HCHS/SOL Physical Activity Data Overview, Methods and Guidelines

May 2014
Version 2

Prepared by
HCHS/SOL Coordinating Center
Collaborative Studies Coordinating Center
UNC Department of Biostatistics

Daniela Sotres-Alvarez
Kelly R. Evenson
Natalia Gouskova
Sonia Davis

Please send questions, suggestions and comments to dsotres@unc.edu

Table of Contents

Updates to Physical Activity Data and Documentation

PA_DERV dataset is being released again to HCHS/SOL investigators in May 2014 (INV4). A new variable for Actical IPW sampling weight (WEIGHT_PA_IPW_OVERALL) was added. This weight allows accounting for missing data due to either participants not returning the accelerometer or participants not being compliant. As a consequence, the document "Physical Activity Data Overview, Methods and Guidelines" is updated.

Updates in Version 2 (May 2014)

- NEW section 6 describes the development of the Actical IPW weights in detail, provides sample SAS and SUDAAN code, as well as suggested text for statistical methods sections of manuscripts.
- Table 3.2 data correction.

FORWARD

Note to User of the HCHS/SOL Physical Activity Data Overview, Methods, and Guidelines:

This document is not intended for direct citation or distribution outside of the immediate HCHS/SOL Study. It should be considered confidential and proprietary to HCHS/SOL investigators.

- **Physical Activity Data Overview Sections 1 and 2:** present a brief summary of physical activity assessment at baseline and the corresponding datasets.
- **Methods Section 3:** describes the methods used to clean and create the objectively measured (Actical) physical activity datasets for analytic use and to produce derived physical activity variables.
- **Methods Section 6:** describe the development of the Actical IPW weights.
- **Guidelines** for using the Actical data are in sections 3 to 6.

1. Physical Activity Assessment in HCHS/SOL

Physical activity (PA) was measured using two different approaches:

- **Objective assessment** of physical activity obtained by the participant wearing an accelerometer (Actical).
- **Self-reported** participant assessment (Global Physical Activity Questionnaire, GPAQ) of their physical activity levels.

The protocol for the **accelerometer component** is described in Chapter 18 of Manual of Operations 2 (Field Center Procedures). Briefly, participants were asked to wear an accelerometer (Actical) for 7 days to assess the frequency, duration, and intensity of their physical activity during that time period. Field center staff gave participants the accelerometers

after all physical examinations during the clinic visit were completed. Participants were fitted with a belt with the appropriate size waist and instructed to wear the accelerometer above the iliac crest, the location most sensitive to vertical movements consistent with ambulation. Participants were told to undertake their usual activities for 7 days while wearing the monitor (see section 3 for further description), and to remove it only for swimming, showering, and sleeping. They were also provided written instructions and a phone number to call if any questions arose and were called mid-week to answer any questions about the device, to make sure the instructions were clear, and to remind them to wear the monitor. Participants left the clinic visit wearing the accelerometer. The monitor was returned to the field center using a pre-paid envelope, where the data were downloaded and the monitor was re-initialized for reuse.

The **self-report assessment** of participant's physical activity during a typical week was performed during the interview portion of the clinic visit. A modified version of the World Health Organization (WHO) GPAQ (found here: http://www.who.int/chp/steps/GPAQ/en/index.html) was used to obtain information about participants' habitual activities in three domains: work-related (both moderate and vigorous levels), transportation (moderate level), and leisure or recreational (both moderate and vigorous levels). It also included a question about how much time the participant spent sitting or reclining on an average day.

The dataset PAEA has the self-report physical activity data. For dataset and variable naming conventions see "HCHS/SOL Investigator Use Database Overview". The following changes were made from the WHO original GPAQ for use in HCHS/SOL. Lead-in text to the questionnaire and major questionnaire sections differed from the original.

Work:

- (original #3, PAEA3) essentially the same question but stated in a longer format
- (original #6, PAEA6) essentially the same question but stated in a longer format

Transport:

- (original #8, PAEA8) "for at least 10 minutes continuously" was omitted
- (original #9, PAEA9) question was worded slightly differently. It originally stated "How much time do you spend walking or bicycling for travel on a typical day?"

Recreation/Leisure:

- (PAEA11) new item that was not asked originally that quantifies the types of vigorous activities that are done
- (PAEA15) new item that was not asked originally that quantifies the types of moderate activities that are done
- (original #15, PAEA17) essentially the same question but stated in a longer format

Sitting:

- No changes

A trained interviewer conducted the interview and obtained the physical activity information using clarification and prompts within the form. Interviewers consistently reminded participants during the interview that activities should only be reported if the duration was at least ten minutes. Interviewers helped participants recall their activities during a typical week by asking them to list their activities during the previous week and then clarifying whether that was a typical week.

