Journal of Physical Activity and Health, 2012, 9(Suppl 1), S56-S67

Modeling Errors in Physical Activity Recall Data

Sarah M. Nusser, Nicholas K. Beyler, Gregory J. Welk, Alicia L. Carriquiry, Wayne A. Fuller, and Benjamin M.N. King

Background: Physical activity recall instruments provide an inexpensive method of collecting physical activity patterns on a sample of individuals, but they are subject to systematic and random measurement error. Statistical models can be used to estimate measurement error in activity recalls and provide more accurate estimates of usual activity parameters for a population. ***Methods:*** We develop a measurement error model for a short-term activity recall that describes the relationship between the recall and an individual's usual activity over a long period of time. The model includes terms for systematic and random measurement errors. To estimate model parameters, the design should include replicate observations of a concurrent activity recall and an objective monitor measurement on a subsample of respondents. ***Results:*** We illustrate the approach with preliminary data from the Iowa Physical Activity Measurement Study. In this dataset, recalls tend to overestimate actual activity, and measurement errors greatly increase the variance of recalls relative to the person-to-person variation in usual activity. Statistical adjustments are used to remove bias and extraneous variation in estimating the usual activity distribution. ***Conclusions:*** Modeling measurement error in recall data can be used to provide more accurate estimates of long-term activity behavior.

Keywords: measurement error, recall bias, attenuation

Physical activity researchers are often interested in evaluating typical levels of physical activity behavior in populations. This information is needed for improved public health surveillance, for examining associations with health outcomes, and for evaluating effectiveness of behavioral interventions.[1] These types of goals are best served by considering long-term or habitual activity patterns for individuals.[2] We refer to an individual's long-run average activity level per unit of time (eg, day or week) as the individual's *usual activity*, where "activity" may refer to total energy expenditure, time spent at particular MET-levels, or some other activity metric per unit of time.

Unfortunately, it is impractical to directly estimate the usual activity of an individual. To do so would require accurately measuring activity levels for an individual over a long period of time. As a result, most studies turn to simpler measures, such as physical activity recall questionnaires that ask a respondent to provide a list of the activities that s/he has engaged in, along with information on the duration, frequency, and intensity of the physical activity. In some cases, the participant is asked to provide a global self-reported usual activity level over a long period of time. The appeal of self-reported usual activity levels is that they attempt to directly estimate the usual activity level for a person. However, researchers have noted the potential large error associated with such an approach.[3] An alternative method is to ask about very recent activity (eg, yesterday's activity) to reduce the burden and error associated with a long-term recall. The quality of the shorter recall is higher,[4] but recalled activity over a short period of time is not a good measure of long-term behavior for a person.

In this paper, we discuss how statistical models can be used to describe the relationship between a short-term activity measurement and the target concept of an individual's usual activity. We show how models describe the effect of various sources of measurement error in recall data, including systematic bias and variation associated with nuisance factors and measurement error. The approach sets the stage for generating estimates from the recall data that are adjusted for these errors. The resulting estimates offer more accurate descriptions of relationships between health outcomes and physical activity and of parameters such as the percentage of the population that fail to meet a physical activity threshold.

Researchers and policy makers in dietary assessment use a similar approach to estimate distributions of usual daily intake for nutrients and other food components. In the 1980s, Beaton and colleagues[5] noted that the day-to-day variation in short term dietary recall data swamps the person-to-person variation associated with usual intakes. Shortly thereafter, the National Academies issued a report recommending research into statistical methods to

Nusser, Carriquiry, and Fuller are with the Dept of Statistics, Iowa State University, Ames, IA. Beyler is with the Dept of Statistics, Mathematica Policy Research, Washington, D.C. Welk is with the Dept of Health and Human Performance, Iowa State University, Ames, IA. King is with the Human Biology Program, Stanford University, Stanford, CA.

adjust for nuisance effects in 24-hr dietary recall data.[6] In response to this call, the statistical methods were developed to estimate distributions of usual intakes for nutrients and foods.[7–10] Instead of asking for self-reported long-run intakes (eg, via a food frequency questionnaire), surveys such as the National Health and Nutrition Examination Survey (NHANES) obtain data from a large sample of individuals about their dietary intake in the last 24 hours. A smaller sample is asked to provide a second 24-hr intake on another day. A statistical model is used to describe an individual's 24-hr dietary recall in relation to the individual's usual intake and the measurement error associated with a single day's dietary recall. Data from the large sample of single 24-hr measurements plus the smaller sample of duplicate measurements are then used to estimate parameters in the model and ultimately the distribution of usual daily intakes for the food component. These methods are now used by the US to estimate the percent of population groups who have inadequate or excessive intakes.

An advantage in physical activity research is that more tools are available for modeling the errors in physical activity recall data than for dietary surveys. Doubly labeled water and calorimetry can be used to obtain accurate measures of energy expenditure,[1] although these methods are too expensive and burdensome for most population studies. A promising alternative for energy expenditure and other behavioral metrics can be found in relatively inexpensive activity monitors. Objective activity measurements can be made via accelerometers, multisensor monitors and other devices, providing a reference measure for actual activity levels.[1] While this doesn't eliminate error, it allows error components to be modeled in more detail, particularly in describing the potential bias in self-reported data.

Currently, the use of measurement error models in physical activity literature is limited. Exceptions include work by Ferrari and colleagues[11] and Spiegelman and colleagues.[12] One barrier to adopting these methods is a lack of familiarity with how models are constructed and used in practice to answer public health questions. The goal of this paper is to provide a conceptual understanding of how statistical models can express the error in an activity recall, and to provide insight into study designs that support estimation of parameters related to usual activity. Because this approach can be used for any type of activity metric, we are intentionally vague about the metric for activity for much of the discussion. However, we illustrate the approach by modeling total energy expenditure from a preliminary data set associated with a physical activity survey being conducted in Iowa.

We begin by describing classes of errors that occur when collecting data from a sample of respondents. We then construct a series of error models that account for errors that occur in an individual's short-term activity recall as it relates to the individual's usual activity. Designs that support estimation of error parameters and of usual activity distributions are discussed, which involve the use of a reference measure such as an activity monitor. Finally, we briefly illustrate these ideas with preliminary data on daily energy expenditure collected from adult women using a 24-hr activity recall and a multisensor activity monitor, and discuss the implications of this approach.

