Improving Self-reports of Active and Sedentary Behaviors in Large Epidemiologic Studies

Charles E. Matthews[1], **Steven, C. Moore**[1], **Stephanie M. George**[2], **Joshua Sampson**[3], and **Heather R. Bowles**[2]

[1]Nutritional Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD

[2]Applied Reseach Program, Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD

[3]Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD

Abstract

Questionnaires that assess active and sedentary behaviors in large-scale epidemiologic studies are known to contain substantial errors. We present three options for improving measures of physical activity behaviors in large-scale epidemiologic studies, discuss the problems and prospects for each of these options, and highlight a new direction for measuring these behaviors in such studies.

Keywords

Exposure assessment; Exercise; Sitting; Measurement error; Disease prevention

INTRODUCTION

Passmore and Durnin's (34) painstakingly developed a methodology to estimate energy expenditure in free-living humans, involving direct observation, time diaries, and metabolic measures. This approach was simplified in the course of developing physical activity questionnaires that were designed to examine the relation between usual physical activity levels and disease in large-scale epidemiologic studies (29). These questionnaires, which typically rely on long-term recall to estimate usual levels of exposure, have been invaluable in demonstrating the numerous health benefits of physical activity (36), and more recently the adverse effects of sedentary behaviors (33). Yet, the questionnaires used in these studies are likely to contain substantial measurement error and, in terms of physical activity, at best only capture 50% of the variation in objectively measured activity energy expenditure (31).

Measurement errors in prospective epidemiologic studies usually attenuate, or reduce the magnitude, of observed behavior-disease associations, resulting in a loss of statistical power for the hypothesis being tested (39). Furthermore, quantitative measures of the amount (or

Corresponding Author: Charles E. Matthews, PhD, Nutritional Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer, Institute, 6120 Executive Blvd, EPS 3028, Bethesda, MD 20892-7335, charles.matthews2@nih.gov.

dose) of exposure associated with either benefit (physical activity) or risk (sedentary time) may be biased because of these errors (41). If the errors are sufficiently large (15), measurement error could pose considerable challenges for translating results from epidemiologic studies to physical activity guidelines that inform health promotion efforts and public policy.

In this paper, we focus on the measurement error problem for self-reports of "usual" levels, of active and sedentary behaviors in studies designed to provide quantitative estimates of health risks associated with a given level of these exposures. We use the term "usual" to indicate a long-term average dose or volume of these behaviors (e.g., over one year), and make a distinction between self-report methods that employ long-term recall and averaging to estimate usual behavior (i.e., Questionnaires) and methods that employ short-term recall of behavior to estimate usual levels of activity or sedentary behavior. Given their limited ability to evaluate dose-response relationships, we do not consider questionnaires that were designed only to classify individuals into broad categories of activity, (e.g., instruments such as the Lipid Research Clinics and Stanford Usual Activity Questionnaires).

In the first section of the paper, we review the strengths and limitations of existing questionnaires that commonly assess usual levels of active and sedentary behaviors. Next, we describe the consequences of measurement errors in these questionnaires in epidemiologic studies and then consider the available options for minimizing these consequences and/or reducing the level of error in the exposures by using better measures. In the final section, we discuss the potential utility for short-term recalls to provide less error prone estimates of usual levels of exposure in large-scale epidemiologic studies.

Strengths and Weaknesses of Physical Activity and Sedentary Behavior Questionnaires

There is ample evidence from observational studies that questionnaire-based physical activity measures are associated with reduced risk for many chronic diseases such as diabetes, cardiovascular disease, and osteoporosis, as well as certain cancers (e.g., colon, breast, and endometrial) (30, 36). In addition, relative to a broad range of biological (e.g., fitness, fatness), objective (doubly labeled water, accelerometers), and other self-report (e.g., diaries) comparison measures there is evidence that many physical activity questionnaires are able to capture valuable information (45). Results from these studies suggest that many questionnaires can provide a useful ranking of active or sedentary behaviors, but their major limitation is that the level of error in quantifying dose or absolute volume is large.

Reporting errors in assessments of active and sedentary behaviors emanate from misreporting of two basic elements of dose: (1) the usual duration of the behaviors reported, or (2) the intensity of the activities reported (34) in relation to relevant exposure metrics (e.g., metabolic equivalents [METS], bone loading)(3, 14). For the sake of simplicity, we will consider errors in duration and intensity separately, although we recognize that errors in determining intensity can affect the errors in duration. In general, the approach to assessing the usual amount of time spent engaged in specific types of behavior has been to directly ask about the usual duration (per week or per day) of the activity, or to use a decomposition strategy that asks for information about activity frequency (i.e., number of months, days per week) and duration (average time per occasion) separately. Reporting errors in one or both of these decomposed elements can result in large errors in the estimate of usual duration. Interestingly, Passmore and Durnin (34) were keenly aware of the importance of obtaining accurate duration estimates in their measures: "In estimating the expenditure of any individual, it is our experience that larger errors are likely to arise from the failure to determine correctly the length of time spent in any activity rather than in any assessment of the metabolic cost of that activity." Doing more to reduce the magnitude of the errors in

reported duration in active and sedentary behaviors may be one opportunity to substantially reduce the errors in our measurements.

In order to consider the influence of activity intensity on health, reports of the usual activity duration are typically combined with standard intensity values, such as METs or bone loading units(14), to estimate a duration-intensity weighted metric for the activities reported (e.g., MET-hrs/d). It is recognized that intensity values may not reflect the relative intensity of the activity performed, and that for many activities there can be a large inter- and intra-individual variation in the physiological effects of a given activity (2, 41). This latter caveat may be exacerbated for questionnaire items that ask about a broad range of activities (e.g., household chores), or that employ physiologic cues to help classify the energy cost of the activity (e.g., increased heart rate, sweating). *Analytic* errors in the intensity components may arise from errors that emerge when a fixed tabular value (e.g., a MET value) is applied to an individual's report of an activity, while *reporting* errors in intensity arise when respondents misclassify a behavior in the wrong intensity category (e.g., reporting a light activity as moderate). Reducing intensity-reporting errors may also be an important approach to reducing overall measurement errors in self-report instruments.

