
Household (HH)						
Members	2.35	2.35	2.35	2.35	2.35	2.36
Children	28.6%	30.1%	28.4%	29.0%	29.7%	24.6%
Owned HH	90.9%	90.9%	90.9%	90.9%	90.9%	89.8%
HH income <$35,000	24.2%	24.2%	24.1%	24.0%	23.6%	22.9%
HH income $35,000–$75,000	33.3%	34.0%	33.7%	33.7%	33.9%	33.0%
HH income >$75,000	37.7%	37.2%	37.2%	37.2%	37.4%	37.4%
Location						
Within metropolitan region	85.0%	85.0%	85.0%	85.0%	85.0%	83.7%
Northeast U.S. region	20.4%	20.4%	20.4%	20.4%	20.4%	20.4%
Midwest U.S. region	21.7%	21.7%	21.7%	21.7%	21.7%	21.7%
South U.S. region	29.7%	29.7%	29.7%	29.7%	29.7%	29.7%
West U.S. region	28.2%	28.2%	28.2%	28.2%	28.2%	28.2%

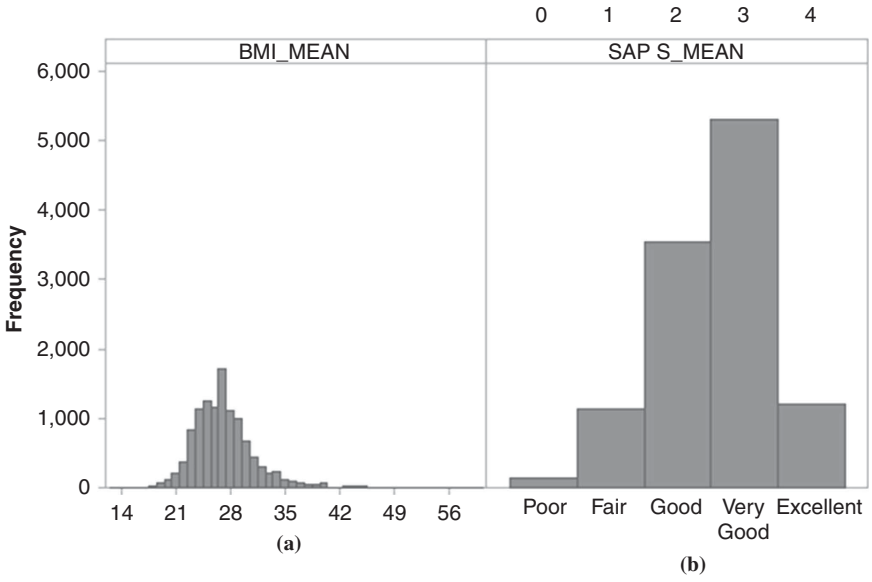

FIGURE 5 BMI and SAPHS distribution.

The distribution of the imputed BMI and SAPHS (averaged over the five runs) is presented in Figure 5. While true distributions of these measures are not available for the NHTS, the overall patterns appear reasonable and consistent with what were observed from the ATUS samples in the validation exercise.

Figures 6 and 7 present a cross tabulation of the imputed (from the ATUS) health measures against long-term uses of the walk, bike, auto, and transit modes directly obtained from the NHTS. Although BMI is intrinsically a continuous variable (and is treated as such in the models later on), it has been converted into a categorical variable for cross-tabulation purposes in Figure 6. Following the National Institutes of Health classification scheme (*17*), a person is classified into one of eight categories based on their BMI values: severely underweight (SU, BMI <14.9), very underweight (VU, BMI = 15–15.9), underweight (U, BMI = 16–18.4), normal (N, BMI = 18.5–24.9), overweight (OW, BMI = 25–29.9), obese 1 (Ob 1, BMI = 30–34.9), obese 2 or very obese (Ob 2, BMI = 35–39.9), or obese 3 or severely overweight (Ob 3, BMI >40). Because of the very small share of underweight individuals, categories SU, VU, and U are combined into an aggregate category with an overall share as 1.7% of the sample.

On visual inspection, the trends from Figures 6 and 7 do not seem consistent with the hypothesis that increased use of active modes should be correlated with better health. Therefore, disaggregate models are developed to capture the marginal impacts of modal usage after controlling for other systematic effects that can also impact health. BMI is modeled using a linear regression model (Table 3) and SAPHS is modeled (Table 4) using ordered-probit (0 = poor, 1 = fair, 2 = good, 3 = very good, and 4 = excellent) recognizing the nature of these data. Each mode was modeled individually (bicycle, walking, public transit, vehicle, etc.) for a total of eight models. Note that the level of use of each mode is described by five categorical variables representing no use and the four quartiles of use (see Table 1).

