American Journal of Epidemiology

Vol. 181, No. 9
DOI: 10.1093/aje/kwu470
Advance Access publication:
April 8, 2015

Practice of Epidemiology

Measurement Error of Self-Reported Physical Activity Levels in New York City: Assessment and Correction

Sungwoo Lim*, Brett Wyker, Katherine Bartley, and Donna Eisenhower

* Correspondence to Dr. Sungwoo Lim, Office of the Executive Deputy Commissioner, Division of Mental Hygiene, New York City Department of Health and Mental Hygiene, Gotham Center, 42-09 28th Street, Queens, NY 11101-4132 (e-mail: slim1@health.nyc.gov).

Initially submitted August 5, 2014; accepted for publication October 29, 2014.

Because it is difficult to objectively measure population-level physical activity levels, self-reported measures have been used as a surveillance tool. However, little is known about their validity in populations living in dense urban areas. We aimed to assess the validity of self-reported physical activity data against accelerometer-based measurements among adults living in New York City and to apply a practical tool to adjust for measurement error in complex sample data using a regression calibration method. We used 2 components of data: 1) dual-frame random digit dialing telephone survey data from 3,806 adults in 2010–2011 and 2) accelerometer data from a subsample of 679 survey participants. Self-reported physical activity levels were measured using a version of the Global Physical Activity Questionnaire, whereas data on weekly moderate-equivalent minutes of activity were collected using accelerometers. Two self-reported health measures (obesity and diabetes) were included as outcomes. Participants with higher accelerometer values were more likely to underreport the actual levels. (Accelerometer values were considered to be the reference values.) After correcting for measurement errors, we found that associations between outcomes and physical activity levels were substantially deattenuated. Despite difficulties in accurately monitoring physical activity levels in dense urban areas using self-reported data, our findings show the importance of performing a well-designed validation study because it allows for understanding and correcting measurement errors.

bias; motor activity; New York City; self report; statistical model

Abbreviations: CI, confidence interval; GPAQ, Global Physical Activity Questionnaire; PR, prevalence ratio.

Editor's note: *An invited commentary on this article appears on page 656, and the authors' response appears on page 659.*

Physically active adults have lower rates of obesity, various chronic diseases, and premature death than do those who are less active (1, 2). However, according to the accelerometer data from 2005–2006 National Health and Nutritional Examination Survey, only 10% of adults in the United States meet the recommended physical activity level of 150 minutes or more per week of moderate physical activities performed in increments of at least 10 minutes (3). Monitoring and promoting physical activity has been a focus of public health efforts in recent years. Yet, objectively measuring population-level physical activity is challenging because it requires tracking a large number of people using expensive devices, such as accelerometers, and imposing strict data collection protocols. As an alternative surveillance tool, self-reported measures of physical activity can be used, and their reliability and validity have been confirmed in national and international studies (4, 5).

Little is known about the validation of this type of tool in populations living in dense urban areas. Because survey response behaviors were found to be unique among residents in these areas in a recent study about nonresponse bias, which might be attributed to individual and environmental characteristics, it is important to assess whether this tool can be adopted as a valid surveillance tool for this type of population (6). Another limitation in the current literature is that the use of the practical calibration method to adjust for measurement error in studies of self-reported physical activity is rarely

demonstrated. Thus, the purpose of the present study was to assess the validity of self-reported physical activity data against accelerometer-based measurements among a sample of adults from New York City who participated in both a telephone survey and a device follow-up study. An additional aim was to demonstrate a practical approach to adjust for measurement error in complex sample data using a regression calibration method.

METHODS

Physical Activity and Transit Survey

The Physical Activity and Transit Survey was a dual-frame random digit dialing telephone survey of adults in New York City. In the present study, we used an overlapping landline and cell phone sample frame to contact adults in residential households in New York City, with equal-sized samples from the 5 boroughs (which became disproportionate because the population sizes in the boroughs differed) plus oversamples from areas with higher levels of obesity. Interviews were conducted in 2010–2011 (American Association for Public Opinion Research response rate 3 = 38%), and they generated a final data set of 3,806 completed interviews (7). Of the participants who completed the survey, 2,488 persons who were able to walk more than 10 feet were identified, and we asked them to participate in the device follow-up study in which their activity levels would be recorded objectively using GT3x accelerometer devices (Actigraph, Pensacola, Florida). Of this subsample, 803 (32%) agreed to participate and returned the devices with data. The minimum accelerometer wear-time for a reliable estimate of weekly activity was 10 or more hours on 4 or more days (8), and use of this cutoff as the inclusion criterion resulted in 679 participants who completed the device follow-up study.

