Articles in PresS. J Appl Physiol (September 1, 2011). doi:10.1152/japplphysiol.00309.2011

Evaluation of artificial neural network algorithms for predicting METs and activity type from accelerometer data: Validation on an independent sample

Patty S. Freedson[1]
Kate Lyden[1]
Sarah Kozey-Keadle[1]
John Staudenmayer[2]

[1]Department of Kinesiology, University of Massachusetts, Amherst, MA
[2]Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA

Send all correspondence to:
Patty Freedson, Ph.D.
Department of Kinesiology
Totman Building
30 Eastman Lane
Amherst, MA 01003
psf@kin.umass.edu
phone: 413.545.2620
fax: 413.545.2906

Running head: Physical activity machine learning algorithms

Abstract

Previous work from our laboratory provided a 'proof of concept' for use of artificial neural networks (nnets) to estimate METs and identify activity type from accelerometer data. The purpose of this study was to develop new nnets based on a larger more diverse training dataset and apply these nnet prediction models to an independent sample to evaluate the robustness and flexibility of this machine learning modeling technique. The nnet training dataset (UMass) included 277 participants who each completed 11 activities. The independent validation sample (n =65) (UTenn) completed one of three activity routines. Criterion measures were: a) measured METs assessed using open circuit indirect calorimetry b) observed activity to identify activity type. The nnet input variables included five accelerometer count distribution features and the lag one autocorrelation. The bias and root mean square errors (rmse) for the nnetMET trained on UMass and applied to UTenn were + 0.32 METsand 1.90 METs, respectively. Seventy seven percent of the activities were correctly classified as sedentary/light, moderate or vigorous intensity. For activity type, household and locomotion activities were correctly classified by the nnetACT98.1% and 89.5% of the time respectively, and sport was correctly classified 23.7% of the time. Use of this machine learning technique operates reasonably well when applied to an independent sample. We propose the creation of an open access activity dictionary including accelerometer data from a broad array of activities leading to further improvements in prediction accuracy for METs, activity intensity and activity type.

Key words: physical activity, wearable activity monitors, intelligent prediction models

Introduction

Accelerometer sensors are popular tools to estimate physical activity behavior. The devices are easy to use and impose nominal subject and researcher burden. These sensors provide objective estimates about physical activity (PA) features such as point estimates of energy expenditure (EE) and categorically defined activity intensity levels. Despite their popularity, the traditional regression methods used to translate accelerometer output to estimates of energy expenditure or time spent in different activity intensity levels remain problematic. For example, traditional regression approaches are not accurate across a range of activity types and intensities (3,7,14,22), and although they often produce relatively small or non-significant mean differences between estimated and actual EE, the individual estimation errors are often substantial (4, 15). Recent advances in motion sensor technology permit accelerometers to capture and store more detailed information than originally possible, leading several groups to explore more advanced data processing methods, such as hidden Markov models (HMM) (18), decision trees (3), cross-sectional time series (5), multivariate adaptive regression splines (5)and artificial neural networks (nnet) (21,23).

Hidden Markov models, decision trees, and nnets are adaptive machine learning systems capable of 'learning' the shape of complex data. When applied to accelerometer output these machine learning methods do not assume a simple parametric relationship (e.g. linear, exponential, cubic) between accelerometer counts and energy expenditure. This inherent flexibility allows such techniques to use more information from the acceleration signal than the counts$\cdot$min^{-1} used in the traditional regression approaches. These two factors suggest machine learning approaches will improve estimates of accelerometer-based PA metrics across a range of activity types and intensities when applied to large diverse samples. These methods also allow us to identify activity type which is not possible with simple regression methods. A review of several different machine learning activity classification methods and algorithms can be found in a review by Preece and colleagues (19).