Modeling Errors

Classes of Errors in Survey Data

Survey methodology is a scientific field that develops methods for collecting information from human and other types of populations in a way that minimizes errors to the extent possible.[13] A survey or study that follows this approach results in a data set that can be used to create statistically valid estimates to describe the population of interest. One of the basic tenets of the field is to recognize and account for errors that occur in the survey data collection process. The survey process for human populations involves selecting a random sample (often called a "probability sample") of individuals or perhaps households, recruiting the randomly selected household or person in the household to participate in the study, obtaining specified measurements or answers to questions, editing and processing the data, and creating estimates from the final data set. The major error types that result from this process include coverage error (when the list from which the random sample is selected does not include the entire population of interest, and the omitted part of the population behaves differently from the rest of the population), sampling error (due to the fact that our estimates are based on a sample of the population, rather than all of the population), nonresponse error (when randomly selected persons or households don't participate and the nonrespondents behave differently from the rest of the sample), specification error (when the question or measurement doesn't directly measure on the concept of interest), measurement error (when the response provided or measurement taken is erroneous), and processing error (when the algorithms or procedures used to prepare the data for analyses contain errors).[13]

All of these types of errors occur in physical activity studies. We may, for example, fail to include more sedentary or less healthy individuals in our sampling list, which will result in biased estimates of health and activity behaviors (coverage error). Similarly, if nonrespondents are disproportionately inactive relative to the responding part of the sample, estimates will be biased (nonresponse bias). Sampling error is the main type of error discussed in statistics classes, and arises from the fact that we are making estimates for the population based on a random sample or subset of the population; that is, we do not expect the estimates to be exactly the same as the value we would get if we could observe all members of the population.

In this paper, we are primarily interested in errors that impact the measurement process. For example, the 24-hr recall does not measure the usual long-run average activity for an individual. This would be a form of

specification error in studies where habitual activity is of interest. We also know that respondents tend to overestimate good behavior and underestimate poor behavior,[14,15] inducing a form of measurement error. The recall process is also prone to errors of cognition and memory,[3,16] particularly for routine and sedentary activities that are not encoded in memory as discrete events.[17] Finally, even if there were no measurement error, data on time spent performing a specific activity is processed into a summary measure, such as total energy expenditure, MET–hrs, or time spent at a particular MET-level. This involves, for example, assigning a MET value to an activity via the Compendium of Physical Activity,[18] which offers a single value for all respondents that engage in that activity (processing error). In our modeling discussions, we will focus primarily on specification and measurement error, since a different kind of study would be needed to evaluate processing error.

Developing a Model for Activity Recall Data

In the following 4 parts of this section, we develop a statistical model that represents the relationship between an activity recall and the true usual activity for an individual. We focus specifically on modeling a 24-hr recall in relation to the usual daily activity for a person, although this approach can be applied to data from other forms of recall instruments. To motivate the model structure and notation, we build the model for a 24-hr recall sequentially, starting with a model that relates a hypothetical "error-free" 24-hr activity recall to an individual's usual daily activity. We extend this model to include systematic bias, and then introduce 2 types of random measurement errors in a 24-hr recall relative to the usual activity of a person. The implications of using a single 24-hr recall as a surrogate for usual daily activity are discussed before turning to study designs that will support estimation of this model.

A Simple Model for an Error-Free 24-hr Activity Recall

We start with an unrealistic model that assumes the 24-hr activity recall is error-free. In other words, the respondent always provides a perfect recall that equals the true activity level on that day. To relate the 24-hr recall to usual activity, we construct a model that recognizes that the true activity on a particular day is equal to the individual's usual daily activity plus a deviation that reflects the difference between that day's activity and the individual's usual daily activity.

We denote the 24-hr recall (R) for individual k on day j as R_{kj}, and the usual activity (U) for individual k as U_k. Recall that the usual activity is the average daily activity over a long time period, such as a year. For individual k, we expect that on some days, the activity level will be higher than the usual activity, while on other days, the individual will engage in less activity than her/his usual activity. We represent the deviation between the 24-hr activity recall for individual k on day j and individual k's usual intake as D_{kj}. We can express these concepts in a statistical model for the 24-hr recall for individual k on day j as follows:

$$R_{kj} = U_k + D_{kj}. \quad (1)$$

As we expand this model to include errors in recall data, it is useful to think of $U_k + D_{kj}$ as the true activity level for individual k on day j. A key idea in this model formulation is that the error-free 24-hr recall can be viewed as an *unbiased*, but "noisy" estimate of individual k's usual activity. By noisy, we mean that the recall for any one day will be an unreliable (or imprecise) estimate of the individual's usual activity.

While this model describes the relationship of a person's recall to her/his usual activity, we also want the model to capture our assumptions about how the usual activity level varies across individuals in the population of interest. If we were able to obtain the usual activity value for each individual in the population, we could average them to calculate the *mean usual activity for the population*, which we denote as μ_U. A population that is generally quite active will have a higher mean usual activity than a population whose members tend to be sedentary. Likewise, we describe the person-to-person variation in usual activity by the *variance of usual activity in the population*, denoted by σ_U^2. A population where individuals have consistent physical activity behaviors (eg, athletes in training) has smaller usual activity variance than, say, the general population of adults in the US, who would vary widely in activity levels from person to person. Since our primary interest is in describing the patterns of usual activity in a population, the usual activity mean and the person-to-person variance are important model parameters that we will estimate. The notation we add to the model to describe the population mean and variance for the usual activity levels in a population is $U_k \sim (\mu_U, \sigma_U^2)$, which is read as "the usual activity for individuals in the population has a mean of μ_U and a variance of σ_U^2."

The mean and variance of usual activity summarize 2 features of the *distribution of usual activity*, namely, the central tendency for persons in the population (mean) and the variability from person to person in usual activity (variance). The usual activity distribution is a more complete description of the usual activity patterns in a population. The distribution is often described using a function, such as the normal distribution (a bell-shaped curve centered at the mean) or the lognormal distribution (skewed to the right, indicating that a relatively small proportion of individuals have higher usual activity levels than the rest of the population). Other summary parameters of the usual activity distribution include the median usual activity (half of the population has usual activity levels below the median and half above) and the percentage of individuals in the population whose activity level falls below a threshold of healthy behavior. One goal in measurement error modeling is to provide an estimate of the usual activity distribution so that any summary of

the distribution can be estimated, including the mean, variance, median, and percentiles.

Returning to model (1), we also need to describe the distribution of the deviation of day j's activity from the usual activity for individual k, D_{kj}. Because usual activity is the long-run mean activity for individual k, we expect that on some days, the individual's activity level will be higher than her/his usual activity, while on other days, the individual will engage in less activity than her/his usual activity. Thus, for model (1), we assume that the mean of the daily deviations (actual activity on a day minus usual activity) for an individual is 0. We also assume that the variance of the deviations is the same for all individuals in this paper (more complicated assumptions can be made). That is, the deviations in the daily activity level from an individual's usual activity level are of the same magnitude as other individuals in the population. Using σ_D^2 to denote the variance of the deviations, we say that $D_{kj} \sim (0, \sigma_D^2)$.