2. Physical Activity (PA) Datasets

Table 2.1 describes two derived datasets (PA_CNTS and PA_DERV) available for objectively measured physical activity (counts/min from Actical), one dataset (PAEA) for self-reported physical activity assessed with the modified GPAQ, and derived variables for QPAQ available in the main participant derived dataset (PART_DERV). These GPAQ derived variables are described in Section 11 of the "HCHS/SOL Derived Variable Dictionary" document. In contrast, variables from the Actical data are defined in the document "Data Dictionary for Objectively Measured Physical Activity".

Table 2.1. SAS Datasets for Physical Activity.

Assessment	HCHS/SOL SAS Dataset	Key fields (identify each record uniquely)		Dataset description
Objective physical activity assessment (Actical)	PA_CNTS	ID + DAY	Participant ID + Consecutive day of Actical wear	- Multiple records per participant; one record per worn DAY per ID. - Each ID has between 1 and 6 records, one for each day they wore the Actical. Day 1 is the next day after the clinic visit day. Data collected on the partial day of the clinic visit are not included in the file (see Figure 3.1). - Dataset has the original activity counts (variables CNT1 to CNT1440) corresponding to each minute (epoch length) of a calendar day where CNT1 corresponds to activity counts at midnight (12:00am) and CNT1440 corresponds to activity counts at 11:59pm. For day 1, all counts are set to special missing (.L) from midnight to 5 am (see Table 3.1). - There are 89,574 records (days).
	PA_DERV	ID	Participant ID	- Actical data (average counts/min) at the participant level (N=16,415). - Dataset includes those participants with no available actical (N=1,502) and those with less than three adherent days (N=2,163), but all summary variables are missing. - ACTICALYN is an indicator variable that identifies participants with at least some actical data (N=14,913) - ADHERENTYN is an indicator variable that identifies participants with at least three adherent days (N=12,750). Another indicator variable identifies whether a weekend day was included or not. - Adherence is defined in section 3 of this document.
Self-report physical activity assessment (GPAQ)	PAEA	ID	Participant ID	- Participants' self-assessment of their physical activity (N=16,335) (i.e. the GPAQ in three domains: work, transportation, and recreation/leisure, plus sitting)
	PART_DERV	ID	Participant ID	- HCHS/SOL participant derived variables include some derived variables from GPAQ

PA refers to Physical Activity, PART to participant, and DERV to derived.

3. Methods to Clean Actical Data and to Derive Variables

The HCHS/SOL protocol stated that the accelerometer (Actical) should be **initialized (start recording the counts)** at 5am on the day of the clinic visit and participants should wear it for 7 days (Chapter 18 of Manual of Operations 2). Epoch length was to be set to 1 minute and the step function was to be enabled. However, in practice to accommodate site-specific challenges (no show-ups and Acticals already initialized), several sites modified the protocol regarding date and start time. In general,

- Chicago set up the monitor (date and start time) while the participant was in the clinic. They told participants to wear the monitor for 7 days including the day of the clinic visit.

- The Bronx and Miami sites set up the monitors on the day of the clinic visit with the start time set-up for 5am the next day. They told participants to wear the monitor for 7 days starting the next day.

- San Diego set up the monitors with the start date set to the day of the clinic visit and 5:00am on that same day. They told participants to wear the monitor for 7 days starting the day of the clinic visit.

Hence, to clean the data, the Coordinating Center and a Technical Physical Activity Subcommittee decided to standardize the data by specifying DAY 1 as the next day of the clinic visit and keeping 6 days at the most (Table 3.1). Clinic visit days, most of the time, were partial days and did not represent a typical day. This day was not intended to be kept and by excluding it, we were consistent across all four sites. For Chicago and San Diego, the times from midnight to 5:00am the day after the clinic visit were set to missing to match the other two sites. Day 1 for all participants is thus the day after the clinic visit, with counts starting at 5:00 am. Similarly, the last day for Bronx and Miami was dropped to standardize across sites, and therefore a maximum of 6 days of data is available for all participants. Also, participant data (N=304) was excluded from PA_CNTS if the start day of the Actical was not within +/- 1 day of the clinic visit, since we are not certain that the Actical data recorded was for that participant.

Table 3.1. Start date, start time and requested days to wear the Actical by site.

	Day 0 (Clinic Visit)	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
Bronx		X (5am)	X	X	X	X	X	X
Chicago	X (clinic visit time)	X	X	X	X	X	X	
Miami		X (5am)	X	X	X	X	X	X
San Diego	X (5am)	X	X	X	X	X	X	

Shaded columns (days) are kept in PA_CNTS. So, for Chicago and San Diego, data collected prior to 5 am on Day 1 is excluded from PA_CNTS.