The extent of use of bike mode has no statistically significant impact on either BMI or SAPHS. This could be because of the very low levels of biking in the NHTS sample.

The models for walking indicate that those who are in the second quartile of walking (4 to 5 trips last week) have a lower BMI and feel better (higher SAPHS) compared with those who do not walk or walk fewer than 4 trips (first quartile). This trend is consistent with the expectation that walkers should be healthier. However, the model also indicates that those who walk more than 6 trips per week (quartiles 3 and 4) are in poorer health (both BMI and SAPHS) than those who walk 4 to 5 trips. Perhaps these people are having to walk despite their poor health.

The BMI model with transit use mirrors the results for the model with walking. That is, those in the second quartile of transit use (2 to 3 trips per month) have a lower BMI compared with others. In the case of the model for SAPHS, those with more than 2 to 3 transit trips per month are generally happier with their health than those with fewer than 2 transit trips per month. However, those in the fourth quartile of transit use are less satisfied than those in the second and third quartiles.

The BMI model with annual vehicle miles driven indicates that those who drive between 5,000 to 15,000 mi per year have lower BMI than those who drive more than 15,000 mi. This trend is consistent with the expected relationship that passive modes of travel are associated with higher BMI. At the same time, the model also indicates that those who drive less than 5,000 mi are not in a better health condition (BMI) compared with those who drive more. This is probably because poor health conditions are limiting the person's driving in the first place. In the case of SAPHS, the model indicates that people are happier with their health with increasing levels of driving.

Overall, the statistical analysis provides a rather interesting insight into the cross-sectional correlation patterns between the use of different modes and health. While increasing walking and transit use is associated with better health (relative to nonusers of the mode), those with the highest levels of walking and transit use are also in poor health relative to moderate users of the mode. Although the current study did control for several factors, incorporating additional controls can help further disentangle the complex relationships between health and the use of active modes.

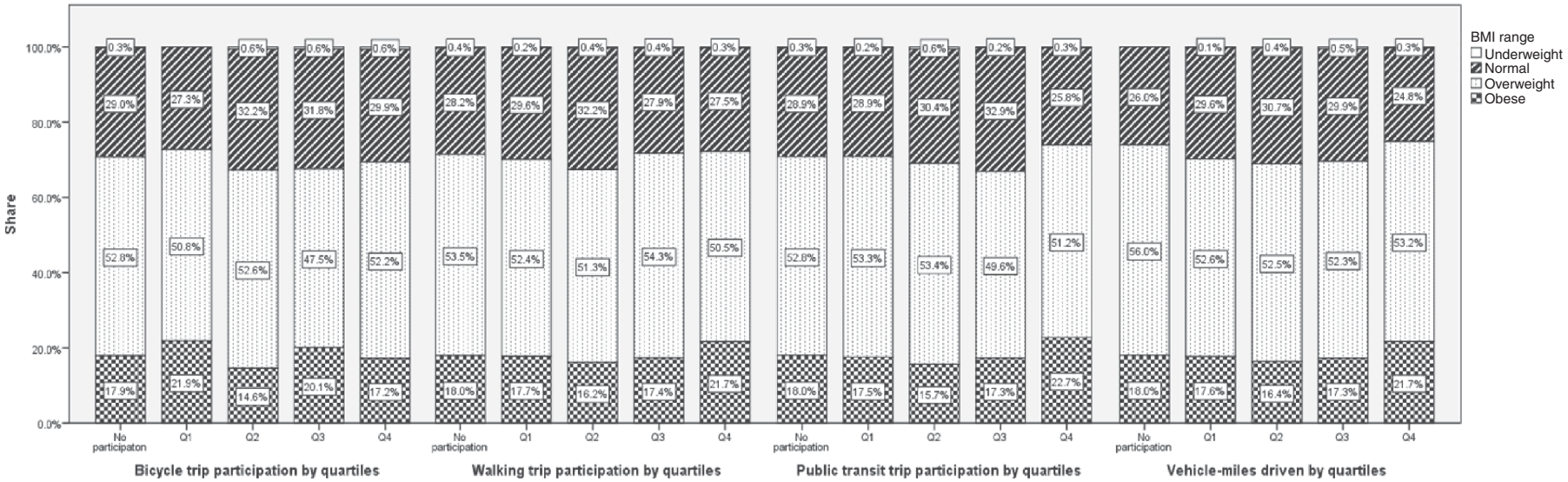

FIGURE 6 BMI range by participation rates of modes.