Our group and others previously reported success in applying hidden Markov models to identify specific modes of activity (6, 13, 18). The HMM method is relatively complex and relies on custom software that may be a barrier for many applied researchers. Our group (23) and de Vries et al (8) have used nnet models to successfully identify different activity types (23) and specific activities (8). Rothney et al. (21) developed annnetusing raw acceleration input features that improves EE estimates compared to traditional regression techniques. This approach is promising, but at present, it requires expensive analytical software (Matlab, Mathworks, Cambridge, MA) and a very complex multiple accelerometer system (Intelligent Device for Energy Expenditure and Activity (IDEEA), MiniSun LLC, Fresno, CA). Thus, its application to free-living environments and large-scale epidemiologic studies remains impractical. De Vries et al (8) used nnet models from one or two Actigraph accelerometers positioned on the hip and wrist to successfully identify activity type. However, theirnnets do not predict energy expenditure which is of interest to the research community.

Our group recently published a proof of concept paper for two nnet's using the Actigraph 7164. One nnet estimated METs, and another nnet identified activity type (22). Our model improved MET estimates compared to three traditional regression approaches (7, 9, 24) and successfully differentiated activity type into four general categories (sedentary, locomotion, lifestyle or vigorous sport). Unique features of our nnet prediction models is that we used a single hip-mounted accelerometer (ActiGraph 7164; ActiGraph, Pensacola, FL) and the open-source computing language and statistics package R (20) to process the data. The ActiGraph is a commonly used activity monitor in the field, and R is a free statistics package, making this model readily accessible to applied researchers without requiring expensive monitors or skills in advanced statistical methods.

Our methodology established that advanced data processing techniques (artificial neural networks) improved accelerometry-based PA measurement without compromising the capacity of applied researchers to implement these tools in the field. However, our original paper was limited in that the nnet's were validated on the same sample (n =48) in which the models were developed (using cross-validation), and we used an ActiGraph accelerometer model (ActiGraph 7164) that is no longer available and is known to produce different output than more recent accelerometer hardware upgrades (e.g. ActiGraph GT1M) (10). Thus, the purpose of this study was to evaluate the robustness and flexibility of the nnet method for processing GT1M accelerometer data to estimate activity METs and activity type on an independent sample.

Methods

Data collection

At both sites participants read and signed an informed consent document that was approved by the Institutional Review Boards at the respective universities. Participants completed a health history questionnaire to ensure eligibility criteria were met.

University of Massachusetts (UMass) Study protocol

The study sample at UMass included 277 participants. The sample was 50.2% female and 17% minorities. The average age was (mean ± SD) 38 ± 12.4 years, and average BMI was 24.6 ± 4.01 $kg \cdot m^{-2}$.

On the day of the testing, participants reported to the laboratory in a 4-hour fasted state having not consumed caffeine nor participated in exercise for the previous 4hours. Participants completed 11 out of 23 activities (each activity was performed for seven minutes continuously with a four minute rest period between activities) that were divided into two sections: treadmill activities and sport/activities of daily living (ADL). Between each activity section, participants rested for 15 minutes to avoid the possibility of the physiological responses elicited by prior activity influencing the responses of the subsequent activity bout. Furthermore, the order of presentation of the activity bouts was balanced across subjects.

The treadmill section consisted of six conditions; three speeds (1.34, 1.56, 2.23 $m \cdot sec^{-1}$) performed at 0% and 3% grade. The ADL portion included five self-paced ADL'swith each activity being performed for seven minutes continuously. All participants ascended and descended stairs and moved a 6.0 kg box from a shelf to the floor 8m away. The additional two ADLs were randomly selected from a menu of common household activities and sport activities using a blocked randomized design to ensure activities were completed equally among age and sex groups. There were 14 possible household and sport activities including sweeping, mopping, gardening, trimming, mowing, raking, dusting, laundry, vacuuming washing dishes, painting, tennis (with a partner), and basketball. A detailed description of the activities and study protocol has been published elsewhere (11).

Oxygen consumption during activities was measured using a portable metabolic system (Oxycon Mobile, Cardinal Health, Yorba Linda, CA). This portable device is a battery-operated, wireless unit that measures breath-by-breath gas exchange. It was secured to the body using a vest similar to a backpack (950 grams). A face mask (Hans Rudolf, Inc., Kansas City, MO) was connected to the flow sensor unit which measured samples of expired air using a microfuel $O2$ sensor and a thermal conductivity $CO2$ sensor. Immediately prior to data collection, and during the break between protocol sections, a two-point (0.2 and 2.0 $L \cdot s^{-1}$) air flow calibration was performed using the automatic flow calibrator, and the gas analyzers were calibrated using a certified gas mixture of 16 % $O2$, 4.01% $CO2$. The system has been shown to be valid for measurement of respiratory gas exchange during exercise (17).