Non-wear time was determined using the Choi algorithm (Choi et al., 2011), defined as at least 90 consecutive minutes of zero counts, with allowance of 1 or 2 minutes of nonzero counts if no counts were detected in a 30 minute window upstream and downstream of the 90 minute

period. After this cleaning procedure, an adherent day for inclusion was defined as at least 10 hours of wear time. To summarize data at the participant level, at least 3 adherent days of accelerometer data were required. The PA_DERV dataset has a variable called "wkendday_included" which indicates whether a weekend day was included in the summarized data, in case investigators wish to conduct sensitivity analyses to further exclude participants that did not provide a weekend day.

Intensity of activity was defined using the following intensity level cutpoints (Colley et al., 2011; Wong et al., 2011):
Sedentary: <100 counts/minute
Light activity: 100-1534 counts/minute
Moderate activity: 1535-3961 counts/minute
Vigorous activity: >=3962 counts/minute

Data cleaning for "spurious" data

We used Colley et al (2010) to guide us in identifying problematic observations. Colley et al identified potentially spurious data based on the Actical as any minute with counts/min above 20,000. We did not find any counts/min above 20,000. Next, we explored any non-zero count that was sustained for more than 10 consecutive minutes. We identified 124 participants with at least one instance of non-zero count/minute that was sustained for more than 10 minutes. However, all counts/min were below 200 and most were below 100, which is the threshold for sedentary behavior. According to the manufacturer (via word of mouth from their technical department), due to the nature of the Actical processing, counts below 100 are not as precise, often recorded using only a few unique counts that appear repetitively rather than being truly continuous. This phenomenon can lead to sustained repetitions of the same count that would not be considered spurious. After close examination of these 124 cases, we only excluded 3 participants for which the same few counts/min were repeated for a very large percentage of the day for all their adherent days. In addition, we excluded 5 participants that did not have a single minute of sedentary behavior in any of their 6 adherent days. We did not exclude 4 participants with no sedentary time on one day only. In summary, 92% (15,153 participants out of 16,415) returned the Actical (Table 3.2), 78% (n=12,750 out of 16,415) had at least three adherent days, and hence summarized physical activity data.

Table 3.2. Sample size due to not wearing the Actical, data cleaning, or not enough days to summarize

Description of Exclusion	**Excluded**	**Included**		
	N	N	%	%
HCHS/SOL Cohort	0	16,415	100	
No Actical	1,262	15,153	92	100
Difference between clinic date and Actical date >1 or missing	232	14,921	91	98
No sedentary time on all 6 days	5	14,916	91	98
Same count/min sustained repeatedly	3	14,913	91	98
<3 adherent days	2,163	12,750	78	84
Depending on the manuscript, investigators may wish to require that the accelerometer was worn at least one weekend day. This exclusion is optional and is manuscript-dependent. The exclusion was not applied to the data release.				
≥ 3 adherent days but no weekend day	3,585	11,328	69	74

4. Variables available from the Actical and Global Physical Activity Questionnaire (GPAQ)

Table 4.1 shows the variables available from Actical accelerometer and from self-reported physical activity (GPAQ); some are available from both sources. Due to the format of the GPAQ questions it was not possible to construct derived BOUT variables for the self-reported data; however, the GPAQ data does offer information on the type of physical activity (work, recreation, transport) that the Actical does not provide. Both datasets provide information on the intensity and duration of the participant's physical activity as well as whether the participant's level of physical activity met the 2008 Physical Activity Guidelines for Americans (http://www.health.gov/paguidelines/).

Table 4.1. Relationship of the variables available from Actical accelerometer and the self-reported physical activity (GPAQ)

Variable description	Units	Actical		Self-reported physical activity (GPAQ)
		PA_CNTS (by ID and DAY)	PA_DERV (by ID)[1]	PART_DERV
Counts per minute	count/min	CNTS_MIN	CNTS_MIN _DAY	
Hours monitor was worn	Hours	TOTHRS	HRS_DAY	
Adherent (≥ 10 hrs/day) (1=Yes, 0=No)	Yes/No	ADHERENT	ADHERENTYN	
Sedentary behavior (<100 counts/min)	min/day	TOTSED	SED_DAY	
Light activity (100-1534 counts/min)	min/day	TOTLIGHT	LIGHT_DAY	
Moderate activity (1535 - 3961 counts/min)	min/day	TOTMOD	MOD_DAY	GPAQ_TOTAL_MOD, GPAQ_REC_MOD, GPAQ_WORK_MOD
	min/week		MOD_WEEK	GPAQ_MOD_WEEK
Vigorous activity (>=3962 counts/min)	min/day	TOTVIG	VIG_DAY	GPAQ_TOTAL_VIG, GPAQ_REC_VIG, GPAQ_WORK_VIG
	min/week		VIG_WEEK	GPAQ_VIG_WEEK
Moderate or vigorous activity (>=1535 counts/min)	min/day	TOTMV	MV_DAY	
	min/week		MV_WEEK	GPAQ_MV_WEEK
Activity level per 2008 PA guidelines (categorical - 4 levels)			PAG2008	
Meets 2008 activity level guidelines (1=Yes, 0=No)	Yes/No		PAG2008YN	
Moderate activity in bouts[2]	min/day	TOTMOD_BOUT	MOD_DAY_BOUT	
	min/week		MOD_WEEK_BOUT	
Vigorous activity in bouts	min/day	TOTVIG_BOUT	VIG_DAY_BOUT	
	min/week		VIG_WEEK_BOUT	