The Cognitive Demands Involved in Reporting Long-term Averages Are Extraordinary

Reporting autobiographical information on a questionnaire about usual participation in active and sedentary behaviors forces respondents to retrieve and organize a great deal of information in order to formulate a response (27). It has long been known that vigorous activities (often more structured exercise) tend to be more reliably reported than moderate intensity activities (37, 45), and that other lower intensity daily activities (e.g., non-exercise activity), often done in several short bouts within a day, are the least reliably reported. Indeed, questions about household activities were dropped from early questionnaires because of the difficulties associated with reporting them (29). A striking example of the challenges associated with reliably assessing common daily activity was observed by Dipietro (12) in her examination of the test-retest reliability of the Yale Physical Activity Survey. Figure 1 illustrates that test-retest reproducibility (i.e., reliability), indicating the ability of respondents to provide consistent answers for specific activities on the questionnaire, is best for less frequent activities done in specific episodes and worst for the most prevalent daily activities (27). Instruments to assess sedentary behaviors are starting to appear, and consistent with physical activity, more structured sedentary behaviors appear to be more reliably reported (17).

Studies using advanced activity monitors provide insight into the magnitude of the cognitive demands associated with reporting usual levels of activity, particularly common daily activities. Levine (24) recently reported that adults engaged in an average of 47 bouts of active and sedentary behaviors each day, and that the average amount of time spent upright and ambulatory was about 6.5 hrs per day; mostly accumulated in short bouts of activity. Assuming these estimates are representative for adults, in order to literally report what they usually did over one month a respondent would have to cognitively process information about 1,400 bouts of activity and nearly 200 hours of active time. Clearly, the cognitive demands are staggering, and thus it is not surprising that errors in reporting physical activity by questionnaire, particularly common daily activities, is large.

Measurement Error in Questionnaires Attenuates Behavior-disease Associations

Studies that have concurrently evaluated risk for mortality associated with low levels of objectively measured physical activity energy expenditure and activity reported by

questionnaire, have indicated that associations with measured activity energy expenditure are much stronger than those obtained by self-report. Manini (26) examined mortality outcomes in relation to physical activity energy expenditure measured by doubly labeled water (DLW) among older adults and noted nearly a 70% reduction in risk among the most active participants as measured by DLW, but no association with self-reported activity. In addition, studies that have measured cardiorespiratory fitness as well as physical activity reported by questionnaire have indicated that associations with objectively measured fitness are consistently stronger than those with self-reported physical activity (8). Collectively, these data are consistent with the notion that measurement errors in physical activity questionnaires attenuate the strength of associations, and indicate that the impact of the errors may be substantial. While we know less about the potential measurement error in reported sedentary behaviors, it is likely that attenuation due to error may obscure these associations as well.

While attenuation of the strength of the true associations between active and sedentary behaviors and disease are often discussed as a limitation in etiologic studies, the actual level of attenuation is unknown. Measurement error models can quantify these effects. Here we introduce a simple model to describe these errors and use information derived from the model to assess impact of random errors on epidemiologic associations (i.e., attenuation). To quantify these parameters, and the magnitude of the attenuation, consider the simple model where Q_i is an unbiased estimate of the true value (T_i) for individual i. The additional term (ε_i) is random error with a mean of 0 and variance (σ_ε^2).

$$Q_i = T_i + \varepsilon_i$$ [Equation 1]

For example, a study might be interested in testing the hypothesis that time spent sitting and watching television is associated with increased risk for endometrial cancer. Investigators would use a questionnaire to estimate the true amount of exposure (T_i), but with some level of random error. The questionnaire-based estimate of television viewing (Q_i) would then be used to quantify any association with this health outcome. If the level of random error in questionnaire is small, then Q_i is a good approximation of T_i, or the true amount of sitting and watching television and any real signal between television and endometrial cancer would be observable. However, if the amount of random error on the questionnaire was large, say one hundred percent of the true value, then the questionnaire would provide a poorer approximation of T_i, and the signal between television watching as measured by questionnaire and the outcome would be obscured by the "noise" associated with random errors. In this simple model, the amount of attenuation of the true behavior-disease association that is due to measurement error in the questionnaire can be quantified as an attenuation factor (4, 22). Specifically, the attenuation factor (λ) is defined as:

$$\lambda = \frac{1}{1 + \sigma_\varepsilon^2 / \sigma_T^2}$$ [Equation 2]

where the variance of the true measure is σ_T^2 (4). When the measurement errors are very small, the attenuation factor is close to 1.0, but as these errors increase, the attenuation factor typically gets smaller, as does the strength of the associations that can be observed. As an approximation, if we let the relative risk (RR), or the risk for disease comparing high to low levels of an exposure, denote the strength of the underlying association between the true exposure (T_i) and the outcome, then the magnitude of the RR that is observable with the questionnaire can be estimated as RR^λ (4). Therefore, if the attenuation factor is 0.5, and the true RR for endometrial cancer is increased 1.20 times for each additional hour of television viewing, we would only observe a RR of 1.10 using the questionnaire (i.e., RR =

1.20$^{0.5}$=1.10). Similarly, if the true RR for television viewing and heart disease is 4.0, we would only observe a RR of 2.0 using the questionnaire (i.e., RR = 4.0$^{0.5}$=2.0).

In addition to random error, self-reports can also include systematic errors or biased reports of active and sedentary behavior, and these errors can further decrease the attenuation factor, and can quickly reduce the magnitude of the relative risks that are observable in etiologic studies to an undetectable value.

Improving Self-report Measures and Obtaining More Accurate Behavior-disease Associations

In Figure 2 we present three basic options for reducing the impact of measurement errors in epidemiologic research on physical activity and health. The first uses statistical methods to quantify and correct for errors in questionnaires, while the latter options reflect exposure assessment methods that are simply less error prone. The options are: (1) Use measurement error correction methods to minimize the impact of reporting errors on questionnaires (42), (2) Use objective indicators of active and sedentary behaviors to eliminate reporting errors; or (3) Use short-term recalls to reduce the magnitude of the reporting errors in estimates of usual levels of behavior. Hybrids of these basic options are also possible. For example, a calibration study outlined in Option 1 (below) also could be applied to Option 3 in order to adjust for random and systematic errors present in short-term recalls (32), and measurement error correction approaches also could be applied to minimize intra-individual error in activity monitor data (46). In the remainder of the report we discuss the problems and prospects associated with the three basic options outlined in Figure 2.