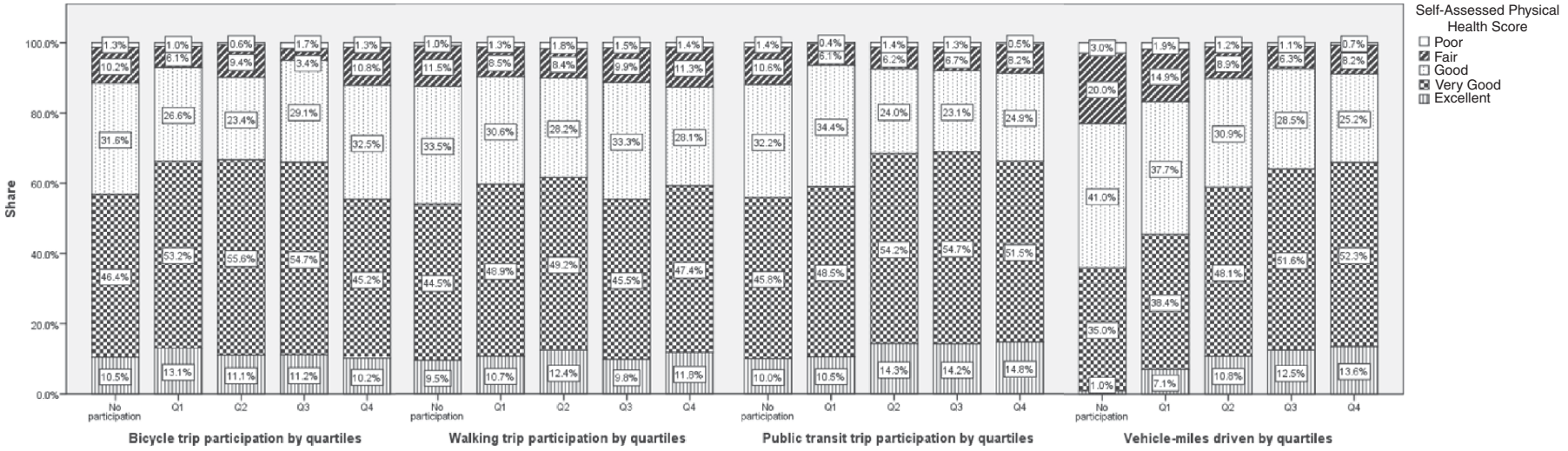

FIGURE 7 SAPHS values by participation rates of modes.

TABLE 3 Modal Impact on BMI

	Weekly Bicycle Trips		Weekly Walking Trips		Monthly Transit Trips		Yearly Vehicle Miles	
Variable	B	Significance	B	Significance	B	Significance	B	Significance
Participation								
Q1	—		—		—		—	
Q2	—		–0.27	0.01	–0.37	0.04	–0.36	<0.01
Q3	—		—		—		–0.35	<0.01
Q4	—		—		—		—	
Gender (base male): female	–1.28	<0.01	–1.28	<0.01	–1.28	<0.01	–1.29	<0.01
Age (base young adult)								
Middle adult (30–49)	1.20	<0.01	1.19	<0.01	1.20	<0.01	1.20	<0.01
Late adult (50+)	0.75	<0.01	0.75	<0.01	0.75	<0.01	0.75	<0.01
Census region (base west)								
Midwest region	1.29	<0.01	1.28	<0.01	1.28	<0.01	1.29	<0.01
Northeast region	1.00	<0.01	1.00	<0.01	1.00	<0.01	1.00	<0.01
South region	0.55	<0.01	0.55	<0.01	0.54	<0.01	0.54	<0.01
Metro (base nonmetro): metropolitan area	–0.38	<0.01	–0.38	<0.01	–0.38	<0.01	–0.37	<0.01
Vehicle ownership (base shared-vehicle household)								
Nonshared vehicles	–0.38	<0.01	–0.38	<0.01	–0.38	<0.01	–0.34	<0.01
Household adults	–0.18	<0.01	–0.18	<0.01	–0.18	<0.01	–0.18	<0.01
Constant	27.36	<0.01	27.39	<0.01	27.38	<0.01	27.48	<0.01
Adjusted R^2	.05		.05		.05		.06	

NOTE: Q = quarter; B = unstandardized coefficient; — = insignificant at $p = .05$ values.