Subjective physical activity measure

Self-reported physical activity levels were measured using the Global Physical Activity Questionnaire (GPAQ). The GPAQ was developed by the World Health Organization to measure physical activity levels across 3 domains of activity that occur on a typical day (work around the home and in the workforce, transportation, and recreation) in culturally diverse populations (4, 9). The version of the GPAQ that was used for the interviews (Web Appendix 1, available at http://aje.oxfordjournals.org/) was slightly modified to obtain information on activity performed as part of paid work separately from activity related to house work rather than for all work (including unpaid work, study/training, and household chores) combined. Respondents were asked whether their work (if employed), house work, transit, and recreation time involved physical activity, including walking, during the past 7 days. If so, they were then asked to report the number of days that the activity caused an increase in breathing or heart rate (considered to be moderate physical activity) and the average time spent doing the activity each day. Additionally, in the workforce and recreational domains, respondents were asked to report the number of days and average length of time spent doing vigorous activity, defined as activity that caused a large increase in breathing or heart rate. For each domain, total weekly minutes of moderate and vigorous activity were calculated by multiplying the number of days the respondent engaged in physical activity by the daily average minutes of moderate or vigorous physical activity. Because vigorous activity requires approximately twice the energy expenditure of moderate activity (1), the number of minutes of vigorous activity was multiplied by 2 to create "moderate-equivalent minutes." Activity across all domains was summed to create a variable of total weekly moderate-equivalent minutes for each participant. If a participant reported a daily average of 960 minutes or more (≥16 hours) of activity in any domain or if there were any inconsistent values in a participant's response, such as reporting more than 7 days of activity in a week or reporting 0 days of activity but then having a value greater than 0 for minutes or hours of activity ($n = 174$), the participant was not included in the analysis as per the analytic guidelines established by the World Health Organization (10).

Objective physical activity measure

Accelerometers were worn by participants in the device follow-up study for 1 week throughout the day other than while sleeping, swimming, or bathing. Because the devices were worn at home, while working, while in transit, and during recreational time, the data were comparable to the self-report data on activity in these domains. To process the accelerometer data for analysis, activity thresholds from the National Health and Nutrition Examination Survey were used to categorize minutes as moderate or vigorous. These thresholds were established in calibration studies that compared activity counts recorded by accelerometers to measured energy expenditure during walking and running on a treadmill or track (11–13). All accelerometer minutes that ranged from 2,020 through 5,998 counts per minute were considered moderate and minutes at or above 5,999 counts per minute were considered vigorous. To create moderate-equivalent minutes, vigorous minutes from the accelerometer data were multiplied by 2 using the same process that was used for the survey data. To obtain weekly physical activity values for participants who had days with missing or invalid data or those who had 4–6 days of valid data ($n = 336$), the minutes of moderate-equivalent activity on valid days (≥10 hours of wear time) were summed and divided by the number of valid days of wear-time to create a daily average moderate-equivalent activity variable. The daily average was multiplied by 7 to create a weekly total.

Other measures

We included demographic characteristics that have been found to be associated with physical activity levels and with measurement error in self-reported physical activity levels in previous studies, including age, sex, race/ethnicity, educational level, household poverty status, and employment status (14–16). We assessed validity of the GPAQ measure stratified by these variables as well as time spent wearing the device. We also included these demographic variables as covariates of the measurement model for a regression calibration

process. To demonstrate the regression calibration method, we used obesity (based on self-reported height and weight) and self-reported current diabetes as outcomes.

Statistical analysis

Validation. To evaluate validity of the self-reported GPAQ measure, we quantified the extent to which the GPAQ measure was valid against the accelerometer-based measure by using a measurement error model. We first constructed the linear regression model with the GPAQ measure as the dependent variable and the accelerometer-based measure as the independent variable (17). Because total values for physical activity minutes were highly skewed and there was a 0 value, we replaced the actual measures with those transformed via the inverse hyperbolic sine function to satisfy linearity and normality assumptions. This function addressed skewness in a fashion similar to that of log transformation except for an additional capacity to map 0 to 0. The measurement error model was specified as