University of Tennessee (UTenn) Study Protocol

The validation sample was from the University of Tennessee (n = 65) (58% female and 38.2% minorities). Of the 68 participants who completed the protocol, data from 65 participants were included in the analysis. Three participants were excluded due to technical problems in synchronizing the metabolic and accelerometer data. There were 18 different activities in the testing protocol.

The average age of the sample was (mean $\pm$ SD) 40.1 ± 13.0 yrs and average BMI was 27.1 ± 5.61 $kg \cdot m^{-2}$. Age in the UTenn sample was not significantly different from the UMass sample ($p = 0.6064$), and BMI was significantly higher than the mean of 24.6 $kg \cdot m^{-2}$ in the UMass sample ($p = 0.005$). Testing occurred on campus or at the participant's or investigator's home. Participants performed one of three routines, each of which included six different physical activities. For all routines, each activity was performed for 10 minutes with a 3- to 5-minute break between activities.

For routine one (n = 25), participants did laundry including gathering clothes, loading the machines, folding clothes, and putting clothes away. They also ironed, did light cleaning and aerobics. For routine two (n = 22), participants drove through a residential neighborhood, played frisbee golf, trimmed grass using an electric trimmer, gardened and moved dirt with a wheelbarrow. Participants also walked with a 6.8 kg box in their arms, set it down, picked it up, and carried it to another location. For routine three (n = 18), participants played singles tennis and completed self-paced walking and running activities. Distance was recorded to determine speed for each subject in these activities. Participants walked and ran on a track and a road course that included sidewalks, cross-walks, and a slightly hilly terrain. Participants also performed a self-paced walk carrying a 6.8 kg over-the-shoulder laptop computer case. The mean (SD) speeds for the road and track walks were 1.49 (0.18) and 1.52 (0.19) $m{\cdot}sec^{-1}$, 2.70 (0.54) and 2.73 (0.62) $m{\cdot}sec^{-1}$ for the road and track runs and 1.43 (0.17) $m{\cdot}sec^{-1}$ for the walk carrying the computer bag.

The criterion method for measuring oxygen consumption was the CosmedK4b^2 (Cosmed, Rome, Italy) portable metabolic system. The Cosmed K4b^2 is a breath-by-breath gas analysis system consisting of a face-mask, analyzer unit, and battery. Before testing each subject, the unit was warmed up for 45-60 minutes and then calibrated according to the manufacturer's instructions. Calibration of instrument included four parts: room air calibration, reference gas calibration (16.03% O2 and 3.98% CO2), turbine flow-meter calibration with a 3.0 L syringe (Hans-Rudolph), and CO2/O2 analyzer delay calibration with the participant wearing the face mask. To reduce analyzer drift caused by extreme temperatures, the outdoor routines were not performed when the temperature was below 50°F (10°C) (7).

At both UMass and UTenn the ActiGraph GT1M (ActiGraph, Pensacola, FL) accelerometer was used. The device is a small (3.8 x3.7x1.8 cm), lightweight (27grams), uniaxial accelerometer. Detailed specifications of the monitor are published elsewhere (1). Each participant wore an ActiGraph GT1M initialized to collect data in 1-second epochs and secured on the anterior superior iliac spine along the anterior axillary line on the non-dominant hip.

Nnet training and development

For both the development (UMass) and validation datasets (UTenn), we used the identical data cleaning methods as described by Staudenmayer et al. (23). Data points where the coefficient of variation of the counts was greater than 90% different than the mean coefficient of variation for a given activity were eliminated from the final datasets. We removed 16 of 2745 (0.60%) subject/activity combinations for the development group data set (UMass) and 3 of 368 (0.80%) subject/activity combinations for the validation group (UTenn). The accelerometer count features used to develop the nnets were those used in Staudenmayer et al. (22) and included variables representing the signal distribution (10th, 25th, 50th, 75th and 90th percentiles of the second-by-second accelerometer counts) and the temporal dynamics (lag-one autocorrelation). Each subject contributed one set of features for each activity, and those features were calculated from the second-by-second accelerometer counts excluding the first two minutes and last ten seconds of accelerometer data. Each subject performed each activity for seven minutes in the UMass study and 10 minutes in the UTenn study. The METs for each unique subject and activity combination in each study were calculated using the mean measured VO2 ($ml \cdot kg^{-1} \cdot min^{-1}$) divided by 3.5 $ml \cdot kg^{-1} \cdot min^{-1}$, excluding the first two minutes and last 10 seconds of measurements. As in (23), we did not find the inclusion of subject specific characteristics such as age, sex, height, weight, or body mass index to improve the performance of the model.