Variable description	Units	Actical		Self-reported physical activity (GPAQ)
		PA_CNTS (by ID and DAY)	PA_DERV (by ID)[1]	PART_DERV
Moderate or vigorous activity in bouts	min/day	TOTMV_BOUT	MV_DAY_BOUT	
	min/week		MV_WEEK_BOUT	
Activity level per 2008 PA guidelines from bouts only (categorical - 4 levels)			PAG2008_BOUT	
Meets 2008 activity level guidelines from bouts only (1=Yes, 0=No)	Yes/No		PAG2008YN_BOUT	

1 Variables for counts and minutes are the average across adherent days.
2 Bouts defined as durations of at least 10 minutes in which 8 out of each 10 rolling minutes meets the activity criteria.
Blank cells mean there are no corresponding variables available.

5. Methods Section for Objectively Measured Physical Activity (Actical) Suggested for Publications

Investigators are encouraged to paraphrase for use in manuscripts the following technical description. Shortening may also be necessary, depending on the paper.

During the HCHS/SOL baseline clinic visit, participants were asked to wear an Actical B-1 version (model 198-0200-03; Respironics Co. Inc., Bend, Oregon) accelerometer for one week. The Actical is an omnidirectional accelerometer measuring 1.14'' x 1.45'' x 0.43'', weighing 16 grams, and powered by a CR2025 lithium coin cell battery. The device has 32MB of non-volatile flash memory, a sampling rate of 32 Hz, a dynamic range of 0.05 to 2.0G, and bandwidth of 0.035 Hz to 3.5 Hz. Prior studies have shown the Actical to have acceptable technical reliability for counts (Esliger and Tremblay, 2006; Welk et al., 2004) and steps (Esliger et al., 2007).

Participants were fitted with a belt with the appropriate size waist and instructed to wear the accelerometer above the iliac crest, the location most sensitive to vertical movements consistent with ambulation. Participants were told to undertake usual activities for 7 days while wearing the monitor, and to remove it only for swimming, showering, and sleeping. They were also provided written instructions and a phone number to call if any questions arose and were called mid-week to answer any questions about the device, to make sure the instructions were clear, and to remind them to wear the monitor. Participants left the clinic visit wearing the accelerometer. The monitor was returned field center using a padded pre-paid envelope, where the data were downloaded and the monitor was reinitialized for reuse.

The Actical was programmed to capture accelerations in counts and steps in one-minute epochs. The four sites programmed the monitor to start between 5:00am of the clinic visit day to 5:00am of the day following the clinic visit. To standardize, we included time for all sites beginning at 5:00am of the morning following the clinic visit and kept a maximum of six days. Non-wear time was determined using the Choi algorithm (Choi et al., 2011), defined as at least 90 consecutive minutes of zero counts, with allowance of 1 or 2 minutes of nonzero counts if no counts were detected in a 30 minute window upstream and downstream of the 90 minute period. An adherent day for inclusion was defined as at least 10 hours of wear time.

To be included in the analysis, each participant needed to contribute at least 3 adherent days of accelerometer data. The reliability of 3 or more adherent days was 0.88 for counts/day and 0.82 for total hours of wearing, using the Spearman Brown prophecy formula (Table 5.1).

Table 5.1. Reliability of three or more adherent days for counts/day and total hours of wearing.

Outcome	6 days	5 days	4 days	3 days	2 days	1 day
Counts/day	0.94	0.93	0.91	0.88	0.83	0.71
Total hours	0.90	0.89	0.86	0.82	0.76	0.61

The raw data were summarized using average counts/minute. Intensity of activity was defined as follows (Colley et al., 2011; Wong et al., 2011): sedentary <100 counts/minute, light activity 100-1534 counts/minute, moderate activity 1535-3961 counts/minute, and vigorous activity >=3962 counts/minute.