Finally, we consider distribution assumptions for the 24-hr activity recalls. If we are willing to assume that the daily deviations are unrelated to usual activity (ie, statistically independent), then we can derive the mean and variance of the error-free 24-hr recall values for the model. Under these conditions, the 24-hr activity recalls have the same mean as the usual activity distribution, μ_U, and the variance of the 24-hr activity recalls is $\sigma_U^2 + \sigma_D^2$. These assumptions can be expressed as $R_{kj} \sim (\mu_U, \sigma_U^2 + \sigma_D^2)$.

Adding Systematic Measurement Error to the Model

Studies have shown that recall data may provide biased indicators of the individual's true activity on a particular day.[19,20] A linear model is a common method of expressing a systematic bias in a measurement. Under the assumptions of linear bias and no random measurement error, our model takes the form

$$R_{kj} = \beta_0 + \beta_1 (U_k + D_{kj}), \quad (2)$$

where β_0 is the intercept and β_1 is the slope for the linear bias, $U_k \sim (\mu_U, \sigma_U^2)$, $D_{kj} \sim (0, \sigma_D^2)$, and $R_{kj} \sim (\beta_0 + \beta_1\mu_U, \beta_1^2\sigma_U^2 + \beta_1^2\sigma_D^2)$ assuming U and D are independent. The intercept parameter represents an overall bias in reported daily activity that is present regardless of the true activity for that day. The slope parameter indicates how bias changes in relation to the true activity for that day. Many bias patterns are represented by model (2). The simplest case is when no systematic bias is present, which occurs when the intercept is 0 and the slope is 1. Figure 1 illustrates other bias patterns including underreporting of small activity values and overreporting of large values, general overreporting of activity with more bias for larger levels of activity, and overreporting of small activity values and underreporting of large values.

The systematic measurement error associated with model (2) results in a different distribution for the 24-hr activity recall R_{kj} than under model (1). The assumption of linear bias leads to a mean of the 24-hr recall distribution, $\beta_0 + \beta_1\mu_U$, that is biased for the true usual activity mean, μ_U (unless $\beta_0 = 0$ and $\beta_1 = 1$). In addition, the bias influences the variation in 24-hr activity recalls, $\beta_1^2\sigma_U^2 + \beta_1^2\sigma_D^2$.

Adding Random Measurement Error to the Model

Here, we consider 2 forms of random measurement error, one expressed as person-to-person variation reporting biases and the other expressed as remaining measurement error associated with the recall from a specific day.

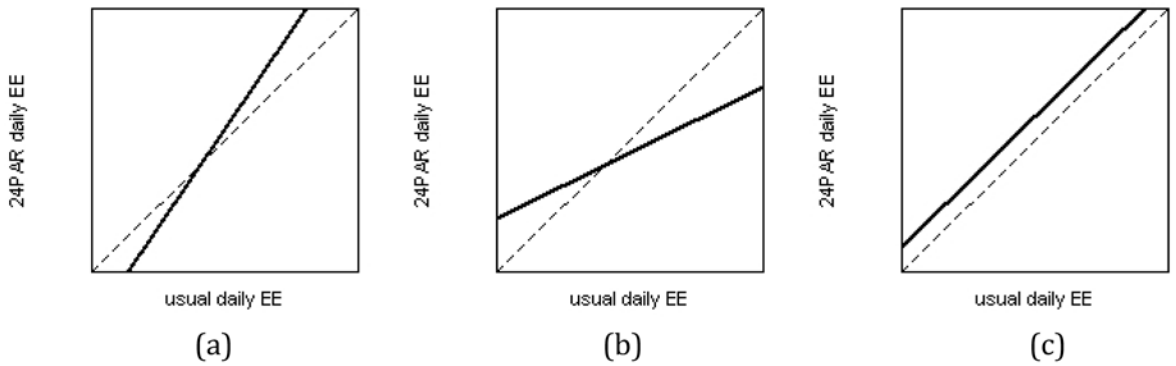

Figure 1 — Examples of bias patterns when the systematic bias in a 24-hr physical activity recall (PAR) is a linear function of the true activity on that day, using energy expenditure (EE) as the activity metric. The dashed line represents the case where the 24-hr recall accurately measures the true activity for that day, where the intercept is 0 and the slope is 1 in model (2). The solid line represents a linear bias in model (2) for the cases of (a) negative intercept and slope larger than 1, where lower activity levels are underreported and larger activity levels are overreported; (b) positive intercept and slope less than 1, where smaller activity levels are over reported and larger activity levels are underreported; and (c) positive intercept and slope equal to 1, where all activity levels are overreported by the same amount, regardless of usual activity level.

Model (2) implies that each individual's reporting bias is the same, which we know to be untrue. In fact, individuals vary in their misreporting tendencies. One person may be very aware of her/his activity level and offer a more accurate report relative to the general population bias from model (2). Another person may be susceptible to social desirability and overstate good behavior, such as time spent exercising. We use S_k to denote the deviation of an individual's average reporting bias relative to the average bias for the population expressed in model (2). The person-to-person variation in subject-specific bias, S_k, is a form of random measurement error. Typically, we assume that this deviation has a mean of 0 and a variance σ_S^2, reflecting the random variation in recall bias from person to person. Adding a term to represent the presence of random subject-specific bias, the model becomes

$$R_{kj} = \beta_0 + \beta_1 (U_k + D_{kj}) + S_k. \quad (3)$$

A second source of random error arises because, on any given day, reporting errors vary from a person's average reporting bias as expressed in model (3). For example, a person may have poorer recall on a day when s/he is fatigued, and more accurate recall on a day when s/he is more alert. We represent this variation with another model term, E_{kj}, which represents the deviation between individual k's 24-hr recall for day j, R_{kj}, from individual k's average reporting bias as expressed in model (3). This term represents another form of random measurement error, which is assumed to have mean 0 and variance σ_E^2.

Summary of Activity Recall Model

Combining all of these terms in an additive model for individual k's 24-hr activity recall on day j, we have

$$R_{kj} = \beta_0 + \beta_1 (U_k + D_{kj}) + S_k + E_{kj,} \quad (4)$$

where β_0 is the intercept and β_1 is the slope for the linear bias, $U_k \sim (\mu_U, \sigma_U^2)$, $D_{kj} \sim (0, \sigma_D^2)$, $S_k \sim (0, \sigma_S^2)$, $E_{kj} \sim (0, \sigma_E^2)$, and $R_{kj} \sim (\beta_0 + \beta_1\mu_U, \beta_1^2\sigma_U^2 + \beta_1^2\sigma_D^2 + \sigma_S^2 + \sigma_E^2)$, assuming that the random model terms are independent (more complex assumptions are possible). The model parameters and the random variables and their distributions are summarized in Table 1.