TABLE 4 Modal Impact on Self-Assessed Physical Health Score

	Weekly Bicycle Trips		Weekly Walking Trips		Monthly Transit Trips		Yearly Vehicle Miles	
Variable	B	Significance	B	Significance	B	Significance	B	Significance
Participation								
Q1	—		—		—		—	
Q2	—		0.07	0.01	0.20	<0.01	0.26	<0.01
Q3	—		—		0.21	<0.01	0.32	<0.01
Q4	—		—		0.16	0.01	0.32	<0.01
Gender (base male): female	—		—	<0.01	—		0.09	<0.01
Age (base young adult)								
Middle adult (30–49)	–0.21	<0.01	–0.21	<0.01	–0.21	<0.01	–0.22	<0.01
Late adult (50+)	–0.52	<0.01	–0.52	<0.01	–0.51	<0.01	–0.48	<0.01
Census region (base west)								
Midwest region	–0.05	0.05	—		—		–0.05	0.04
Northeast region	—		—		—			
South region	–0.09	<0.01	–0.08	<0.01	–0.07	<0.01	–0.10	<0.01
Metro (base nonmetro): metropolitan area	0.25	<0.01	0.25	<0.01	0.24	<0.01	0.25	<0.01
Vehicle ownership (base shared-vehicle household)								
Nonshared vehicles	0.28	<0.01	0.28	<0.01	0.29	<0.01	0.22	<0.01
Household adults	0.15	<0.01	0.15	<0.01	0.15	<0.01	0.05	<0.01
Household children	0.06	<0.01	0.06	<0.01	0.06	<0.01	0.14	<0.01
Threshold = 0	1.88	<0.01	1.85	<0.01	1.83	<0.01	1.71	<0.01
Threshold = 1	–0.85	<0.01	–0.82	<0.01	–0.79	<0.01	–0.67	<0.01
Threshold = 2	0.23	<0.01	0.26	<0.01	0.28	<0.01	0.42	<0.01
Threshold = 3	1.73	<0.01	1.76	<0.01	1.79	<0.01	1.93	<0.01
Log likelihood	–2,495.55		–2,334.60		–2,541.98		–5,435.87	

NOTE: — = insignificant at $p = .05$ values.

SUMMARY AND CONCLUSIONS

While the overall volume of studies on time use, transportation choices, and health is extensive, these studies often focus on a single mode or a specific aspect of time use. Studies also often rely on short-term or one-day travel information. Given the significant day-to-day variabilities in travel patterns, the travel pattern of a single day may not be representative enough to be a predictor of long-term health patterns. Travel surveys do collect multimodal or long-term travel information, but practically none of the travel surveys collect data on health, while health surveys are generally limited in the travel data collected.

This study demonstrates the feasibility of using data fusion in the context of large-scale travel and health surveys, and subsequently used the new comprehensive data set generated to model the relationship between health and multimodal (walking, biking, transit, and vehicle usage) and long-term (weekly, monthly, and yearly) travel choices. Two measures of health are fused from a health survey onto a travel survey at the disaggregate level.

The probabilistic record linkage software Link Plus was used for the data fusion purposes. The methodology was validated using the EH module of ATUS. Subsequently, the algorithm was used to match the health information from the ATUS to the NHTS and the resulting master data set was used to develop models for multimodal travel choices and health.

The statistical analysis indicates that while increasing walking and transit use is associated with better health (relative to nonusers of the mode), those with the highest levels of walking and transit use are also in poor health relative to moderate users of the mode. Similarly, those at the two ends of the driving spectrum (first and fourth quartiles) have higher BMI compared with those in the middle of the spectrum. There were no statistically significant effects of weekly bike trips on health measures.

Overall, this study is envisioned as a proof of concept of how data fusion techniques may be used to integrate multiple data sets to facilitate a comprehensive study of multimodal travel choices and health.

Since the intent of this exercise was to impute health measures, variables that are most strongly correlated to health as matching variables were used. Several such socioeconomic factors were used which are both readily available in the surveys and have also been shown to be correlated to health from past studies. However, it is reasonable to expect that there are several other physiological, genetic, and nutritional factors that could also affect health. Having data on some of these other factors could improve the matching process. It is envisioned that future empirical studies will add to the knowledge of the most important matching variables.

The Link-Plus software from the CDC was used for this study. This software was not developed to match travel survey records (their focus was on matching medical records). Even though simultaneous blocking variables and multiple runs of the software were incorporated with re-sorting of the donor data set to address the requirements, using off-the-shelf software generally imposes limits. It is envisioned that future studies will use newly developed code that can be more flexible in its algorithms.