6. Actical outcomes missingness

In HCHS/SOL baseline data, there is a high percentage of missing Actical (accelerometer) data (22.3%) due to either participants not returning the accelerometer (N=1262, 7.8%) or participants not being compliant (having less than 3 of 6 adherent days; N=2403, 14.6%). Further, there are differences between participants who have Actical data and those who do not have it, as reported in manuscript 213 “Accelerometer Adherence and Performance in a Population-based Cohort Study of US Hispanic Adults” (Evenson K, Sotres-Alvarez D, *et al*). Hence, it is necessary for all analyses of Actical data to address the missingness.

The HCHS/SOL Coordinating Center (CC) recommends using inverse probability weighting (IPW) to adjust for the missing Actical data. The HCHS/SOL PA_DERV dataset (released May 2014) contains the Actical IPW weight (WEIGHT_PA_IPW_OVERALL) for each participant. Adjusted analyses are performed by using the Actical IPW weight instead of the standard HCHS/SOL sampling weight, combined with a subpopulation statement to accurately limit analysis to the compliant participants. This section describes the development of the Actical IPW weights in detail and provides sample SAS and SUDAAN code, as well as suggested text for statistical methods sections of manuscripts.

As of May 2014, this method is currently in use for manuscripts 109 “Distribution of Self-report and Objectively Measured Physical Activity” (Arredondo E, Sotres-Alvarez D, *et al*) and 57 “Self-reported and Objectively Measured Physical Activity in Association with Body Mass Index” (Davis SM, Gouskova N, *et al*).

6.1. Actical IPW derivation

Inverse probability weighting (IPW) is one method which allows correcting for the bias of the estimates obtained by complete-case analyses and can be implemented for complex survey designs. The complete cases (i.e., 3 or more adherent days) are weighted by the inverse of their probability of being a complete case. For further background, Seaman and White (2013) provide a review of the implementation and advantages and disadvantages of using IPW to handle missing data in epidemiological research. The method has been implemented for physical activity (Actical) data in HCHS/SOL due to the ease with which it can be applied to complex survey data.

The probability of being a case (i.e., having 3 or more adherent days of Actical wear) is calculated by fitting a logistic regression model where the outcome is the binary variable which takes value 1 if the participant has three or more adherent days and 0 otherwise. The model is fitted without weighting by the HCHS/SOL sampling weight because what we want to predict is the conditional probability of having 3 or more adherent days, given inclusion into the HCHS/SOL sample. We then use the predicted probability of 3 or more adherent days to compute an inverse probability weight for each participant. This additional weight is combined with HCHS/SOL sampling weight into a single weight for analyses involving objectively measured physical activity (Actical) data. The 16 covariates that we included in the logistic regression model to predict the probability that the participant has 3 or more adherent days are shown in Table 6.1.

Since for those participants who have missing covariate values (5.6%, N=917 considering all 16 covariates simultaneously), we cannot compute predicted probabilities to have complete Actical data based on the fitted model, we first imputed missing covariates using multiple imputation (MI). Note that although 5.6% of all study participants had one or more missing covariates, the percentage of individual data items which we needed to impute is only 0.63% (N=1,693) of the total number of data items (16,415*16; see table 1 for more detail). Hence, in step 1 we created 5 datasets with imputed values and in step 2 we fit the logistic regression model using each of the imputed datasets to generate the regression coefficients. Finally, we averaged the 5 linear predictors from the imputed datasets to obtain a single IPW weight. Details of steps 1 and 2 are described below.

Table 6.1. Percent of missing covariate data and scale

Covariate	Description	Scale	Missing data	
			N	%
GENDER	Gender	Binary	0	0
AGE	Age	Continuous	0	0
INCOME_C3_nomiss	Missing were combined with refused creating 3-level nominal variable: <$30K, ≥$30K and unknown	Nominal	0	0
LANG_PREF	Language of preference	Binary	0	0
WEIGHT_FINAL_NORM_OVERALL	Sampling weight	Continuous	0	0
STRAT	Stratification	Nominal	0	0
BMI	Body mass index	Continuous	71	0.4
SITE_BKGRD	17-level cross-classification of center and Hispanic/Latino background	Nominal	87	0.5
MARITAL_STATUS	3-level marital status	Nominal	88	0.5
EDUCATION_C3	3-level education level	Ordinal	91	0.6
IMGEN_C2	Immigrant generation (first, 2nd or higher)	Binary	94	0.6
GPAQ_REC	Recreational physical activity (min/day)	Continuous	146	0.9
GPAQ_TRSPORT	Transportation-related physical activity (min/day)	Continuous	176	1.1
GPAQ_WORK	Work-related physical activity (min/day)	Continuous	395	2.4
AGG_PHYS	Aggregate Physical Health Scale	Continuous	239	1.5
EMPLOYED	4-level employment status	Nominal	306	1.9