This model reflects the same linear bias in the mean as expressed in model (2), $\beta_0 + \beta_1\mu_U$, indicating that the mean of the 24-hr recalls may provide a biased estimate of the usual activity mean, μ_U. With the presence of systematic and random measurement error, it is also clear that the variance of the 24-hr recall data, $\beta_1^2\sigma_U^2 + \beta_1^2\sigma_D^2 + \sigma_S^2 + \sigma_E^2$, has the potential to greatly exceed the between-person usual activity variance, σ_U^2, associated with the usual activity distribution for the population. If other types of recall instruments were used, model (4)

Table 1 Summary of Parameters and Variables Used in Measurement Error Models for Activity Recalls (Model 4) and for Monitor-Based Activity Measurements (Model 5)

Parameter	Interpretation		
μ_U	Mean of usual activity for population		
β_0	Intercept for population-level systematic bias in recall		
β_1	Slope for population-level systematic bias in recall		
σ_U^2	Variance of usual activity for population (between-person variation)		
σ_D^2	Variance of daily deviations in true activity relative to an individual's usual activity (within-person variation)		
σ_S^2	Variance of subject-specific bias (between-person variation in person-level systematic bias relative to population bias)		
σ_E^2	Variance of measurement error for activity recall (after accounting for other model terms)		
σ_F^2	Variance of measurement error for monitor-based activity measurement (after accounting for other model terms)		
Variable	**Interpretation**	**Mean**	**Variance**
U_k	Usual activity for person k (unobserved mean daily activity level for person k over a long time period)	μ_U	σ_U^2
R_{kj}	Self-reported activity recall on day j for person k	$\beta_0 + \beta_1\mu_U$	$\beta_1^2\sigma_U^2 + \beta_1^2\sigma_D^2 + \sigma_S^2 + \sigma_E^2$
M_{kj}	Monitor-based activity measurement on j for person k	μ_U	$\sigma_U^2 + \sigma_D^2 + \sigma_F^2$
D_{kj}	Deviation in true activity on day j for person k relative to person k's usual activity	0	σ_D^2
S_k	Deviation from population-level bias for activity recall on day j for person k (subject-specific bias)	0	σ_S^2
E_{kj}	Measurement error in activity recall for person k on day j	0	σ_E^2
F_{kj}	Measurement error in monitor-based activity measurement for person k on day j	0	σ_F^2

may take the same form, although the magnitude of variance and bias terms may be different.

Implications of Using Recall Data to Estimate the Distribution of Usual Activity for a Population

In public health problems that focus on long-term activity behaviors, we are primarily interested in parameters associated with the usual activity distribution.[1] Model (4) makes clear that the distribution of raw recall data does not accurately represent the distribution of long-run average activity for individuals. In particular, without adjusting for systematic bias, the distribution of recalls has the potential to be shifted to the right or left of the true usual activity distribution, and thus is likely to yield an inaccurate estimate of the mean usual activity for a population. Further, because of systematic and random sources of error in recall data, the variation in activity recalls can greatly exceed that of the true usual activity distribution. This has many practical consequences. For example, estimates of the percentage of the population that fails to meet a specific usual activity threshold will likely be biased. In addition, regressing health outcomes on covariates that include activity recall data as a covariate will yield biased estimates of the relationship between physical activity and the health outcome.[11,12] If the activity recall is the only covariate in the regression model, the relationship between the health outcome and activity (expressed by a slope) will be understated and may be declared insignificant when in fact there is a relationship between the health outcome and true usual activity. If additional explanatory variables are present in the regression model, the relationship between activity and the health outcome could be over- or understated.

These implications underscore the need to use measurement error models in analyzing recall data so that parameters associated with usual activity are accurately estimated. In the remainder of this section, we discuss study design considerations and an approach for estimating the parameters of our measurement error model (4).

Study Design Considerations

Many studies record only 1 activity recall per individual, but a single recall per subject does not support estimation of error properties. To estimate the distribution of usual activity or adjustment factors for health outcome regressions, we need to consider other study designs that gather more information on the error properties of the recall measure. Spiegelman and colleagues[21] outline a number of designs and possible models for estimating measurement model parameters. For illustration, we will focus on a design that addresses 2 important features of model (4).

First, we want to separate person-to-person variation in activity levels from day-to-day variation in activity levels for a person. Day-to-day variation is a nuisance factor when considering long-run behaviors such as usual activity. Because we are primarily interested in the person-to-person variation in usual activity, it is important to remove the daily variation in activity levels. To estimate within-person sources of (daily) variation, we need to obtain 2 or more 24-hr recalls on at least a subsample of study respondents. NHANES uses this approach in dietary intake estimation by randomly selecting a sample from NHANES respondents to participate in a second 24-hr dietary recall. The extra recall allows analysts to estimate the within-person variation in daily intakes.

The design must also facilitate estimation of systematic measurement error in the recall relative to true activity during the same time period. This requires a second type of measurement that provides an unbiased observation of activity during the recall time period. The most accurate measures of energy expenditure (eg, doubly-labeled water, calorimetry) are costly and burdensome, and do not provide data on other metrics, such as activity behaviors. However, activity monitors offer a lower cost method of gathering objective data on energy expenditure or on other facets of physical activity for which no other biomarker exists, such time spent sitting or in moderate to vigorous activity. The most useful devices will be sufficiently accurate to provide approximately unbiased measurements on an individual for the activity metric that is being studied. A multisensor monitor such as a Sensewear Pro armband provides a reasonable option because it avoids some of the biases inherent in accelerometers worn on the hip for activity measures.[22] Because the device does not provide a perfect measure of the true activity during the time period it is worn, the objective measure represents a "reference" measure rather than a gold standard. We will present a model below that assumes device measurements are unbiased for the activity metric being studied and are subject to random measurement error. Having concurrent replicate measures of the recall and the objective measurement allows us to estimate true daily variation in activity as distinct from within-person measurement error.

Finally, when estimating the distribution of usual activity, the design should involve selecting a random (ie, probability) sample of participants from the target population to the degree possible. When random sampling is used, estimates can be credibly generalized to the larger population. Probability (or random) sampling involves knowing the probability (or likelihood) of including each member in the sample. Most researchers are familiar with simple random samples for which each member of the population has an equal chance of being included. However, many alternative designs exist that more effectively address operational constraints and estimation goals than simple random samples. For example, an "unequal probability" sample design may oversample minority groups by setting a higher inclusion probability in areas with higher minority populations. Survey weights are used to ensure that the data from the oversampled (and undersampled) areas are "weighted" to represent their true proportion in the population. See Lohr[23] for more details on random sampling and weighting.