Option 1. Use Measurement Error Correction to Minimize Impact of Errors in Questionnaires

The first option is to evaluate the measurement error in questionnaires that assess usual levels of active and sedentary behaviors through a calibration study, and then adjust the strength of the associations observed using measurement error correction methods, e.g., (21, 32, 42). The calibration study measures the level of relevant behaviors on a small subset of study participants with a reference instrument, which is presumed to be more accurate than the questionnaire used in the larger study. With this information, we can reconstruct an estimate of the true effect size from our study. In the simplest case described earlier (Equation 1), we could estimate the true relative risk by exponentiation of the naïve relative risk using the inverse attenuation factor (1/λ). However, usually, such reconstruction requires more complex measurement error models. Here, we expand Equation 1 to accommodate this complexity. General "activity-related" bias, or systematic errors that are expressed over the range of the exposure, can be accounted for by including an intercept β_0 and a slope β_1 term to describe the relation between the questionnaire (Q_i) and the true value derived from a reference measure (T_i). Examples include...

$$Q_i=\beta_0+\beta_1 T_i+\varepsilon_i \qquad \text{[Equation 3]}$$

Although, each individual, by definition, must continue to have only a single true value of usual exposure, they might receive a questionnaire at multiple time-points. Therefore, we require an additional subscript and let Q_{ij} be the questionnaire value reported for individual i at time j. When this occurs and multiple measurements are taken on each individual, it is possible to estimate systematic reporting errors within the same individual over time (i.e., "person-specific" bias, r_i)(22, 32). For example, individual i may consistently underestimate her true time sitting watching a television on the questionnaire. We now relate the questionnaire value(s) and the true value for individual i by (22):

$$Q_{ij}=\beta_0+\beta_1 T_i+r_i+\varepsilon_{ij} \qquad \text{[Equation 4]}$$

We generally assume that r follows a normal distribution with mean 0 and variance σ_r^2. The attenuation factor resulting from the above model would be

$$\lambda=\frac{\beta_1}{\beta_1^2+\frac{\sigma_r^2}{\sigma_T^2}+\frac{\sigma_e^2}{\sigma_T^2}} \qquad \text{[Equation 5]}$$

Close inspection of the model in equation 4 reveals that the quantities derived for two of the three error terms estimated for Q (i.e., activity-related and person-specific biases) are dependent on the value of the reference measure, which is taken to be an unbiased estimate of the true value (T_i). While the reference measures commonly used in physical activity studies, such as physical activity monitors and doubly labeled water, can provide insight into the ability of self-report instruments to rank-order individuals, greater scrutiny of these methods—and the questionnaires against which they are compared—is necessary in the context of estimating the bias terms in measurement error models.

If systematic errors are present in the reference measures then the instrument may not provide accurate estimates of the bias terms in the model, and thus may not provide accurate estimates of validity of the instrument, or the attenuation factors derived from the results. For example, the first generation physical activity monitors that employed one minute epoch data and linear regression calibrations to estimate energy expenditure performed well in laboratory studies of walking and running, but they clearly underestimated the energy cost of many common daily activities requiring less ambulation, such as household chores (28). Consistent with this finding, recent comparisons against doubly labeled water indicate that this class of accelerometers may underestimate physical activity energy expenditure by at least 10% (e.g., (10, 20)). Results from studies that employ this class of activity monitors should be interpreted accordingly. Considerable progress is being made in the assessment of common daily activities by accelerometer (e.g. (16, 43)), and we are hopeful that studies in free-living subjects will demonstrate that the accuracy of these devices will improve sufficiently to meet the requirements of a valid reference measure in this context. New devices that measure body position and sedentary behavior with better accuracy appear to be promising options and should be evaluated for this purpose (e.g., 17, 21).

After accounting for resting metabolism and dietary thermogenesis, DLW can be used to estimate the average level of physical activity energy expenditure and many consider this method to be the best available reference measure of overall physical activity energy expenditure. But, there is an important caveat for using this method in the context of measurement error modeling from questionnaires of usual physical activity levels. DLW is an integrated measure of the energy expenditure resulting from all of the different activity behaviors that participants engage in during the measurement period. In contrast, most questionnaires asses only a select subset of activities generally believed to contribute most to overall physical activity energy expenditure. Neilson (31) recently showed that many if not most questionnaires substantially *underestimated* activity energy expenditure in comparison to DLW, most likely because they fail to assess common daily activities that contribute to overall energy expenditure. Thus, potential differences in the scope of the activities assessed by questionnaires and DLW estimates of overall physical activity energy expenditure warrants careful consideration when using DLW as a reference measure to quantify the error structure in the self-reports of physical activity.

The recent focus on the adverse health effect of sedentary behaviors (33) have highlighted the need to measure sedentary behaviors in etiologic studies (33). Although time spent sitting is associated with reduced physical activity energy expenditure (25), the inability of DLW to quantify time spent in sedentary behaviors directly suggests a measure of energy expenditure may not be a suitable reference measure in calibration studies designed to determine the error structure of sedentary behavior questionnaires. The next generation of physical activity monitors, which assess body position directly, may be required for this purpose (e.g., (18, 23)).

In summary, implementation of calibration studies and measurement error correction methods to estimate the error structure of questionnaire-based estimates of usual behavior and adjust risk estimates for attenuation may be a valuable approach for future epidemiologic investigations. When the assumptions of the method are met they offer an opportunity to more accurately estimate the true magnitude of association between physical activity and the health outcomes of interest.