Step 1. Multiple imputation for missing covariates

We generated 5 imputed datasets which is an adequate number when there is little missing data. Because the income variable had 9.1% of missing data due to participants refusing to provide their income, we created a special response category for missing income and did not impute any specific level of income. We used the SAS MI procedure with the fully conditional specification (FCS) method which only assumes the existence of a joint distribution for all

variables (van Buuren, 2007). We included the strata and the sampling weight as covariates into the imputation model as a way to account for the complex survey design. The primary sampling unit (PSU) was not included because this would lead to convergence problems. Imputation models for specific types of covariates were:

- Logistic regression was used for binary or ordinal variables: imgen_c2, education_c3, and income_C3_nomiss;
- Discriminant method was used for nominal variables: site_bkgrd, marital_status, and employed;
- Linear regression was used for continuous variables: age, bmi gpaq_rec, gpaq_work, gpaq_trsport, agg_phys, and weight_final_norm_overall.

Step 2. Calculate IPW

For each of the 5 imputed datasets (each with N=16,415) from step 1 we used logistic regression to model the probability of being a case (defined as having 3 or more adherent days of actical wear, each with at least 10 hours of wear time). We included all the covariates in table 1 plus all pairwise interactions, except for STRATA and IMGEN_C2 which were used as main effects only to avoid logistic regression convergence problems due to some sparse cells. Then we:

1. computed the 5 linear predictors (X*beta)
2. averaged them to obtain a single mean linear predictor (mxb)
3. calculated the probability of being adherent: $PROB\_ADHERENCE = \frac{\exp(mxb)}{1+\exp(mxb)}$
4. calculated the IPW: $\text{IPW\_ADHERENCE} = \frac{1}{PROB\_ADHERENCE}$
5. combined the IPW and the HCHS/SOL sampling weight:
   $\text{PA\_IPW} = \text{WEIGHT\_FINAL\_NORM\_OVERALL} * \text{IPW\_ADHERENCE}$

Numerous covariate models were evaluated for optimal fit and convergence. The final model had fairly good prediction properties as measured by the C-statistic of 0.71 averaged across the 5 imputations (0.5 is the same as chance, 1.0 indicates perfect prediction).

Two other methods for handling missing Actical data were evaluated by HCHS/SOL Coordinating Center, and the provided IPW weight was determined to be the most optimal and easy to implement among those investigated. One option was to use IPW as described to adjust for participants who did not returned the Actical, and to use MI to impute Actical outcomes using available partial Actical data for participants with one or two adherent days (N=1,511). The other option was to only use MI and impute Actical outcomes for all study participants with missing Actical data (MI). These two methods have disadvantages in implementation versus the IPW as follows: (1) require more steps for investigators on subsequent data analyses as they have to summarize results from the 5 imputed datasets and (2) must be completed for each Actical outcome separately, while a single IPW weight can be used for all Actical outcomes. An evaluation comparing these two methods to the proposed IPW method for two selected Actical outcomes (min/day in moderate or vigorous activity and meeting 2008 PA Guidelines) showed that the overall estimates from all three methods were very similar (data not shown). Therefore, the CC is providing an IPW weight as a tool to account for missing Actical outcome data.

6.2. Statistical implementation

All 16,415 participants should be included in the analytic dataset and complex survey procedures should be used. Instead of the usual HCHS/SOL sampling weight (WEIGHT_FINAL_NORM_OVERALL), the WEIGHT_PA_IPW_OVERALL variable should be used. Participants who have at least 3 adherent days (N=12,750) should be treated as a subpopulation and selected by using the SUBPOPN (SUDAAN) or DOMAIN (SAS) statements. Note that although only the weights for the compliant subpopulation are relevant in the analysis, the provided WEIGHT_PA_IPW_OVERALL variable is non-missing and strictly greater than zero for all 16,415 participants, as required by the software.