REFERENCES

1. Milne, A., and M. Melin. Bicycling and Walking in the United States: 2014 Benchmarking Report, Alliance for Biking and Walking, Washington, D.C., 2014.
2. Lee, I. M., and D. M. Buchner. The Importance of Walking to Public Health. *Medicine and Science in Sports and Exercise,* Vol. 40, No. 7 Supplement, 2008.
3. McCormack, G. R., and J. S. Virk. Driving Towards Obesity: A Systematized Literature Review on the Association between Motor Vehicle Travel Time and Distance and Weight Status in Adults. *Preventive Medicine,* Vol. 66, 2014, pp. 49–55.
4. Merom, D., H. P. van der Ploeg, G. Corpuz, and A. E. Bauman. Public Health Perspectives on Household Travel Surveys: Active Travel between 1997 and 2007. *American Journal of Preventive Medicine,* Vol. 39, No. 2, 2010, pp. 113–121.
5. Pucher, J., R. Buehler, D. R. Bassett, and A. L. Dannenberg. Walking and Cycling to Health: A Comparative Analysis of City, State, and International Data. *American Journal of Public Health,* Vol. 100, No. 10, 2010.
6. Lugo, M. *Walking and Healthy? On the Relationship Among Utilitarian Walking, Health, and Residential Choice.* PhD dissertation. University of Florida, Gainesville, 2015.
7. Saelens, B. E., A. Vernez Moudon, B. Kang, P. M. Hurvitz, and C. Zhou. Relation Between Higher Physical Activity and Public Transit Use. *American Journal of Public Health,* Vol. 104, No. 5, 2014, pp. 854–859.
8. Wojan, T. R., and K. S. Hamrick. Can Walking or Biking to Work Really Make a Difference? Compact Development, Observed Commuter Choice and Body Mass Index. *PloS One,* Vol. 10, No. 7, 2015.
9. Wolf, J. Applications of New Technologies in Travel Surveys. *7th International Conference on Travel Survey Methods, Costa Rica,* 2004.
10. Kusakabe, T., and Y. Asakura. Behavioural Data Mining of Transit Smart Card Data: A Data Fusion Approach. *Transportation Research Part C: Emerging Technologies,* Vol. 46, 2014, pp. 179–191.
11. Amorim, M., S. Ferreira, and A. Couto. Linking Police and Hospital Road Accident Records: How Consistent Can It Be? In *Transportation Research Record: Journal of the Transportation Research Board, No. 2432,* Transportation Research Board of the National Academies, Washington, D.C., 2014, pp. 10–16.
12. Rosman, D. L. The Western Australian Road Injury Database (1987–1996): Ten Years of Linked Police, Hospital and Death Records of Road Crashes and Injuries. *Accident Analysis & Prevention,* Vol. 33, No. 1, 2001, pp. 81–88.
13. Williamson, A., and S. Boufous. A Data-Matching Study of the Role of Fatigue in Work-Related Crashes. *Transportation Research Part F: Traffic Psychology and Behaviour,* Vol. 10, No. 3, 2007, pp. 242–253.
14. Thakuriah, V., J. Lee, and S. Niumpradit. Motor Carrier Safety Databases: Strategies for Developing Linkages, Assessing Completeness and Making Imputations. *International Truck and Bus Safety Research and Policy Symposium,* 2002.
15. Pawlak, J., J. W. Polak, and A. Sivakumar. An Imputation Approach to the Fusion of Travel Diary and Lifestyle Data: Application to the Analysis of the Interaction of ICT and Physical Mobility. *New Techniques and Technologies for Statistics Conference,* Brussels, Belgium, 2013.
16. Kressner, J. D., and L. A. Garrow. Using Third-Party Data for Travel Demand Modeling: Comparison of Targeted Marketing, Census, and Household Travel Survey Data. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2442,* Transportation Research Board of the National Academies, Washington, D.C., 2014, pp. 8–19.
17. Fellegi, I. P., and A. B. Sunter. A Theory for Record Linkage. *Journal of the American Statistical Association.* Vol. 64, No. 328, 1969, pp. 1183–1210.
18. Dempster, A. P., N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via EM Algorithm. *Journal of the Royal Statistical Society.* Series A, Vol. 153, 1977, pp. 287–320.
19. Jaro, M. A. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association,* Vol. 84, No. 406, 1989, pp. 414–420.

The Standing Committee on Environmental Justice in Transportation peer-reviewed this paper.