$$Q_i = \alpha_Q + \beta_Q T_i + \varepsilon_{Q_i},$$

where Q_i indicates the transformed GPAQ measure, T_i indicates the transformed accelerometer-based measure, and i indicates 1, 2, 3, . . ., 679 individuals. We tested the assumption that errors of subjective and objective measures are independent, which is the assumption required for deriving validity indicators from the measurement error model (18). According to the graphical observation and estimated correlation coefficients, this assumption was unlikely to be violated (data not shown). Then, using parameters from the model, we calculated the validity coefficient $(\hat{\rho}_{QT})$ and attenuation factor $(\hat{\lambda})$ as follows:

$$\hat{\rho}_{QT} = \sqrt{\frac{1}{1 + (\hat{\sigma}^2_{\varepsilon_Q}/\hat{\beta}^2_Q\hat{\sigma}^2_T)}}, \quad \hat{\lambda} = \frac{\hat{\beta}_Q}{\hat{\beta}^2_Q\hat{\sigma}^2_T + \hat{\sigma}^2_{\varepsilon_Q}}.$$

The validity coefficient quantifies the validity of the GPAQ measure relative to the accelerometer-based measure (18). The attenuation factor represents the extent to which the relationship between the GPAQ measure and disease outcome is attenuated because of measurement error in GPAQ measure (18). Both indicators range from 0 to 1, and those close to 1 represent a higher level of validity and less attenuation due to measurement errors. In addition, we plotted the data using a Bland-Altman plot, a graphical tool that can be used to assess the validity of measurement. Because we considered the accelerometer-based measures to be the reference values, we plotted the accelerometer data on the x-axis of the plot instead of the combined mean scores of GPAQ and accelerometer-based measures (19). If difference between GPAQ and accelerometer-based measures is less than 2 standard deviations of mean difference, we can conclude that agreement between the 2 measures is good (20).

Regression calibration method. In order to correct measurement error in the GPAQ measure after assessing its validity, we adopted a regression calibration method proposed by Rosner et al. (18). The first step in this method was to create 2 types of regression models: 1) a measurement error model to estimate measurement errors and 2) a main model to adjust estimates for measurement errors. The measurement error model was a linear regression for a transformed accelerometer-based measure using the device follow-up data. The main model included 2 separate log-Poisson models for obesity and diabetes outcomes. Both the measurement and main models shared the same set of independent variables, including the transformed GPAQ measure, age, sex, race/ethnicity, household poverty status, educational level, and employment status.

Once the 2 types of models were created, we tested whether our data met the assumptions required for the regression calibration method. We specifically used the procedures described by Horick et al. (21). First, linearity assumptions in both the measurement error and the main models were tested. Specifically, we tested 2 hypotheses: 1) that the transformed GPAQ measure was linearly related to the transformed accelerometer-based measure in the measurement error model and 2) that the transformed GPAQ measure was linearly related to the log of health outcomes in the main models. In each hypothesis, a linear model was compared with a nonlinear one (e.g., a model with a squared GPAQ measure), and the results of F tests showed that the linear model was preferable to the nonlinear one. Second, we tested the assumption that the variance was homoscedastic. No clear pattern was detected in the residual plot compared with the fitted plot, suggesting homoscedastic variance. Third, adding the GPAQ measure to the model with the accelerometer-based measure did not improve power, suggesting that the GPAQ measure was properly considered to be a surrogate of the accelerometer-based measure. Lastly, we assessed the severity of the measurement error, which was considered moderate according to the estimated multiple correlation coefficient $(\sqrt{R^2} = 0.54)$ in the measurement model.

After confirming that the data satisfied all of the assumptions, we first obtained estimates of prevalence ratios (PR) of obesity and diabetes by physical activity levels $(\exp(\hat{\beta}))$ from the main models and then adjusted these estimates to correct the measurement error by taking the exponent of $(\hat{\beta}/\hat{\gamma})$, where $\hat{\gamma}$ represents a coefficient of the transformed GPAQ measure in the measurement error model (18, 21). We further adjusted estimates for the complex sample design by using the bootstrap method (22). Specifically, we constructed 1,000 bootstrap replicates that consisted of random samples of size $n_h - 1$ per each stratum (h) with replacement and calculated variances using replicate weights using the formula $w_i \times n_h/n_h - 1 \times m_i$, where w_i is an original survey weight for individual i and m_i is the number of times that individual i is selected in each bootstrap sample (22). There were some missing data (for diabetes, 5%; for obesity, 5%; for age, <1%; for educational level, <1%; for employment status, <1%; and for household poverty status, 10%), and to reduce potential bias resulting from excluding missing data in the analysis, we performed multiple imputations using *IVEware* software (University of Michigan, Ann Arbor, Michigan) that adopted the sequential regression method (23). Multiple imputations generated 5 imputed data sets, and the combined results of 5 estimates that accounted for within- and between-imputation variability were reported according to Schafer's approach (24). Except for imputation, all the analyses were performed