A Model for the Reference Measure

In developing a model for the reference measure, in our case for the 24-hr recall, we denote the 24-hr reference measurement for individual k for day j as M_{kj}, where both the 24-hr recall and the reference measurement are taken on the same day j for the same individual. The reference measure is assumed to be unbiased for the target activity measure, in contrast to the self-report model (4). Thus, no intercept and slope are needed in the model for measured activity during the 24-hr period, M_{kj}. In addition, the measurement is not subject to individual biases in self-reporting, so the subjective random error term S_k is also not needed. However, we do need a term for random measurement error in the device, F_{kj}, which represents the deviation of individual k's reference measurement from the true activity on day j. The resulting model for the reference measure can be written as

$$M_{kj} = (U_k + D_{kj}) + F_{kj}, \quad (5)$$

where $U_k \sim (\mu_U, \sigma_U^2)$, $D_{kj} \sim (0, \sigma_D^2)$, $F_{kj} \sim (0, \sigma_F^2)$, and $M_{kj} \sim (\mu_U, \sigma_U^2 + \sigma_D^2 + \sigma_F^2)$. Because the reference measure has fewer sources of error (no population bias or random variation in subject-specific bias), model (5) has a simpler form than model (4) for the 24-hr recall. The reference measure's variance, $\sigma_U^2 + \sigma_D^2 + \sigma_F^2$, is still larger than the usual activity variance, σ_U^2, but smaller than the 24-hr recall variance. The parameters and variables for model (5) are summarized in Table 1.

Estimation Approach

Given data from the type of design we have outlined above, a number of estimation approaches can be applied. Most involve assuming a distribution for the observed activity recall and reference measurements, and for the unobserved usual activity. As with dietary intake data, physical activity data in heterogeneous populations are likely to be right skewed. Some methods begin by transforming the data to normality to simplify the model and estimation approach.[7] After transforming the data to normality, the additive models (4) and (5) are assumed to have normal errors, which implies that the transformed usual activity distribution is normal. Further assumptions are needed regarding the correlations among error terms for the 2 models. Many approaches are possible,[12] and a detailed discussion of these estimators is beyond the scope of this paper. For the purposes of illustration, we develop an example using a method of moments approach developed by Beyler and colleagues[24] for energy expenditure that assumes the original energy expenditure measurements are lognormally distributed. In the example below, the log-scale error variances are considered independent. The individual usual activity distribution is expressed in the original data units by applying a back-transformation, which accounts for the bias that arises in nonlinear back-transformations of means.

Example

The Physical Activity Measurement Survey is a survey conducted in 4 Iowa counties over the course of 8 quarters (3-month waves). The goal of the survey is to collect simultaneous 24-hr physical activity recall and objective activity monitor data on a sample of approximately 1200 adults to support research in measurement error modeling approaches for usual physical activity distributions. Each quarter, a probability sample of households is selected, and an adult aged 21 to 70 is randomly selected from each household to participate in the study. The study participant is asked to wear a Sensewear Pro Minifly armband (Bodymedia Inc., Pittsburgh, PA) from before midnight of a randomly selected target day until after midnight on the subsequent day. On the day after wearing the armband, a telephone interviewer obtains a 24-hr recall from the respondent using an instrument modeled after a computerized 24-hr protocol developed by Matthews and colleagues[25] and validated by Calabro and colleagues.[26] Roughly 10 days later, the protocol is repeated and a second concurrent armband and recall is collected for a 24-hr period. Recall data on activities are processed using a reduced set of MET values from the Compendium of Physical Activity,[16] which was adapted for use in the survey setting. MET-values and corresponding durations for each activity are translated to total energy expenditure (EE) by assuming 1 MET = 0.0175 kcal/kg/min. Monitor data are also converted to total energy expenditure using proprietary algorithms developed by the manufacturer. Recent validation studies conducted with doubly-labeled water demonstrated validity of Sensewear monitors for estimating free-living energy expenditure.[27]

Preliminary data on 171 women from the first sample wave are used to illustrate the measurement error modeling framework, using results from Beyler.[24] Exploratory plots of the average of the 2 recalls per person and the average of the 2 monitor values for each person suggest that the models we posit for the recall and monitor data are reasonable for women in this sample. Consistent with our hypothesized models, the EE distribution for 24-hr recalls differs from that of the monitor (Figure 2). Most recall values exceed the corresponding monitor values, and the median for recall data (thick horizontal line in box plot) is higher than for the monitor data. Because the recall data appear to overstate EE relative to monitor data for most participants, systematic bias terms are needed. In addition, the variability in the recall data are higher than in the monitor data, as indicated by the larger spread in the recall relative to the monitor data (Figure 2). This suggests that the variance for recall data, which includes the impacts of systematic measurement error, person-to-person variation in reporting bias, and random measurement error, is larger than the variance for monitor data, which is subject only to the random measurement error associated with the objective device. It is also clear that the data are skewed to the right, and analyses

indicated that a natural log transformation of both the monitor and the recall data result in approximately normal distributions.

Because age is a factor in EE, Beyler[24] explored whether measurement error parameters vary with age by dividing the preliminary sample into 4 equal-sized groups with age spans of 23–42, 43–52, 53–59, and 60–70 years. We stress that age group estimates are based on sample sizes too small to provide meaningful estimates of activity patterns, and that these results are presented simply to demonstrate the utility of estimating error characteristics and adjusting for them to more accurately estimate the usual activity distribution. Figure 3 presents the linear bias model for log-transformed 24-hr recalled

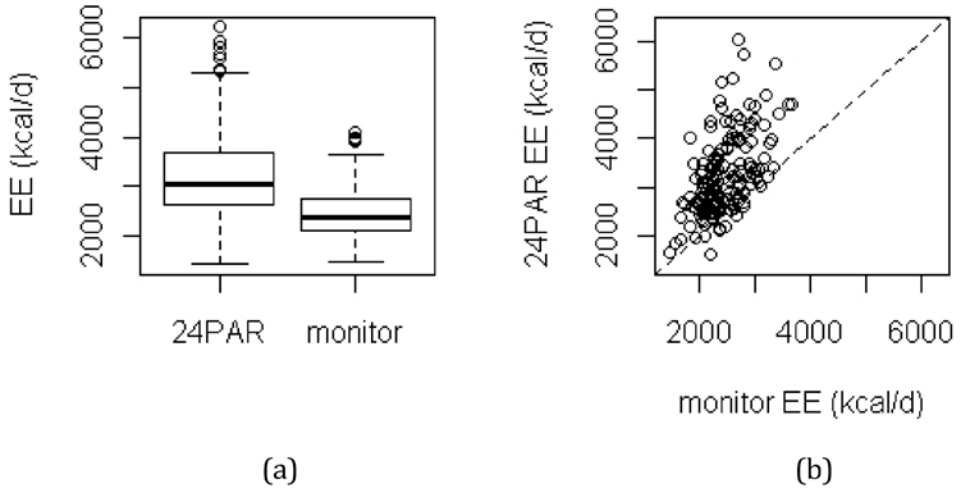

Figure 2 — The relationship between the average of 2 24-hr physical activity recall (24PAR) values and the average of 2 monitor-based EE values for individuals, as expressed by (a) a boxplot depicting the distribution of individual recall and monitor means and (b) a plot of the individual recall means vs. the monitor means.