Option 2. Use Objective Indicators of Behavior to Eliminate Reporting Errors

One attractive option for dealing with errors associated with self-report would be to completely eliminate this source of error by opting to use objective indicators of behavior rather than self-report instruments. We use the phrase "objective indicators of behavior" to describe measurements derived from physical activity monitors, which measure body motion and/or position in order to make inferences about behavior, and DLW which can measure physical activity energy expenditure resulting from time spent in different behaviors (11, 26). The major strength of these measures, of course, is that errors associated with self-report are completely removed, the analytic errors inherent in the measures are relatively low (e.g., laboratory error for DLW, technical reliability of accelerometers), and accordingly the level of attenuation in the associations observed would be expected to be greatly reduced (11)(Figure 2). However, as noted previously accelerometers data can also contain systematic and random measurement error, and a single DLW assessment is subject to errors associated intra-individual variation. An additional limitation of using objective indicators of behavior alone in large studies is the general absence of contextual information provided by the measures. Contextual information may include insight about the type of activity (e.g., aerobic vs. strengthening activities), as well as information about the behavioral setting within which participants engage in a given behavior (e.g., at home or work, sitting in a car). Key scientific questions of public health importance relate as much to the amount of a behavior as the context within which the behavior occurs. The value of contextual information cannot be underestimated because this data element facilitates translation of the evidence for specific behavior-disease association to health interventions, and to public policy.

The relatively higher cost and logistical demands associated with implementing objective measures in large-scale studies also can limit the use of these methods. Objective measures have been extremely valuable in providing new insights into physical activity and health in small to moderate-sized studies (e.g., (24, 26)), but in very large studies designed to examine rare health outcomes such as cancer (40), cost and feasibility often remains a limiting factor. For these reasons, reliance on objective indicators of behavior alone is not always the best measurement option, particularly in studies that seek to understand the context in which active and sedentary behaviors occur and in very large studies where costs associated with activity monitoring are more difficult to manage.

Option 3. Use Short-term Recalls and Reduce Reporting Errors in Behaviors

This approach to improving self-reported measures of active and sedentary behaviors is to use a more accurate and detailed self-report instrument that is capable of reducing the magnitude of the errors in the information reported (Figure 2). The application of measurement error correction models can further minimize the impact of random error, as well as systematic errors if a calibration study is conducted with valid references measures (32) (i.e., a hybrid combining Option 1 and 3, Figure 2). Given the cognitive demands associated with reporting usual activity levels via questionnaires, significant advances in reducing reporting errors in these questionnaires appears unlikely. The question is whether there are other more accurate self-report methods that might be considered.

Following the lead of nutritionists (39) and time-use researchers (7), multiple 24-hour recalls could be used to improve assessment of active and sedentary behaviors. Because they have been generally assumed to be less error prone and more detailed, short-term recalls have commonly been used in energy requirement studies (34), and to examine the measurement properties of physical activity questionnaires (e.g., (19)). An important advantage of short-term recalls is that they rely more extensively on the recollection of specific behaviors/events using episodic memories, whereas questionnaires of usual behaviors often force respondents to rely on generic memories of past events and to employ estimation strategies to report past behavior (27). Among time use researchers there is some consensus that short term recalls are a preferred method of capturing information about the kinds of unstructured common daily behaviors (e.g., housework) that traditionally have proven the most difficult for physical activity researchers to measure (7).

In particular, short-term recalls have the potential to reduce errors in the duration of the activities reported as compared with estimates derived from questionnaires of usual levels of exposure. For example, by reducing the recall interval on the previous day to specific segments within the day (e.g., morning, afternoon, evening), short-term recalls begin to limit the scope of allowable reporting errors (5). If a respondent is allowed to report more specifically the duration of the individual bouts of active or sedentary behavior they engage in, rather than daily totals, then the information provided can be tallied by the data collection system, which should further reduce mathematical errors in the reporting process. Thus, a major advantage of short-term recalls may be their ability to rein in errors in estimating the duration of active and sedentary behavior on days for which the reports are provided.

Use of Short-term Recalls of Active and Sedentary Behaviors in Epidemiologic Studies

Over the last decade, 24-hour physical activity recalls (24PARs) have been administered by phone in a number of studies, the results of which provide insight into the potential utility for their use in etiologic studies. A study among middle-aged adults, found that 24PARs were correlated with accelerometer measures of physical activity and that only two to three 24PARS were required to achieve reasonable correlations (32) with a questionnaire that had previously been found to explain 45% of the variance physical activity energy expenditure as measured by DLW (35) In a study of postmenopausal women that compared seven 24PARs to DLW measures over 14 days, no significant differences in physical activity energy expenditure between measures were found, and reporting errors were not associated with body mass index or social desirability (1). Cabalaro (9) compared estimates of total energy expenditure (kcal/d) and time spent in moderate-vigorous activity from two different pattern recognition activity monitors to similar metrics derived from the 24PAR. The 24PAR-based estimates of total energy expenditure were not significantly different from, and were highly correlated with (r ~ 0.9), expenditure from the monitors. Correlations for moderate-vigorous activity duration were lower, but still relatively high (r ~ 0.6). Results from a recent study that employed the 24PAR are consistent with objective monitoring

studies indicating that adults spend little time in moderate-vigorous activity, the majority of their time in sedentary behaviors, and a considerable amount of time in light activity, suggesting that short-term recalls may be particularly useful in gathering information about sedentary behaviors and common daily activities (Figure 3). Collectively, this series of studies and other recent reports (44) using similar methods suggests that there may be considerable utility in using short-term recalls of active and sedentary behaviors in epidemiologic studies.

Obstacles to Using Short-term Recalls in Large Epidemiologic Studies

Although short-term recalls, such as diaries or previous day recalls, are generally considered to be less error prone they have rarely been used as a primary assessment of activity behaviors because of the costs of obtaining a sufficient number of repeated measures to estimate usual activity levels, the high participant burden and coding and data entry costs associated with diaries. Furthermore, study participants may not comply with protocols for completing diaries, thereby potentially introducing reporting errors. For example, a diary protocol may require participants to record their activities at set intervals over a day to minimize forgetting, but participants may put off recording for a more personally convenient time. Recall errors may be introduced by delaying the recording of activities beyond specified windows of recall and report. Computer assisted interviews by phone can reduce costs associated with coding and data entry, and may limit the participant burden, but the expense of conducting the interviews can be high. However, mobile devices (e.g., phones, tablets) and computers linked to internet-based data collection methods for short-term recalls may resolve these problems because self-administration by participants and automated data collection processes has the potential to obviate the need for interviewers (39).