Table 1. Demographic Characteristics and Physical Activity Profiles of Participants, Physical Activity and Transit Survey, New York City, 2010–2011

Characteristic	Main Survey, % (n = 3,806)	Device Follow-up Study, % (n = 679)
Sex		
Male	40	39
Female	60	61
Race/ethnicity		
Non-Hispanic white	43	44
Non-Hispanic black	24	27
Hispanic	23	22
Asian	8	5
Other	2	2
Age, years		
18–24	7	6
25–44	30	33
45–64	38	42
≥65	25	20
Employment status		
Employed	52	60
Unemployed	9	9
Not in the labor force[a]	40	31
Educational level		
Not a high school graduate	14	9
High school graduate	25	23
Some college	21	23
College graduate	40	45
Household income[b]		
≤200% of the federal poverty level	38	35
200%–399% of the federal poverty level	17	19
≥400% of the federal poverty level	38	42
Body mass index category[c]		
Underweight/normal weight	40	38
Overweight	34	35
Obese	26	27
Physical activity level		
Weekly moderate-equivalent activity minutes from the Global Physical Activity Questionnaire	380 (120, 840)[d]	435 (175, 940)[d]
Weekly moderate-equivalent activity minutes from accelerometer		188 (86, 331)[d]

[a] This category included homemakers, students, and retirees.
[b] Missing data were not presented.
[c] Weight (kg)/height (m)^2.
[d] Values are expressed as median (25th, 75th percentiles).

using SAS, version 9.2 (SAS Institute, Inc., Cary, North Carolina). Statistical significance was tested with 2-sided P values <0.05.

RESULTS

Demographic characteristics were generally similar between survey participants (n = 3,806) and those in the subsample who wore the accelerometer devices (n = 679) except that more subjects in the latter group had higher levels of education and household income and were more likely to be currently employed (Table 1). On the other hand, participants in the device follow-up study appeared to be more physically active than survey participants. The median number of weekly moderate-equivalent activity minutes from the GPAQ survey was 435 minutes (25th percentile, 175; 75th percentile, 940) among the subsample and 380 minutes (25th percentile, 120; 75th percentile, 840) among survey participants. Lastly, the median of weekly moderate-equivalent activity minutes was 188 minutes according to the accelerometer measure.

The overall validity coefficient for the GPAQ measure compared with the accelerometer-based measure was 0.19 (95% confidence interval (CI): 0.13, 0.25) (Table 2). When stratified by demographic characteristics, the validity coefficient varied between 0.09 and 0.26. In particular, women, non-Hispanic whites, and individuals with high school degrees or higher appeared to more accurately report their physical activity levels than did others. Additionally, being younger (18–24 years vs. 25–44 years), being employed (vs. unemployed), having a lower body mass index (underweight/normal weight vs. overweight), and wearing the device for a shorter period of time were positively associated with the accuracy of self-reported physical activity levels. The overall attenuation factor was 0.14 (95% confidence interval (CI): 0.10, 0.17), which indicated that the measurement error would substantially attenuate the PR associated with physical activity levels. If the true PR were 0.50, the biased PR would be 0.96. Similar to the validity coefficient, the attenuation factor varied across demographic characteristics. For example, the PR associated with physical activity levels would be less likely to be underestimated for those who were unemployed than for those not in the labor force. In contrast, persons who were 65 years of age or older were more likely to report biased physical activity levels than were younger participants, leading to greater attenuation of the association between health outcomes and physical activity levels.

The Bland-Altman plot in Figure 1 shows that self-reported physical activity levels among most participants were relatively similar to the accelerometer-based measurements and were within 2 standard deviations of the mean difference between 2 measures, indicating good agreement. Although there were a few individuals who had differences between the 2 measures that were greater than 2 standard deviations of mean differences, indicating substantial over-reports of their physical activity levels (7% of 679 device follow-up participants), those with higher accelerometer values (i.e., located toward the right side of x-axis) were more likely to underreport their physical activity on the GPAQ and were more likely to be located in the area below 0 difference (Figure 1, Web Figure 1). As the objectively measured physical

Table 2. Estimated Validity Coefficients and Attenuation Factors in Moderate-Equivalent Physical Activity Minutes per Week Measured Using Self-Reported Data Versus Accelerometer-Based Data, Physical Activity and Transit Survey, New York City, 2010–2011