Example SUDAAN code. Age-sex-adjusted mean of min/day of moderate or vigorous activity.

```
proc sort data=work.analysis; by strat PSU_ID; run;
proc regress data=work.PA_DERV filetype=sas design=wr;
    nest strat PSU_ID;
    subpopn adherentYN=1;
    class BKGRD1_C7;
    weight WEIGHT_PA_IPW_OVERALL;
    model MV_day = BKGRD1_C7 gendernum age;
    condmarg BKGRD1_C7;
run;
```

```
DESIGN SUMMARY: Variances will be computed using the Taylor Linearization Method, Assuming a With
Replacement (WR) Design
    Sample Weight: WEIGHT_PA_OVERALL
    Stratification Variables(s): STRAT
    Primary Sampling Unit: PSU_ID

Number of observations read       :  16415    Weighted count:    23413
Observations in subpopulation     :  12750    Weighted count:    16361
Observations used in the analysis :  12724    Weighted count:    16325
Denominator degrees of freedom    :    644

Link Function: Identity
Response variable MV_DAY: min/day (average) of moderate or vigorous activity (>=1535 counts/min)
For Subpopulation: ADHERENTYN = 1
```

Conditional Marginal #1	Conditional Marginal	SE	Lower 95% Limit	Upper 95% Limit	T:Marg=0	P-value
7-level re-classification of Hispanic/Latino Background						
0	31.01	1.24	28.57	33.45	24.93	0.0000
1	24.22	1.08	22.10	26.33	22.47	0.0000
2	15.82	0.70	14.45	17.18	22.74	0.0000
3	24.36	0.69	23.00	25.72	35.23	0.0000
4	32.91	1.56	29.85	35.97	21.10	0.0000
5	26.56	1.61	23.41	29.71	16.55	0.0000
6	27.19	2.03	23.20	31.18	13.38	0.0000

Example SAS code. Age-sex-adjusted mean of min/day of moderate or vigorous activity.

```
proc sort data=work.analysis; by strat PSU_ID; run;
proc regress data=work.analysis filetype=sas design=wr;
    nest strat PSU_ID;
    subpopn adherentYN=1;
    class BKGRD1_C7;
    weight WEIGHT_PA_IPW_OVERALL;
    model MV_day = BKGRD1_C7 gendernum age;
    condmarg BKGRD1_C7;
run;
```

Domain Summary	
Number of Observations	12724
Number of Observations in Domain	12724
Number of Observations Not in Domain	0
Sum of Weights in Domain	16324.6
Weighted Mean of MV_DAY	24.89197
Weighted Sum of MV_DAY	406350.3

BKGRD1_C7 Least Squares Means

7-level re-classification of Hispanic/Latino Background	*GENDERNUM*	*AGE*	*Estimate*	*Standard Error*	*DF*	*t Value*	*Pr > |t|*	*Alpha*	*Lower*	*Upper*
0	0.48	41.07	31.0087	1.2440	641	24.93	<.0001	0.05	28.5658	33.4515
1	0.48	41.07	24.2177	1.0780	641	22.46	<.0001	0.05	22.1008	26.3346
2	0.48	41.07	15.8179	0.6959	641	22.73	<.0001	0.05	14.4514	17.1844
3	0.48	41.07	24.3591	0.6915	641	35.22	<.0001	0.05	23.0012	25.7171
4	0.48	41.07	32.9119	1.5599	641	21.10	<.0001	0.05	29.8487	35.9750
5	0.48	41.07	26.5606	1.6056	641	16.54	<.0001	0.05	23.4078	29.7134
6	0.48	41.07	27.1915	2.0333	641	13.37	<.0001	0.05	23.1987	31.1842

Finally, data users are reminded that manuscripts should include the strategy used to take into account the complex sample design used to select participants in HCHS/SOL. In particular, statistical methods to incorporate stratification factors, sampling weights, and clustering in the data analysis should be described (see "Analysis Methods for HCHS/SOL").

6.3. Suggested wording for methods section when using IPW

Of the 16,415 HCHS/SOL participants, 92.3% (N=15,153) returned the accelerometer, and 77.7% (N=12,750) provided at least 3 adherent days with 10 or more hours of wear time (Evenson et al, in submission). Due to this large amount of missing Actical outcomes, analyses were adjusted for missing data using inverse probability weighting (IPW) (add reference here for the first HCHS/SOL paper to be published, 2014; Seaman and White, 2011). An IPW weight was created from a logistic regression model predicting Actical compliance based on age, gender, income level, marital status, education, employment status, language preference, immigrant generation, self-reported physical activity from GPAQ, BMI, aggregate physical health score, field center by background cross-classification, sampling stratum, and sampling weight. Because some participants (N=917, 5.6%) also had sporadic missing data for one or more covariates in the IPW model, the missing covariates were first imputed by multiple imputation. Finally, the weight used in the analyses of Actical outcomes is the product of the IPW weight (to weight the results for the compliant subset back to the whole HCHS/SOL sample) and the HCHS/SOL sampling weight (to further weight the results back to the Hispanic/Latino population in the target areas).

REFERENCES

Cited and Additional Helpful Resources

Carlson SA, Fulton JE, Galuska DA, and Kruger J. Prevalence of self-reported physically active adults – United States, 2007. *MMWR* 2008;57(48), 1297-1300.