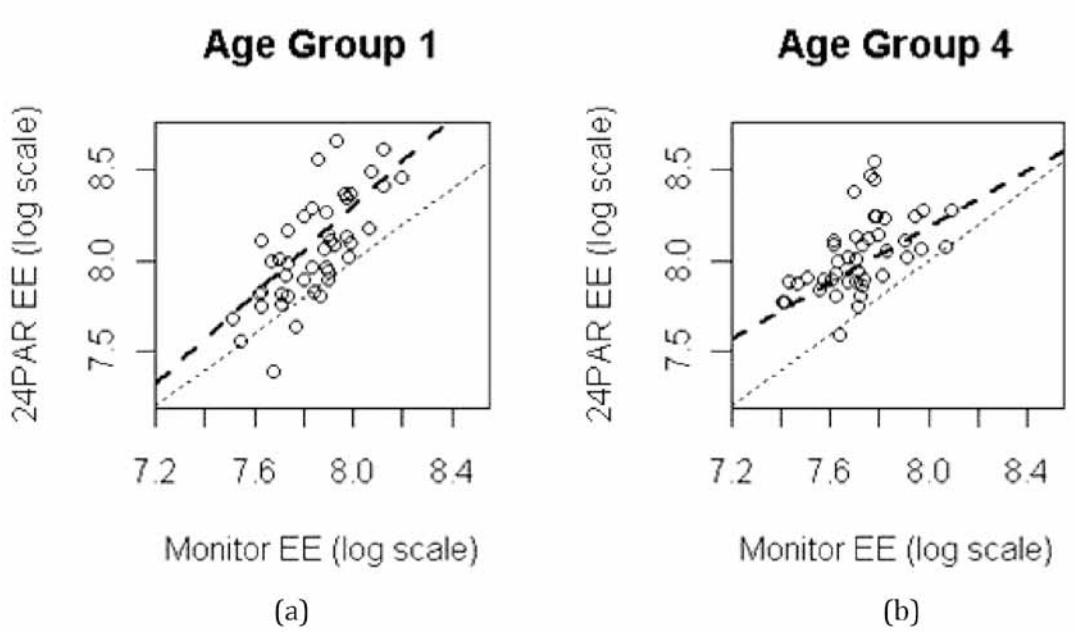

Figure 3 — Estimated bias in average daily EE from 24 to hr physical activity recalls (24PAR) in relation to average daily monitor-based EE for (a) 23- to 42-year-old women (group 1) and (b) 60- to 70-year-old women (group 4). Line with log-dashes is fitted linear bias; line with short dashes is a reference line with an intercept of 0 and slope of 1 (no difference in monitor and recall measurements).

EE relative to 24-hr monitor EE for the youngest and oldest age groups. The fitted bias line (line with long dashes) for both age groups suggested that self-reported data overstate actual EE as estimated by the monitor (line with short dashes). These early data also suggest that the pattern of bias may vary with age. According to the fitted lines, as EE increases, younger adults overstate recalled EE more severely (intercept positive, slope larger than 1 at 1.22, with standard error 0.12), while for older adults, bias in recalled EE recall declines with increasing EE (intercept positive, slope smaller than 1 at 0.73, with standard error 0.08).

Estimated variance components did not vary with age for this small data set. The day-to-day variation in actual activity for a person is only about one-quarter the magnitude of person-to-person variation in usual daily EE (Table 2). This is markedly different from dietary intake data where day-to-day variation in nutrient intake for a person tends to swamp person-to-person variation in usual nutrient intakes, which was the original motivation for constructing measurement error models for dietary intake data.

Turning to the variance components for the monitor, the estimated measurement error variance for the 24-hr monitor means was slightly larger than the true day-to-day variation in the EE for these women. Using the expression for the variance of monitor values under model (5), the estimated variance of a 24-hr monitor EE measurement is about 1.5 times the estimated usual activity variance (Table 2). Thus, while we expect the monitor EE values to have the same mean as the true usual daily EE (due to the unbiasedness assumption), the estimated variation in the monitor-based EE values is clearly larger than the person-to-person variation in usual EE. Even though monitor measurements are prone to less error than the 24-hr recall, the 24-hr monitor values for EE will still overestimate the between-person variation of usual activity, which would affect estimates such as the percent of the population whose activity falls below a threshold level.

For recall data, the major contributor to measurement error variance is the variance associated with the subject-specific bias, which by itself was roughly the same magnitude as the estimated person-to-person variation in usual EE. The random within-person measurement error variance for recalls was slightly larger than for the monitor data. The estimated variance of the 24-hr recall EE for age group 1 (23- to 42-year-olds), which includes effects of systematic and random errors that are not present in the monitor EE variance, was over 3 times larger than the estimated usual activity variance; for age group 4 (60- to 70-year-olds), the recall variance was about twice as large as the variance for usual EE (Table 2). This is a large amount of extra variation in the data relative to the usual activity variance, and further underscores the problems associated with using recall data without adjustment in making inferences related to long-term activity behaviors.

The effects of measurement error variance and bias for age group 1 are presented in Figure 4, which depicts the estimated distribution for 24-hr recalled EE, monitor-based EE, and the usual daily EE distribution. Note that overstated EE values associated with 24-hr recalls shifts the 24-hr recall EE distribution to the right of the usual daily EE distribution. In addition, the large amount of extra variation due to measurement error in recall data are exhibited in the vastly larger spread of the recall distribution relative to the usual daily EE distribution. The monitor distribution also has extra variation due to measurement error, but far less than that of the 24-hr recall EE distribution.