Characteristic	Attenuation Factor[a]	95% CI[b]	*P* Value[c]	Validity Coefficient[d]	95% CI[b]	*P* Value[c]
Overall	0.14	0.10, 0.17		0.19	0.13, 0.25	
Sex						
Male	0.12	0.06, 0.18	Referent	0.14	0.08, 0.22	Referent
Female	0.15	0.10, 0.20	0.13	0.21	0.14, 0.30	0.02
Race/ethnicity[e]						
Non-Hispanic white	0.13	0.08, 0.19	Referent	0.24	0.14, 0.35	Referent
Non-Hispanic black	0.11	0.05, 0.18	0.26	0.13	0.07, 0.20	<0.001
Hispanic	0.19	0.07, 0.31	0.02	0.16	0.06, 0.26	<0.001
Age, years						
18–24	0.42	−0.03, 0.86	Referent	0.19	0.02, 0.41	Referent
25–44	0.14	0.05, 0.23	<0.001	0.09	0.03, 0.16	<0.001
45–64	0.13	0.05, 0.19	<0.001	0.14	0.06, 0.21	0.06
≥65	0.07	0.01, 0.12	<0.001	0.15	0.04, 0.28	0.14
Employment status						
Employed	0.13	0.06, 0.24	Referent	0.12	0.05, 0.27	Referent
Unemployed	0.27	0.13, 0.45	<0.001	0.18	0.09, 0.29	0.01
Not in the labor force[f]	0.07	0.03, 0.11	0.01	0.15	0.07, 0.25	0.17
Educational level						
Not a high school graduate	0.08	−0.05, 0.23	Referent	0.09	−0.04, 0.26	Referent
High school graduate	0.14	0.06, 0.22	0.01	0.18	0.09, 0.38	<0.001
Some college	0.12	0.04, 0.19	0.04	0.15	0.06, 0.30	0.02
College graduate	0.17	0.09, 0.24	<0.001	0.26	0.16, 0.35	<0.001
Household income						
≤200% of the federal poverty level	0.13	0.06, 0.20	Referent	0.18	0.10, 0.29	Referent
200%–399% of the federal poverty level	0.13	0.05, 0.21	0.96	0.19	0.10, 0.30	0.55
≥400% of the federal poverty level	0.12	0.05, 0.19	0.62	0.16	0.07, 0.27	0.56
Body mass index[g] category						
Underweight/normal weight	0.17	0.09, 0.25	Referent	0.22	0.13, 0.35	Referent
Overweight	0.14	0.08, 0.21	0.23	0.15	0.09, 0.23	0.02
Obese	0.10	0.04, 0.17	<0.001	0.17	0.07, 0.30	0.07
Time spent wearing the device, days[h]						
<3.6	0.12	0.05, 0.18	Referent	0.22	0.11, 0.38	Referent
3.6–3.8	0.14	0.07, 0.23	0.23	0.19	0.11, 0.31	0.36
3.9–4.1	0.17	0.07, 0.28	0.03	0.18	0.09, 0.31	0.23
≥4.2	0.11	0.01, 0.22	0.80	0.09	0.01, 0.17	<0.001

Abbreviation: CI, confidence interval.

[a] The attenuation factor indicates the extent to which a relationship between self-reported physical activity levels and health outcome is attenuated by measurement error in self-reports.

[b] Based on 1,000 bootstrap replicates.

[c] *P* values were based on permutation tests for the pair-wise difference between 2 subgroups.

[d] The validity coefficient is an indicator of the validity of the self-reported measure relative to the accelerometer-based measure.

[e] Indicators for Asian and other racial/ethnic groups were not presented because of the small sample size.

[f] This category included homemakers, students, and retirees.

[g] Weight (kg)/height $(m)^2$.

[h] Categorized by quartile of time spent wearing the device.