The other major obstacle associated with using short-term recalls is concern about how effective assessment of only a few days of observation may be in providing useful estimates of usual levels of active and sedentary behaviors. This error, considered intra-individual variation in behavior (or within-person error), is captured by the e term in the models. For our discussion, we shall assume that all ε_{ij} are normally distributed and independent of each other, but repeat measurements may not always satisfy these assumptions (6). For example, measurements recorded within the same week can be correlated due to weather, work, or health. Similarly, measurements recorded on the same day of the week may be correlated due to work schedules, and exercise and television viewing habits. However, if we intelligently design our collection of replicate measures, we can obtain a relatively accurate and unbiased estimate of usual activity levels. In fact, when our assumptions of normality and independence are met for our ε_{ij} term, only a few repeat measures over time can be extremely useful in reducing the impact of intra-individual variation in behavior on our measures. Table 1 describes how the attenuation factor and statistical power increases with the number of replicates under these simplifying assumptions, as a function of the percentage of the total variation attributable to the intra-individual variation in behavior. In this example, we estimate the effect on statistical power for a 100 subject study for each effect size at an alpha=0.01 level. When intra-individual variation associated with a single replicate recall is greatest (i.e. 80%), the addition of two additional recalls (three total replicates) result in an approximate doubling of the attenuation factor (from 0.20 to 0.43), an increase in the strength of the observable association, and an approximate doubling of the statistical power available. Table 1 also shows that as the total number of replicates increase, the benefit of additional replicate measure begins to diminish, particularly beyond three to four recalls. This is consistent with results from nutritional epidemiology demonstrating that four replicate 24-hour dietary recalls can substantially reduce random measurement errors (38). We have presented this simple scenario to highlight the idea that a modest number of replicate measures can substantially reduce measurement error associated with intra-

individual variation in behavior. However, since daily variation can follow specific patterns over time (e.g., seasonality, day of the week effects), real life scenarios are more complicated and the optimal method for quantifying intra-individual variation and the schedule for collecting replicates requires careful thought (6).

There are, of course, some limitations to using short-term recalls in epidemiologic studies. First, this approach may reduce but does not eliminate measurement error, and it only assesses current behavioral exposures during a given measurement period (e.g., a 12 month period at study baseline). Information about historical activity patterns, which could be important for some health outcomes, cannot be measured directly and questionnaire-based approaches would be required to capture this information. Short-term recalls may also be less adept at estimating levels of less frequent behaviors, such as exercise participation or more seasonal activities. However, statistical methods are being developed that may be able to translate a few discrete observations of less frequent behaviors into meaningful estimates of usual levels of dietary behaviors (e.g., (13)).

A New Direction in Assessment of Activity and Sedentary Behavior in Epidemiologic Studies

The Activities Completed Over Time in 24 Hours (ACT24) system is a self-administered web-based physical activity assessment tool that has been developed by investigators at the National Cancer Institute. It asks respondents to report how they spent their time in the previous 24-hours including time sleeping, and in active and sedentary behaviors. The program leads respondents through four 6-hour time-periods, asking them to record their activities on a timeline. They browse and select from over 100 individual activities listed and can search for an additional 110 exercise and sports activities. Follow-up questions determine time spent in each activity, as well as selected activity-specific questions (e.g., body posture, rating of perceived exertion during exercise). Respondents typically report 20 to 30 distinct active/sedentary behaviors in each recall day. Summary values for time spent sleeping and in active and sedentary behaviors, as well as energy expenditure (MET-hrs/d) are derived from the information reported. The goal is to have ACT24 available to interested researchers, providing a website to register studies and to provide access to the system for respondents to complete recalls. A demonstration version of the current instrument is available for review (; http://act24demo.westat.com).

Summary

Existing self-report questionnaires of active and sedentary behaviors that are suitable for use in large-scale epidemiologic studies are known to contain substantial errors. For future large-scale epidemiologic studies of physically active and sedentary behaviors and health, we present three options for improving our assessment of these important behavioral exposures: (1) correcting errors in self-report questionnaires of usual behaviors analytically using calibration studies and measurement error correction models; (2) eliminating reporting errors by using objective indicators of behavior, or (3) by reducing the magnitude of the reporting errors through use of short-term recalls. Given that short term recalls may reduce the magnitude of reporting errors, and because they also offer the opportunity for gathering salient contextual information about the behaviors reported, we highlight the potential for short-term recalls to be used in future epidemiologic studies and discussed how we might overcome obstacles to their use.

Acknowledgments

Reference List

1. Adams SA, Matthews CE, Moore CG, Cunningham JE, Fulton J, Hebert JR. The effect of social desirability and social approval on self-reports of physical activity. American Journal of Epidemiology. 2005; 161:389–398. [PubMed: 15692083]

2. Ainsworth B, Haskell W, Leon A, Jacobs D, Montoye H, Sallis J, Paffenbarger R. Compendium of physical activities: classification of energy costs of human physical activities. Medicine & Science in Sports & Exercise. 1993; 25:71–80. [PubMed: 8292105]

3. Ainsworth B, Haskell W, Whitt M, Irwin M, S A, Strath S, O'Brien W, Bassett D, Schmitz K, Emplaincourt P, Jacobs D, Leon A. Compendium of Physical Activities: An Update of Activity Codes and MET Intensities. Medicine & Science in Sports & Exercise. 2000; 32:S498–S516. [PubMed: 10993420]

4. Armstrong BG. The effects of measurement errors on relative risk regressions. American Journal of Epidemiology. 1990; 132:1176–1184. [PubMed: 2260550]

5. Baranowski T. Validity and reliability of self-report measures of physical activity: an information-processing perspective. Research Quarterly for Exercise and Sport. 1988; 59:314–327.