If we were to estimate the fraction of the population whose EE values fell below a threshold, or even the mean of the usual EE distribution using the 24-hr recall data, we risk making biased estimates. For example, if we wanted to estimate the percentage of 23- to 43-year-old

Table 2 Estimates of Variance Components for Log 24-hr EE, Based on Preliminary Iowa Physical Activity Survey Data From 23- to 70-Year-Old Women (for Illustration Purposes Only)

	100 x Estimate (Standard Error)	
Source of variation	**Recall**	**Monitor**
Usual EE (σ_U^2)	2.16 (0.25)	2.16 (0.25)
Daily deviation from usual EE (σ_D^2)	0.50 (0.09)	0.50 (0.09)
Subject-specific bias (σ_S^2)	2.11 (0.30)	
Recall measurement error (σ_E^2)	0.66 (0.12)	
Monitor measurement error (σ_F^2)		0.50 (0.11)
Variance of log 24-hr EE for 23- to 42-yr-old women (group 1)[a]	6.73	3.15
Variance of log 24-hr EE for 60- to 70-yr-old women (group 4)[b]	4.17	3.15

[a] Variance for log 24-hr recall, $(\beta_{1,\,age\ group})^2\sigma_U^2 + (\beta_{1,\,age\ group})^2\sigma_D^2 + \sigma_S^2 + \sigma_E^2$, assumes common variance components for all age groups and separate bias parameters for each age group.

[b] Variance for log 24-hr monitor measurement, $\sigma_U^2 + \sigma_D^2 + \sigma_F^2$, assumes common variance components for all age groups.

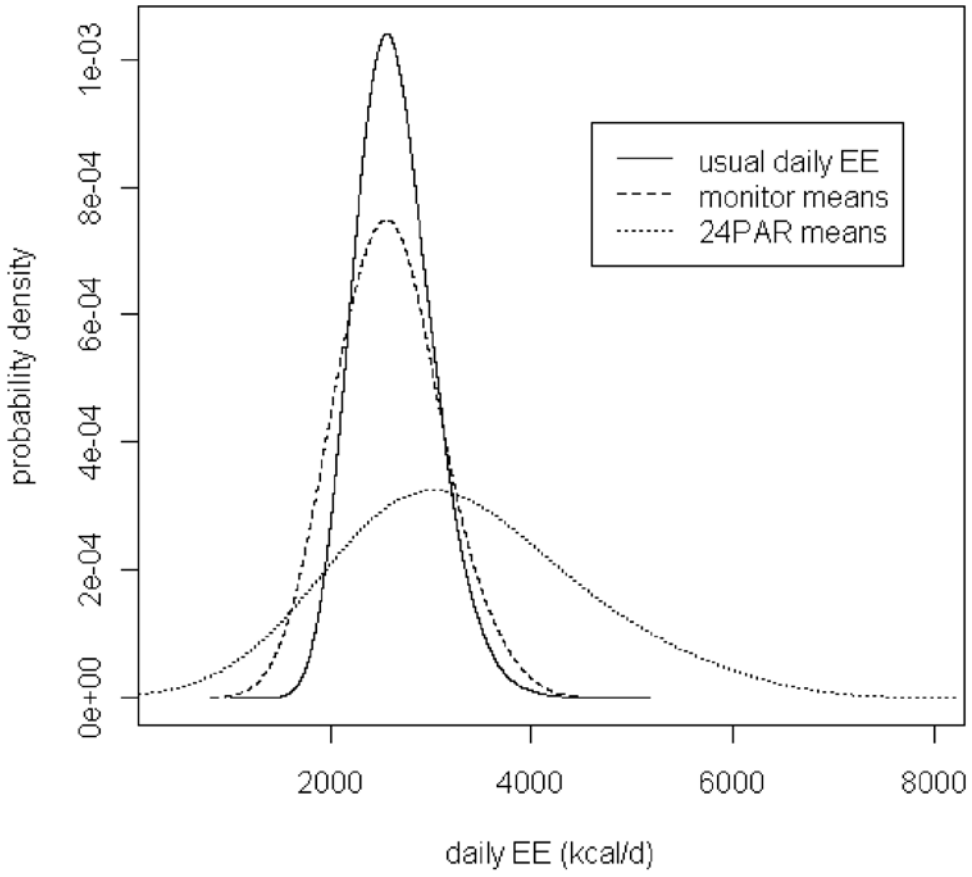

Figure 4 — Estimated distributions of average daily EE for 24-hr physical activity recalls (24PAR) (line with short dashes), monitor-based EE (line with long dashes), and usual EE (solid line).

women whose usual EE values were below 1750 kcals, we would get a much larger estimate if we used 24-hr recall data without adjusting for measurement error than if we were to apply measurement error adjustments in estimating the usual activity distribution.

Conclusions

It is difficult to estimate patterns of long-run behavior of populations. We present an approach that involves making reasonably accurate measurements on short-term behaviors and using statistical models that express the relationship between the short-term and long-term behavior. In constructing this approach, we draw from principles for minimizing error in scientific inferences. These principles rely on preventative measures that avoid error to the extent possible (eg, collecting data that are closely related to the concept of interest, as well as using protocols or questions to obtain data that are as accurate and precise as possible) and adjustment methods that reduce the impact of measurement error in the estimates (eg, measurement error models and associated estimation procedures).

For public health objectives that involve physical activity, it is nearly impossible to get an accurate measure of an individual's usual activity. Fortunately, as with dietary assessment, a practical alternative is to collect reasonably accurate data on very recent behaviors via, eg, a 24-hr recall. The recall is conceptually related to the parameter of interest, usual activity over a long period of time. This relationship can be made explicit via a statistical model that expresses the relatively accurate measure (24-hr recall) as a function of the unobserved target measure (usual activity) and the hypothesized errors in the reasonably accurate 24-hr recall (potential bias, extra variation). The model includes parameters that express the goals of the analysis. In our example, the goals were to estimate the error structure of activity recalls and monitor measurements and to estimate parameters of the usual intake distribution. In other applications, the goal may be to estimate parameters for predicting each respondent's usual activity level for use as a covariate in regression of health outcomes on activity measures. Alternatively, the model parameters may be used to estimate an attenuation coefficient for the regression parameter of the short-term activity measure in a regression model.

The design of the study will depend on the analysis goals. Similarly, the model assumptions and estimation procedures will depend on the design and on the parameters of interest in the investigation. Although a full discussion of the options is beyond the scope of this paper, we have provided an example of how this process would work if credible estimates of usual activity behaviors in a population were of interest, along with estimates of measurement error effects. The model expresses a 24-hr recall as a function of bias and nuisance sources of variation plus the underlying mean and variance of the usual activity distribution. Typical study designs that rely on a single recall per participant are inadequate for estimating the measurement error model parameters. More intensive study is required of at least a subset of individuals to collect data on the potential error in recalls via an objective reference measurement (eg, activity monitor) and on nuisance factors such as daily variation in activity levels for a person. Surveys such as NHANES routinely incorporate this type of protocol for dietary assessment to facilitate more accurate estimation of usual dietary intake distributions. A design that involves replicate concurrent measures of the 24-hr recall and objective reference measure on a portion of the sample will provide the basis for estimating the distribution of usual activity, as well as regression calibration or prediction of usual activity covariates. We also note that scientifically valid inferences should be based on probability samples drawn from a population, rather than convenience samples of available subjects.