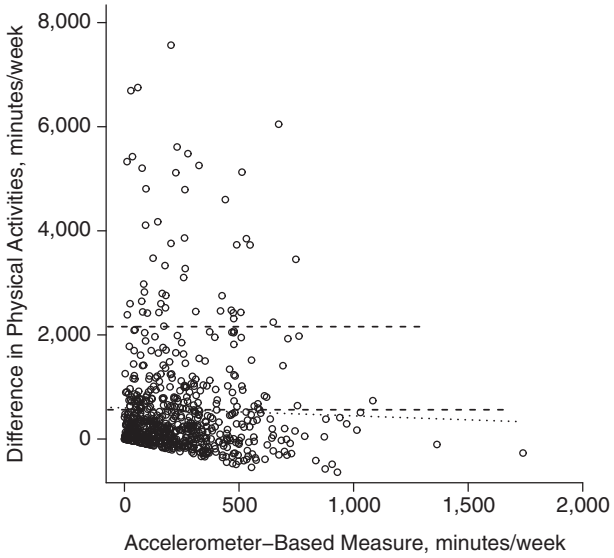

Figure 1. Bland-Altman plot of moderate-equivalent physical activity minutes per week measured by self-report and accelerometers, Physical Activity and Transit Survey, New York City, 2010–2011. The bottom dashed line represents the mean difference between self-reported and accelerometer-based physical activity measures (563 weekly minutes), and the top dashed line represents 2 standard deviations of the mean difference (2,158 weekly minutes). The dotted line represents a regression line between differences in self-reports and accelerometer-based physical activity measures (dependent variable) and accelerometer-based physical activity measures (independent variable). The *y*-axis refers to differences in between self-reported and accelerometer-based physical activity.

activity levels increased, the GPAQ measure was lower than the accelerometer-based measure (Spearman's correlation = −0.11).

Table 3 shows that a 1-unit increase of the transformed self-reported physical activity levels (e.g., an increase from 100 minutes to 1,000 minutes) was associated with a 0.89 times (95% CI: 0.84, 0.95) lower prevalence of diabetes before measurement errors were corrected. Similarly, self-reported physical activity was associated with the uncorrected prevalence of obesity (PR = 0.94, 95% CI: 0.89, 0.99). After adjustment for measurement errors via the regression calibration method, the inverse association between physical activity and the outcomes was strengthened (for obesity, PR = 0.66, 95% CI: 0.41, 1.07; for diabetes, PR = 0.47, 95% CI: 0.24, 0.89). Another impact of this adjustment was that confidence intervals became wider, reflecting additional uncertainties using parameter estimates from the measurement error model and measurement error itself.

DISCUSSION

We assessed the validity of self-reported physical activity data compared with accelerometer-based measurements by analyzing both telephone survey data and validation data from a large sample of the adult population in New York City. The low validity coefficients indicated that there were measurement errors in self-reported data compared with objective data from accelerometer devices, which in turn attenuated the association between health outcomes and physical activity. We demonstrated that these measurement problems could be somewhat ameliorated by using the regression calibration method.

There have been 2 other studies that reported the validity of self-reported physical activity levels using the measurement error model framework (17, 25). The overall validation coefficient in this study was smaller than those reported in these previous studies (17, 25). This difference was partially driven by the 7% of the device follow-up participants who overreported their physical activity levels by 2,158 weekly minutes or more (i.e., 2 or more standard deviations of the mean difference between the GPAQ and accelerometer-based measure). There were no unique characteristics that distinguished overreporters from the other participants except for a greater proportion of participants who were employed (75% vs. 59%).

After we excluded these individuals, the validation coefficient in our study improved, but it was still lower than those in the national populations in Canada (0.20 vs. 0.24) (17) and the United States (for men, 0.16 vs. 0.41; for women 0.23 vs. 0.32) (25). The Bland-Altman plot without outlying cases revealed a tendency toward underreporting among physically active participants (Web Figure 1). It might be attributable to the built environment and widespread public transportation in New York City, which lead to more active body movements than does living in rural and suburban areas, thus potentially leading to people being more physically active than perceived. This explanation is supported by the finding that accuracy of self-reported physical activity levels was negatively associated with time spent wearing the device. In addition, in the recent report from New York City Department of Health

Table 3. Prevalence Ratios for Obesity and Diabetes by Physical Activity Level Before and After Correction for Measurement Errors, Physical Activity and Transit Survey, New York City, 2010–2011

Physical Activity	Obesity				Diabetes			
	Uncorrected PR	95% CI[a]	Corrected PR	95% CI[a]	Uncorrected PR	95% CI[a]	Corrected PR	95% CI[a]
IHS[b] (weekly physical activity minutes)	0.94	0.89, 0.99	0.66	0.41, 1.07	0.89	0.84, 0.95	0.47	0.24, 0.89

Abbreviations: IHS, inverse hyperbolic sine function; PR, prevalence ratio.
[a] Based on 1,000 bootstrapping replicates that accounted for complex sample design.
[b] $\text{IHS} = \log(y + (y^2 + 1)^{1/2})$.

and Mental Hygiene, 36% of adults who lived in densely populated areas of the city met the physical activity guidelines for Americans (i.e., at least 150 minutes of physical activity per week in increments of at least 10 minutes), a percentage that was much higher than that for persons living in city neighborhoods with low densities (10%) and for the United States overall (11%) (1, 26). Another possible explanation is that objective physical activity levels might be underestimated for physically active participants because vigorous minutes were doubled to create moderate-equivalent minutes. Yet, this does not seem convincing because similar results were observed when we re-ran the analyses using aggregated minutes of moderate and vigorous activities (data not shown). A more in-depth study is warranted to identify individual and environmental factors that might influence perceived physical activity levels among people living in dense urban areas.