6. Baranowski T, Masse LC, Ragan BWG. How Many Days Was That? We're Still Not Sure, But We're Asking the Question Better! Medicine & Science in Sports & Exercise. 2008; 40

7. Bianchi SM, Milkie MA, Sayer JP, Robinson JP. Is Anyone Doing the Housework? Trends in the Gender Division of Household Labor. Social Forces. 2000; 79:191–228.

8. Blair SN, Ching Y, Holder SJ. Is physical activity or physical fitness more important in defining health benefits? Medicine & Science in Sports & Exercise. 2001; 33(Supplement-S399)

9. Calabro MA, Welk GJ, Carriquiry AL, Nusser SM, Beyler NK, Matthews CE. Validation of a Computerized 24-Hour Physical Activity Recall (24PAR) Instrument With Pattern-Recognition Activity Monitors. Journal of Physical Activity and Health. 2009; 6:211–220. [PubMed: 19420399]

10. Colbert LH, Matthews CE, Havighurst TC, KIM KYUN, Schoeller DA. Comparative Validity of Physical Activity Measures in Older Adults. Medicine & Science in Sports & Exercise. 2011; 43

11. Colbert LH, Schoeller DA. Expending our Physical Activity (Measurement) Budget Wisely. Journal of Applied Physiology. 2011

12. DiPietro L, Caspersen CJ, Ostfeld AM, Nadel ER. A survey for assessing physical activity among older adults. Medicine & Science in Sports & Exercise. 1993; 25:628–42. [PubMed: 8492692]

13. Dodd KW, Guenther PM, Freedman LS, Subar AF, Kipnis V, Midthune D, Tooze JA, Krebs-Smith SM, Dodd KW, Guenther PM, Freedman LS, Subar AF, Kipnis V, Midthune D, Tooze JA, Krebs-Smith SM. Statistical methods for estimating usual intake of nutrients and foods: a review of the theory. Journal of the American Dietetic Association. 2006; 106:1640–1650. [PubMed: 17000197]

14. Dolan SH, Williams DP, Ainsworth BE, Shaw JM. Development and reproducibility of the bone loading history questionnaire. Medicine and Science in Sports and Exercise. 2006; 38:1121–1131. [PubMed: 16775555]

15. Ferrari P, Friedenreich C, Matthews CE. The Role of Measurement Error in Estimating Levels of Physical Activity. American Journal of Epidemiology. 2007; 166:832–840. [PubMed: 17670910]

16. Freedson PS, Lyden K, Kozey-Keadle S, Staudenmayer J. Evaluation of artificial neural network algorithms for predicting METs and activity type from accelerometer data: validation on an independent sample. Journal of Applied Physiology. 2011; 111:1804–1812. [PubMed: 21885802]

17. Healy GN, Clark B, Winkler EAH, Gardiner PA, Brown WJ, Matthews CE. Measurement of Adults' Sedentary Time in Population-Based Studies. American Journal of Preventive Medicine. 2011 in press.

18. Hustvedt BE, Christophersen A, Johnsen LR, Tomten H, McNeill G, Haggarty P, Lovo A. Description and validation of the ActiReg: a novel instrument to measure physical activity and energy expenditure. Br. J. Nutr. 2004; 92:1001–1008. [PubMed: 15613263]

19. Jacobs D, Ainsworth B, Hartman T, Leon A. A simultaneous evaluation of 10 commonly used physical activity questionnaires. Medicine & Science in Sports & Exercise. 1993; 25:81–91. [PubMed: 8423759]

20. Johannsen DL, Calabro MA, Stewart J, Franke W, Rood JC, Welk GJ. Accuracy of Armband Monitors for Measuring Daily Energy Expenditure in Healthy Adults. [Miscellaneous Article]. Medicine & Science in Sports & Exercise. 2010; 42:2134–2140. [PubMed: 20386334]

21. Kipnis V, Midthune D, Freedman LS, Bingham S, Schatzkin A, Subar A, Carroll RJ. Empirical evidence of correlated biases in dietary assessment instruments and its implications. Am. J. Epidemiol. 2001; 153:394–403. [PubMed: 11207158]

22. Kipnis V, Subar AF, Midthune D, Freedman LS, Ballard-Barbash R, Troiano RP, Bingham S, Schoeller DA, Schatzkin A, Carroll RJ. Structure of Dietary Measurement Error: Results of the OPEN Biomarker Study. American Journal of Epidemiology. 2003; 158:14–21. [PubMed: 12835281]

23. Kozey-Keadle S, Libertine A, Lyden K, Staudenmayer J, Freedson P. Validation of Wearable Monitors for Assessing Sedentary Behavior. Medicine & Science in Sports & Exercise. Publish Ahead of Print, 9000.

24. Levine JA, McCrady SK, Lanningham-Foster L, Kane PH, Foster RC, Manohar CU. The role of free-living daily walking in human weight gain and obesity. Diabetes. 2008; 57:548–554. [PubMed: 18003759]

25. Levine JA, Lanningham-Foster LM, McCrady SK, Krizan AC, Olson LR, Kane PH, Jensen MD, Clark MM. Interindividual Variation in Posture Allocation: Possible Role in Human Obesity. Science. 2005; 307:584–586. [PubMed: 15681386]

26. Manini TM, Everhart JE, Patel KV, Schoeller DA, Colbert LH, Visser M, Tylavsky F, Bauer DC, Goodpaster BH, Harris TB. Daily Activity Energy Expenditure and Mortality Among Older Adults. JAMA: The Journal of the American Medical Association. 2006; 296:171–179. [PubMed: 16835422]

27. Matthews, CE. Techniques for Physical Activity Assessment: Self-Report Instruments. In: Welk, G.; Dale, D., editors. Physical Activity Assessments for Health-Related Research. Human Kinetics; Champaign, IL: 2002. p. 107-123.