We developed two nnets: a) a prediction of METs (nnetMET) and b) a prediction of activity type (nnetACT). We used the same neural network technical specifications as (23). Briefly, nnetMET was fit to minimize the penalized squared difference between the criterion MET values and the model's predictions. The penalization was done to avoid over-fitting, and the penalty value was chosen through cross-validation. The nnetACT was fit to minimize the penalized negative logistic likelihood, and the penalty value was again chosen through cross-validation. In the main analyses, we examine the accuracy and precision of the nnetMET trained on UMass data by computing the bias (mean difference between prediction and criterion measure) and root mean squared error (rmse, square root of the mean of the squared differences between the prediction and the criterion measure) of the predictions for the UTenn data. We also compared the nnetMET prediction bias and rmse to the bias andrmse for the Crouter et al. (7) and Freedson et al. (9) regression equations applied to the UTenn data. The Crouter et al. (7) model development was performed with data that were not part of the UTenn validation dataset. We examined activity intensity classification accuracy by comparing the actual intensity classification from the measured METs to those predicted from the nnetMET and Crouter et al. (7) and Freedson et al. (9) equations. We validated activity type categories predicted from the nnetACT trained on UMass and applied to UTenn.

Results

The mean counts$\cdot$min^{-1}, the coefficient of variation for the accelerometer counts$\cdot$min^{-1}, the averages of the signal distribution input features ,the lag 1 autocorrelation input feature and the mean (SD) METs for each activity for UMass and UTenn are shown in Table 1. The measured MET values for the individual activities performed by the development and validation groups are presented in Table 2 (Table 2a: UTenn data; Table 2b: UMass data). Notable is that the range of mean measured METs was 1.88 (washing dishes) to 9.75 METs (treadmill, 2.23 m$\cdot$sec^{-1}, 3% grade) for UMass and 0.78 METs (driving) to 11.17 METs (track running) for UTenn. Additionally, for UMass, four activities (17%) were below 3 METs, 14 activities were between 3.1 and 6 METs and five activities were above 6 METs. In contrast, for UTenn, there were eight activities (44%) below 3 METs.

Table 1 and Table 2 about here

The validation of the nnetMET trained on UMass is shown in Figure 1 (validated on UTenn). The bias was 0.32 METs, the rmse was 1.90 METs, and the correlation between measured METs and the nnetMET was r = 0.78. Eight of the activities where METs were overestimated were in the light intensity range and four activities greater than 6 METs (vigorous) were underestimated. The bias and rmse for the individual activities for the nnetMET are presented in Table 2.We note that this figure suggests that a simple additive measurement error model does not explain the relationship between the nnetMET estimates and the criterion measures. Further exploration of measurement error models for nnetMET is outside the scope of the current work.

Figure 1 about here

For comparison purposes, we applied the Freedson et al (9) and Crouter et al. (7) regression models to UTenn and UMass data (Table 2). For all activities combined (mean measured METs = 4.32), the biases for Freedson et al. (9) and Crouter et al. (7) were -0.95 and 0.18 METs, respectively when applied to the UTenn data (top panel of Table 2). The lowest mean measured METs was 1.88 for the UMass data (lower panel of Table 2: washing dishes) whereas there were three UTenn activities with mean METs below 1 (top panel of Table 2: driving, watching television, and reading). We investigated whether those differences in activity intensity between UMass and UTenn influenced the performance of the nnet by removing the three sedentary behaviors from the UTenn data and re-running the validation analysis. When sedentary behaviors were removed, the bias for the nnet validation was reduced from 0.32 METs to 0.10 METs and increased the nnetvalidation rmse from 1.90 to 1.99 METs (top panel, Table 2). The bias increased to -1.31 METs (from -0.95) for the Freedson et al. (9) equation and decreased to 0.14 METs (from 0.18) for Crouter et al. (7) equation. The rmses increased to 2.26 (from 2.07) and 2.15 METs (from 1.97 METs) for the Freedson et al. (9) and Crouter et al. (7) equations, respectively.