The validity coefficients varied by some population characteristics. Women were more likely to accurately report physical activity levels than were men, which contradicts the previous findings that women were more likely to overreport socially desirable behaviors, such as physical activity, because women are more conscious of social desirability than are men (17, 25, 27). In the present study, social desirability might not have played a role in the reporting process because we collected data on self-reported physical activities via telephone, whereas the previous data were based on face-to-face surveys (17, 25, 28). On the other hand, the directions of the associations of validity of GPAQ with body mass index or age were similar to those in the previous surveys (17, 25). Individuals with younger ages and lower body mass indices were more likely to report their physical activity levels accurately.

After correcting measurement errors, we found that the associations of physical activity with obesity and diabetes were substantially strengthened. The corrected PR for diabetes remained statistically significant, whereas the PR for obesity became statistically nonsignificant after correction of the measurement error. Using the regression calibration method, we were able to reveal an association between physical activity and health outcomes that would have been considered negligible. The corrected association between diabetes and physical activity in this study was consistent with the results in the United States national population according to the Medical Expenditure Panel Survey, 2000–2002 (28).

The study had a few limitations. First, because data on physical activity levels were only collected once, we were not able to estimate a random component of measurement error. Second, because accelerometers, although objective, could not measure certain forms of physical activity, such as water activities and stationary activities (e.g., weight-lifting), the true physical activity levels might have been underestimated. Yet, there were several notable strengths of the present study. First, because of the large sample size in the device follow-up study, we were able to test all the assumptions required for the regression calibration method with sufficient statistical power and to obtain an unbiased PR. Second, we presented a practical approach to test key assumptions and account for complex sample design in the regression calibration method.

In conclusion, we found substantial measurement error in self-reported physical activity data among urban populations living in densely populated areas. This indicates difficulty in accurately monitoring physical activity levels using self-reported data, which might be attributed to response behaviors among employed individuals or to the unique built environments of New York City. This finding highlights the importance of performing a well-designed validation study because it allows for understanding and correction of measurement error using the regression calibration method.

ACKNOWLEDGMENTS

Author affiliations: Office of the Executive Deputy Commissioner, Division of Mental Hygiene, the New York City Department of Health and Mental Hygiene, New York, New York (Sungwoo Lim); Phoenix Multisport, Boulder, Colorado (Brett Wyker); and Bureau of Epidemiology Services, Division of Epidemiology, the New York City Department of Health and Mental Hygiene, New York, New York (Katherine Bartley, Donna Eisenhower).

The New York City Physical Activity and Transit Survey was funded through the Communities Putting Prevention to Work (CPPW) Enhanced Evaluation Initiative (3U58DP002418-01S1), and the biometric measurement was funded as an expansion to this initiative through the CPPW Obesity Supplemental Evaluation Activities grant (1U58DP002418-01).

We thank Dr. Thomas A. Farley for his support on the New York City Physical Activity and Transit Survey, Dr. Marie Keem for her help in the initial survey design, Enver Holder-Hayes, Dr. Jennifer Norton, and Stephen Immerwahr for their help in putting the databases together, and Dr. Tiffany G. Harris for her helpful comments.

Conflict of interest: none declared.

REFERENCES

1. Physical Activity Guideline Committee. Physical activity guidelines advisory committee report, 2008. Washington, DC: U.S. Department of Health and Human Services; 2008. http://www.health.gov/paguidelines/. Accessed July 1, 2014.
2. Warburton DER, Nicol CW, Bredin SSD. Health benefits of physical activity: the evidence. *CMAJ*. 2006;174(6):801–809.
3. Tucker JM, Welk GJ, Beyler NK. Physical activity in U.S. adults: compliance with the Physical Activity Guidelines for Americans. *Am J Prev Med*. 2011;40(4):454–461.
4. Bull FC, Maslin TS, Armstrong T. Global physical activity questionnaire (GPAQ): nine country reliability and validity study. *J Phys Act Health*. 2009;6(6):790–804.
5. Craig CL, Marshall AL, Sjöström M, et al. International physical activity questionnaire: 12-country reliability and validity. *Med Sci Sports Exerc*. 2003;35(8):1381–1395.
6. Lim S, Immerwahr S, Lee S, et al. Estimating nonresponse bias in a telephone-based health surveillance survey in New York City. *Am J Epidemiol*. 2013;178(8):1337–1341.
7. The American Association for Public Opinion Research. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. 7th ed. Deerfield, IL: The American Association for Public Opinion Research; 2011.