28. Matthews CE. Calibration of Accelerometer Output for Adults. Med Sci Sports Exerc. 2005; 37:S512–S522. [PubMed: 16294114]

29. Montoye HJ. Introduction: evaluation of some measurements of physical activity and energy expenditure. Medicine & Science in Sports & Exercise. 2000; 32:S439–S441. [PubMed: 10993412]

30. Moore SC, Gierach GL, Schatzkin A, Matthews CE. Physical activity, sedentary behaviours, and the prevention of endometrial cancer. Br J Cancer. 2010; 103:933–938. [PubMed: 20877336]

31. Neilson HK, Robson PJ, Friedenreich CM, Csizmadi I. Estimating activity energy expenditure: how valid are physical activity questionnaires? American Journal of Clinical Nutrition. 2008; 87:279–291. [PubMed: 18258615]

32. Nusser SM, Beyler NK, Welk GJ, Carriquiry AL, Fuller WA, King BMN. Modeling Errors in Physical Activity Recall Data. Journal of Physical Activity and Health. 2012; 9:S56–S67. [PubMed: 22287449]

33. Owen N, Healy GN, Matthews CE, Dunstan DW. Too Much Sitting: The Population Health-Science of Sedentary Behavior. Exercise & Sport Sciences Reviews. 2010; 38:105–113. [PubMed: 20577058]

34. Passmore R, Durnin JVG. Human Energy Expenditure. Physiolgical Reviews. 1955; 35:801–840.

35. Philippaerts RM, Westerterp KR, Lefevre J. Doubly labelled water validation of three physical activity questionnaires. International Journal of Sports Medicine. 1999; 20:284–9. [PubMed: 10452224]

36. Physical Activity Guidelines Advisory Committee. Physical Activity Guidelines Advisory Committee Report. U.S.Department of Health and Human Services; Washington, DC: 2008.

37. Sallis JF, Saelens BE. Assessment of Physical Activity by Self-Report: Limitations and Future Directions. Research Quarterly for Exercise & Sport. 2000; 71:S1–S14. [PubMed: 10925819]

38. Schatzkin A, Kipnis V, Carroll RJ, Midthune D, Subar AF, Bingham S, Schoeller DA, Troiano RP, Freedman LS. A comparison of a food frequency questionnaire with a 24-hour recall for use in an epidemiological cohort study: results from the biomarker-based Observing Protein and Energy Nutrition (OPEN) study. Int. J Epidemiol. 2003; 32:1054–1062. [PubMed: 14681273]

39. Schatzkin A, Subar AF, Moore S, Park Y, Potischman N, Thompson FE, Leitzmann M, Hollenbeck A, Morrissey KG, Kipnis V. Observational Epidemiologic Studies of Nutrition and Cancer: The Next Generation (with Better Observation). Cancer Epidemiology Biomarkers & Prevention. 2009; 18:1026–1032.

40. Schatzkin A, Subar AF, Thompson FE, Harlan LC, Tangrea J, Hollenbeck AR, Hurwitz PE, Coyle L, Schussler N, Michaud DS, Freedman LS, Brown CC, Midthune D, Kipnis V. Design and Serendipity in Establishing a Large Cohort with Wide Dietary Intake Distributions: The National Institutes of Health-American Association of Retired Persons Diet and Health Study. American Journal of Epidemiology. 2001; 154:1119–1125. [PubMed: 11744517]

41. Shephard RJ. Limits to the measurement of habitual physical activity by questionnaires. British Journal of Sports Medicine. 2003; 37:197–206. [PubMed: 12782543]

42. Spiegelman D, Schneeweiss S, McDermott A. Measurement error correction for logistic regression models with an "alloyed gold standard. Am. J. Epidemiol. 1997; 145:184–196. [PubMed: 9006315]

43. Staudenmayer J, Pober D, Crouter SE, Bassett DR, Freedson P. An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer. Journal of Applied Physiology. 2009; 00465

44. van der Ploeg HP, Merom D, Chau JY, Bittman M, Trost SG, Bauman AE. Advances in Population Surveillance for Physical Activity and Sedentary Behavior: Reliability and Validity of Time Use Surveys. American Journal of Epidemiology. 2010; 172:1199–1206. [PubMed: 20855469]

45. van Poppel MNM, Chinapaw MJM, Mokkink LB, van Mechelen W, Terwee CB. Physical Activity Questionnaires for Adults: A Systematic Review of Measurement Properties. Sports Medicine. 2010; 40