Using UTenn, we examined the activity intensity classification accuracy for nnetMET, and the Freedson et al. (9) and Crouter et al. (7) regression equations. Based on the measured METs, each activity was placed in an activity intensity category (sedentary/light: less than 3 METs, moderate: 3.0 – 5. 99 METs and vigorous: 6.0 METs and above). Predicted METs from the Freedson et al. (9), Crouter et al. (7) regression equations and nnetMETwere directed to the appropriate intensity level classification. The confusion matrices illustrating these analyses are shown in Table 3. The Freedson et al. (9) and the Crouter et al. (7) regression equations correctly classified activity intensity 72.9% and 72.3% of the time. The nnetMET correctly classified activity intensity 77% of the time, and the classification accuracy was relatively constant across intensity categories. The nnetMET classification accuracy is lowest for vigorous activities (71.9%). This is largely due to aerobics which was classified as a vigorous activity (6.2 METs, on average) which was not included in the UMass training data but was in the UTenn validation data.

Table 3 about here

We validated the nnetACT to predict activity type by developing and training the model on UMass data and applying it to UTenn data. We placed the activities into household, locomotion and sport activity categories, and did not include the UTenn sedentary behaviors since the UMass study did not include sedentary behaviors (see Table 4 for assignment of activity type). Table 5presents a confusion matrix illustrating the percentage of activities correctly classified. Application of the nnetACT trained on UMass to UTenn data, yielded an overall correct classification rate of 80.9% (Table 5a). Correct classification occurred for over 98.1% of the household activities, 89.5% of the locomotion activities, and 23.7% of the sports activities. Sport activities were often misclassified as household activities.Correct classification was 97.3% when we applied the nnetACT trained on UMass data to the UMass data (using hold one out cross-validation) (Table 5b). Classification accuracy for the individual activities (UTenn data) for nnetMET, Freedson et al. (9) and Crouter et al. (7) are shown in Table 6. Activity specific classification accuracy ranged from 24% (frisbee golf, Crouter et al (7) regression equation) to 100% for most of the sedentary behaviors for all three prediction models. Household and locomotion activities were correctly classified 95% of the time while sport activities were correctly classified 76.3% ofthe time.

Table 4, Table 5 and Table 6 about here

We also developed and cross- validated nnetMETusing the hold one out cross-validation method. The training of the nnetMET on UMass data and cross-validation on UMass data yielded a bias of 0 METs. In comparison, the biases were -1.26 and -0.84 METs for Freedson et al (9) and Crouter et al (7), respectively (applied to UMass data). The rmses were also higher for Freedson et al (9) and Crouter et al (7) (2.18 and 2.05 METs, respectively) in comparison to the nnet (1.43 METs). Annnet was trained on a combination of the UTenn and UMass data and evaluated with hold one out cross-validation. Bias and rmse were 0.0 METs and 1.2 METs respectively.

Discussion

The primary aim of this study was to advance the nnet methodology for assessing PA metrics using the GT1M Actigraph accelerometer by 1) training the nnet's on a large, diverse sample using broad range of locomotion, lifestyle and sporting activities and 2) validating the nnets on an independent sample. The nnet methodology produced reasonably valid MET estimates, with an overall bias of 0.32 METs and rmse of 1.90 METs, respectively. The nnet also successfully identified activity intensity category 77% of the time and activity type 80.9% of the time. These data are novel in that they move the neural network methodology from a proof of concept (22) to a viable and validated method for processing accelerometer data. An alternative approach for validating METs using a decision tree prediction model was employed by Albinali et al. (3). They successfully predicted activity type from a decision tree algorithm and then used MET values from the Compendium of Physical Activities (2) to predict METs. For model validation, our approach and that used by Albinali et al (3) are both viable options to examine prediction model performance.