While through our example we have focused on 24-hr EE, this approach is more broadly applicable. The recall instrument could be a 7-day recall that is less intensive (but potentially subject to more error) than the procedure used in the Iowa Physical Activity Measurement Survey. In addition, other physical activity metrics can be used, although some may require different distributional assumptions. For example, when considering time spent in moderate-to-vigorous activity (MVPA), a sample of adults may yield a significant number of 0 values that suggest a mixture of 2 populations: those who do not engage in MVPA and thus always have 0 minutes of MVPA, and those who do engage in MVPA and thus have either 0 or positive values for MVPA. A similar approach is taken for distributions of food intakes for components such as fish.[8] We plan to use data from our Iowa Physical Activity Measurement Survey to explore these methods for a range of activity metrics, including EE, MET-hrs, and time spent in various types of behaviors.

Acknowledgments

The authors wish to thank MA Calabro, JM Larson, D Osthus and B Stanfill for their assistance in preparing the manuscript. This work was supported in part by a grant (HL091024) from the National Institutes of Health. We thank anonymous reviewers for their insightful comments on an earlier draft of this manuscript.

References

1. Welk GJ. Introduction to physical activity research. In: Welk GJ, ed. *Physical activity assessments for health related research*. Champaign, IL: Human Kinetics; 2002:3–18.
2. Shephard RJ. Limit to the measurement of habitual physical activity by questionnaires. *Br J Sports Med*. 2003;37:197–206.
3. Matthews CE. Use of self-report instruments to assess physical activity. In: Welk GJ, ed. *Physical activity assessments for health related research*. Champaign, IL: Human Kinetics; 2002.
4. Ainsworth BE. How do I measure physical activity in my patients? Questionnaires and objective methods. *Br J Sports Med*. 2009;43:6–9.
5. Beaton GH, Mimer I, Corey P, et al. Sources of variance in 24-hour dietary recall data: implications for nutrition study design and interpretation. *Am J Clin Nutr*. 1979;32:2546–2559.
6. National Academies. *Nutrient adequacy: assessment using food consumption surveys*. Washington, DC: National Academy Press; 1986.
7. Nusser SM, Carriquiry AL, Dodd KW, Fuller WA. A semiparametric transformation approach to estimating usual intake distributions. *J Am Stat Assoc*. 1996;91:1440–1449.
8. Carriquiry AL. Estimation of usual intake distributions of nutrients and foods. *J Nutr*. 2009;133:601S–608S.
9. Nusser SM, Fuller WA, Guenther PM. Estimating usual dietary intake distributions: adjusting for measurement error and nonnormality in 24-hour food intake data. In: Lyberg L, Biemer P, Collins M, et al, eds. *Survey measurement and process quality*. New York: Wiley; 1997:689–709.
10. Dodd KW, Guenther PM, Freedman LS, et al. Statistical methods for estimating usual intake of nutrients and foods: a review of the theory. *J Am Diet Assoc*. 2006;106:1640–1650.
11. 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.
12. Spiegelman D, Schneeweiss S, McDermott A. Measurement error correction for logistic regression models with an "alloyed gold standard". *Am J Epidemiol*. 1997;145(2):184–196.
13. Groves RM, Fowler FJ, Couper MP, Lepkowski JM, Singer E, Tourangeau R. *Survey methodology*. 2nd ed. Hobekn, NJ: Wiley; 2009.
14. Warnecke RB, Johnson TP, Chavez N, et al. Improving questionnaire wording in surveys of culturally diverse populations. *Ann Epidemiol*. 1997;7:334–342.
15. Adams SA, Matthews CE, Ebbeling CB, et al. The effect of social desirability and social approval on self-reports of physical activity. *Am J Epidemiol*. 2005;161:389–398.
16. Bassett DR, Cureton AL, Ainsworth BE. Measurement of daily walking distance—questionnaire versus pedometer. *Med Sci Sports Exerc*. 2000;32:1018–1023.
17. Altschuler A, Picchi T, Nelson M, Rogers JD, Hart J, Sternfeld B. Physical activity questionnaire comprehension: lessons from cognitive interviews. *Med Sci Sports Exerc*. 2009;41:336–343.
18. Ainsworth BE, Haskell WL, Whitt MC, et al. Compendium of physical activities: an update of activity codes and MET intensities. *Med Sci Sports Exerc*. 2000;32:S498–S516.

19. Irwin ML, Ainsworth BE, Conway JM. Estimation of energy expenditure from physical activity measures: determinants of accuracy. *Obes Res*. 2001;9(9):517–525.
20. Lagerros YT, Mucci LA, Bellocco R, Nyren O, Balter O, Balter KA. Validity and reliability of self-reported total energy expenditure using a novel instrument. *Eur J Epidemiol*. 2006;21(3):227–236.
21. Spiegelman D, Zhao B, Kim J. Correlated errors in biased surrogates: study designs and methods for measurement error correction. *Stat Med*. 2005;24:1657–1682.
22. Welk GJ, McClain J, Eisenmann JC, Wickel EE, Beier S, Flakol P. Field validation of the MTI Actigraph and BodyMedia Armband monitor using the IDEEA monitor. *Obesity (Silver Spring)*. 2007;15:918–928.
23. Lohr SL. *Sampling: design and analysis*. 2nd ed. Brooks/Cole; 2010.
24. Beyler NK. *Statistical methods for analyzing physical activity data. Unpublished doctoral distribution*. Ames, IA: Iowa State University; 2010.
25. Matthews CE, Ainsworth BE, Hanby C, et al. Development and testing of a short physical activity recall questionnaire. *Med Sci Sports Exerc*. 2005;37(6):986–994.
26. Calabro MA, Welk GJ, Beyler N, Carriquiry AL, Nusser SM, Matthews CE. Validation of a computerized 24 hour physical activity recall (24hPAR) instrument with pattern recognition activity monitors. *J Phys Act Health*. 2009;6(2):211–220.
27. Johannsen DL, Calabro MA, Stewart J, Franke W, Rood JC, Welk GJ. Accuracy of armband monitors for measuring daily energy expenditure in healthy adults. *Med Sci Sports Exerc*. 2010;42(11):2134–2140.