46. Wong MY, Day NE, Wareham NJ. Measurement error in epidemiology: the design of validation studies II: bivariate situation. Stat. Med. 1999; 18:2831–2845. [PubMed: 10523745]
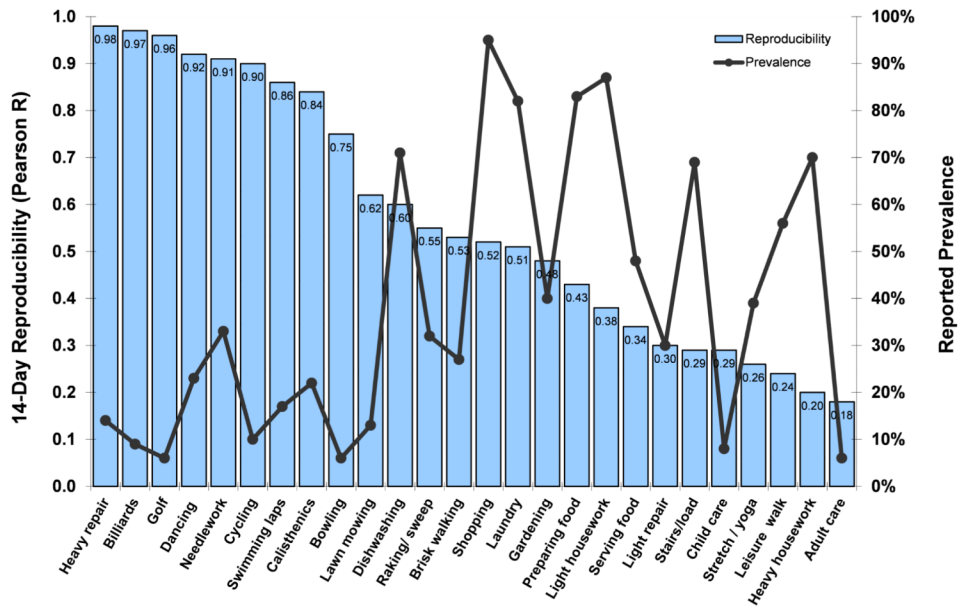
Figure 1.
Reproducibility and prevalence of reporting specific activities on the Yale Physical Activity Survey. Adapted from DiPietro (11).
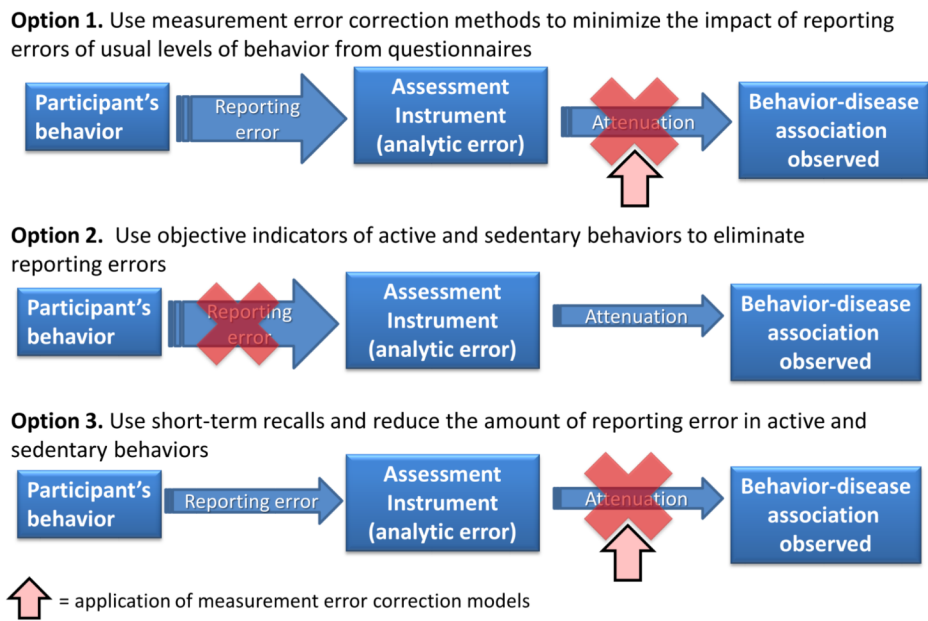
Figure 2.
Options for improving measures of activity-related behaviors and obtaining better estimates of true behavior-disease associations

Men (n=31): 15.6 (0.8) hrs/d   **Women (n=38):** 15.3 (1.2) hrs/d
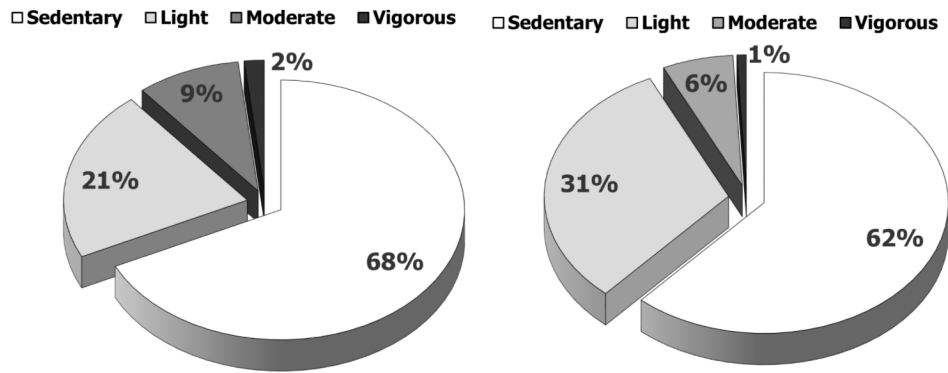

Figure 3.
Allocation of active and sedentary time during waking hours in adults via short-term recalls.
Adapted from Matthews et al. MSSE 37: 986, 2005.

Table 1

Influence of intra-individual variation in behavior and number of replicates on attenuation, bias in observed relative risks, and on statistical power

Intra-individual variation	Replicates	Attenuation (λ)	Naïve relative risks			Statistical Power (α=0.01)		
			Large effect (RR=4.0)	Medium effect (RR=2.5)	Small effect (RR=1.5)	Large effect (RR=4.0)	Medium effect (RR=2.5)	Small effect (RR=1.5)
80%	1	0.20	1.32	1.20	1.08	0.39	0.16	0.03
	2	0.33	1.58	1.35	1.14	0.65	0.29	0.05
	3	0.43	1.82	1.48	1.19	0.79	0.39	0.06
	4	0.50	2.00	1.58	1.22	0.86	0.46	0.07
	5	0.56	2.17	1.67	1.25	0.90	0.52	0.08
	6	0.60	2.30	1.73	1.28	0.92	0.56	0.09
70%	1	0.30	1.52	1.32	1.13	0.60	0.26	0.04
	2	0.46	1.89	1.52	1.21	0.82	0.42	0.07
	3	0.56	2.17	1.67	1.25	0.90	0.52	0.08
	4	0.63	2.39	1.78	1.29	0.93	0.58	0.09
	5	0.68	2.57	1.86	1.32	0.95	0.63	0.10
	6	0.72	2.71	1.93	1.34	0.96	0.66	0.11
60%	1	0.40	1.74	1.44	1.18	0.75	0.36	0.06
	2	0.57	2.20	1.69	1.26	0.91	0.53	0.08
	3	0.67	2.53	1.85	1.31	0.95	0.61	0.10
	4	0.73	2.75	1.95	1.34	0.97	0.66	0.11
	5	0.77	2.91	2.02	1.37	0.97	0.69	0.12
	6	0.80	3.03	2.08	1.38	0.98	0.71	0.12

$\lambda = 1/1 + (\sigma_\varepsilon^2/n/\sigma_T^2)$, adaptation of Equation 2 where n is the number of replicates

Attenuated relative risks calculated as: True RR^λ.

Thus, a true RR of 4.0 with an attenuation factor of $0.5 = 4.0^{0.5} = 2.0$.

Statistical power is calculated for N=100 subjects for each effect size and α = 0.01