Choi, L., Liu, Z., Matthews, C.E., Buchowski, M.S., 2011. Validation of accelerometer wear and nonwear time classification algorithm. *Med Sci Sports Exerc* 43, 357-364.

Colley, R., Garriguet, D., Janssen, I., Craig, C., Clarke, J., Tremblay, M., 2011. Physical activity of Canadian adults: Accelerometer results from the 2007 to 2009 Canadian Health Measures Survey. Statistics Canada no. 82-003-XPE, Health Reports 22, 7-14.

Colley, R., Connor Gorbor, S., Tremblay, M. 2010. Quality control and data reduction procedures for accelerometry-derived measures of physical activity. Health Reports 21(1), 63-69.

Esliger, D., Tremblay, M., 2006. Technical reliability assessment of three accelerometer models in a mechanical setup. Med Sci Sports Exerc 38, 2173-2181.

Esliger DW, Probert A, Gorber SC, Bryan S, Laviolette M, Tremblay MS. 2007. Validity of the Actical accelerometer step-count function. Med Sci Sports Exerc 39(7):1200-1204.

Troiano RP, Berrigan D, Dodd KW, Mâsse LC, Tilert T, McDowell M. Physical activity in the United States measured by accelerometer. Med Sci Sports Exerc. 2008 Jan;40(1):181-8.

Tudor-Locke C, Camhi SM, Troiano RP. A catalog of rules, variables, and definitions applied to accelerometer data in the National Health and Nutrition Examination Survey, 2003-2006. Prev Chronic Dis. 2012; 9:E113.

2008 US Physical Activity Guidelines for Americans:
http://www.health.gov/paguidelines/guidelines/default.aspx

Wong, S.L., Colley, R., Connor Gorber, S., Tremblay, M., 2011. Actical accelerometer sedentary activity thresholds for adults. J Physical Activity Health 8, 587-591.

World Health Organization (WHO) Global Physical Activity Questionnaire (GPAQ)
http://www.who.int/chp/steps/GPAQ/en/index.html

Inverse probability weight (IPW)

Seaman, S. R., & White, I. R. (2011). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, *22*(3), 278-295.

Missing Data Background References

Alhassan S, Sirard JR, Spencer TR, Varady A, Robinson TN. Estimating physical activity from incomplete accelerometer data in field studies. J Phys Act Health. 2008;5 Suppl 1:S112-25.

Catellier DJ, Hannan PJ, Murray DM, Addy CL, Conway TL, Yang S, Rice JC. Imputation of missing data when measuring physical activity by accelerometry. Med Sci Sports Exerc. 2005 Nov;37(11 Suppl):S555-62.

Kang M, Rowe DA, Barreira TV, Robinson TS, Mahar MT. Individual information-centered approach for handling physical activity missing data. Res Q Exerc Sport. 2009 Jun;80(2):131-7.

Lee PH. Data imputation for accelerometer-measured physical activity: the combined approach. Am J Clin Nutr. 2013 May;97(5):965-71.

Morris JS, Arroyo C, Coull BA, Ryan LM, Herrick R, Gortmaker SL. Using Wavelet-Based Functional Mixed Models to Characterize Population Heterogeneity in Accelerometer Profiles: A Case Study. J Am Stat Assoc. 2006 Dec 1;101(476):1352-1364.

van Buuren, S. (2007), "Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification," Statistical Methods in Medical Research, 16, 219–242.

HCHS/SOL Documents

- Manual of Operations 2: Field Center Procedures
- HCHS/SOL Investigator Use Database Overview
- HCHS/SOL Derived Variable Dictionary
- Analysis Methods for HCHS/SOL
- Data Dictionary for Objectively Measured Physical Activity

HCHS/SOL Manuscripts in Progress

MS57
Davis SM, Gouskova N, and *[alphabetically: Carnethon M, Castañeda SF, Evenson KR, Gellman MD, Hankinson AL, Isasi CR, McMurray RG, Schneiderman N, Sotres-Alvarez D, Talavera G]* "Self-reported and Objectively Measured Physical Activity in Association with Body Mass Index").

MS109
Arredondo E, Sotres-Alvarez D, and *[alphabetically: Carnethon M, Castañeda SF, Crespo NC, Daviglus ML, Davis SM, Espinoza RA, Evenson KR, Isasi CR, Marshall SJ, Stoutenberg]* "Distribution of Self-report and Objectively Measured Physical Activity"

MS213
Evenson KR, Sotres-Alvarez D, Deng Y, Marshall SJ, Isasi CR, Esliger D, and Davis, SM. "Accelerometer Adherence and Performance in a Population-based Cohort Study of US Hispanic Adults"