8. Trost SG, McIver KL, Pate RR. Conducting accelerometer-based activity assessments in field-based research. *Med Sci Sports Exerc*. 2005;37(11 suppl):S531–S543.
9. Armstrong T, Bull F. Development of the World Health Organization Global Physical Activity Questionnaire (GPAQ). *J Public Health*. 2006;14(2):66–70.
10. World Health Organization. Global physical activity questionnaire and analysis guide. http://www.who.int/entity/chp/steps/resources/GPAQ_Analysis_Guide.pdf. Accessed February 20, 2013.
11. Freedson PS, Melanson E, Sirard J. Calibration of the Computer Science and Applications, Inc. accelerometer. *Med Sci Sports Exerc*. 1998;30(5):777–781.
12. Leenders NY, Sherman WM, Nagaraja HN, et al. Evaluation of methods to assess physical activity in free-living conditions. *Med Sci Sports Exerc*. 2001;33(7):1233–1240.
13. Yngve A, Nilsson A, Sjostrom M, et al. Effect of monitor placement and of activity setting on the MTI accelerometer output. *Med Sci Sports Exerc*. 2003;35(2):320–326.
14. Haskell WL, Lee I, Pate RR, et al. Physical activity and public health: updated recommendation for adults from the American College of Sports Medicine and the American Heart Association. *Circulation*. 2007;116(9):1081–1093.
15. Van Domelen DR, Koster A, Caserotti P, et al. Employment and physical activity in the U.S. *Am J Prev Med*. 2011;41(2):136–145.
16. Parks SE, Housemann RA, Brownson RC. Differential correlates of physical activity in urban and rural adults of various socioeconomic backgrounds in the United States. *J Epidemiol Community Health*. 2003;57(1):29–35.
17. Ferrari P, Friedenreich C, Matthews CE. The role of measurement error in estimating levels of physical activity. *Am J Epidemiol*. 2007;166(7):832–840.
18. Rosner B, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Stat Med*. 1989;8(9):1051–1069.
19. Krouwer JS. Why Bland-Altman plots should use *X*, not (*Y*+*X*)/2 when *X* is a reference method. *Stat Med*. 2008;27(5):778–780.
20. Sedgwick P. Limits of agreement (Bland-Altman method). *BMJ*. 2013;346:f1630.
21. Horick N, Weller E, Milton DK, et al. Home endotoxin exposure and wheeze in infants: correction for bias due to exposure measurement error. *Environ Health Perspect*. 2006;114(1):135–140.
22. Sitter RR. Comparing three bootstrap methods for survey data. *Can J Stat*. 1992;20(2):135–154.
23. Raghunathan TE, Lepkowski JM, VanHoewyk J, et al. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv Methodol*. 2001;27(1):85–95.
24. Schafer JL. *Analysis of Incomplete Multivariate Data*. London, UK: Chapman Hall; 1997.
25. Tooze JA, Troiano RP, Carroll RJ, et al. A measurement error model for physical activity level as measured by a questionnaire with application to the 1999–2006 NHANES questionnaire. *Am J Epidemiol*. 2013;177(11):1199–1208.
26. Bartley K, Wyker B, Eisenhower D. Physical activity measured by accelerometer: a comparison of New York City and the nation. Epi Data Brief 2013. http://www.nyc.gov/html/doh/downloads/pdf/epi/databrief22.pdf. Published February 2013. Accessed July 1, 2014.
27. Adams SW, Matthews CE, Ebbeling CB, et al. The effect of social desirability and social approval on self-reports of physical activity. *Am J Epidemiol*. 2005;161(4):389–398.
28. Sullivan PW, Morrato EH, Ghushchyan V, et al. Obesity, inactivity, and the prevalence of diabetes and diabetes-related cardiovascular comorbidities in the U.S., 2000–2002. *Diabetes Care*. 2005;28(7):1599–1603.