We previously demonstrated the nnet methodology success in estimating METs (bias = 0.00 METs, rmse = 1.43 METs) and identifying activity type (88.8% correct) using a 'hold-one out' cross validation technique (23). The nnets were validated on a single observation from the original sample and the remaining observations were used for nnet training. This process was repeated such that each observation from the original sample was used once for validation and the results were then averaged to produce a single estimate of the precision and accuracy of nnet model. When used in calibration studies, cross validation provides an estimate of how well the nnet model will generalize to an independent sample. This approach is not ideal since the validation sample was not truly independent and the activity protocol and research procedures were identical for the cross-validation. Thus researchers should expect that the model will not be as successful when applied to an independent sample that is performing different activities.

In the current study, we again demonstrate the nnets' success using cross validation. Although a primary aim of this paper was to validate the nnets on an independent sample, we present these ancillary results (lower panel, Table 2) to make several points. First, the measurement error reported using cross-validation in this study (bias = 0.00 METs, rmse = 1.43 METs) is similar to previous cross validation results (bias = 0.05 METs, rmse = 1.22 METs) (23). This comparison is interesting because in the current study we used a much larger, more diverse sample and broader range of activities to train the nnets, yet the validity remained comparable to the smaller, less diverse sample results. Accommodation to a broad range of activities performed by a diverse population illustrates the adaptive nature of the nnet method. This inherent flexibility is an improvement over the traditional linear and non-linear regression models that assume simple, rigid relationships between accelerometer counts and energy expenditure. It has been repeatedly documented that traditional regression models do not perform well when applied to diverse samples performing a range of activities (7, 14, 22).

The second reason to present cross-validation results is for comparison to the independent sample validation. The error reported when the nnets are cross validated (bias = 0.00 METs, rmse = 1.43 METs) is less than that reported using the independent sample validation (bias = 0.32 METs, rmse = 1.90 METs). The error range is also narrower for the cross-validation compared to the independent sample validation, indicating the nnet performs better for individual activities (see Table 1). These data show the discrepancies that arise when different validation techniques are used and illustrate the need to validate PA measurement techniques with independent samples. Independent sample validation provides a clearer picture of method robustness.

Figure 1 shows the average measured and predicted METs for each activity when the nnet was trained on UMass and applied to UTenn. The closer an activity is to the line of identity, the better the nnet MET estimate is to the truth. The nnet that was trained on UMass and applied to UTenn tended to overestimate METs (positive bias), but this was not statistically significant overall (see Table 2). This is perhaps because the UTenn study included sedentary activities and the UMass study did not. The UMass nnet returns a MET estimate of 1.98 METs when the counts in a minute are all zero.

In the current study, 12 activities are 'different' between development and validation (track run, road run, aerobics, 15lb bag walk, load/unload boxes, moving dirt, track walk, Frisbee golf, road walk, reading, television, driving [see Table 2]). The average rmse for these activities is 2.25 METs. The average rmse for activities that were 'similar' between UMass and UTenn (ironing, gardening, laundry, light cleaning, trimming, tennis [see Table 2]) is 1.32 METs. It is expected that the absolute errors would be larger for higher MET activities; the activities identified as being 'different' had higher measured METs (mean = 4.66 METs) than the activities identified as being 'similar' (mean = 3.64 METs). We also assessed this difference in terms of percent rmse (measured METs/rmse). Using this approach, activities identified as 'different' had a mean percent rmse of 70.8%, while 'similar' activities had a mean percent rmse of 29.4%. This supports the observation that the error was substantially larger for activities not used in the training dataset. This issue is also discussed by Albinali et al (3) who recommend that 'tuning' machine learning algorithms to individual activities to improve the precision of activity type identification.

In our original study (23) we suggested the nnet improved MET estimates compared to traditional regression approaches. This could not be conclusively confirmed given that the nnet was cross-validated while the traditional regressions were being tested on an independent sample. Table 2 presents the rmse for the nnet method, the Freedson cut-point method (9) and the Crouter two-regression method (7) all using an independent sample for validation. These data support that the nnet improves MET estimates compared to simple regression. Although the improvement in rmse was modest for the nnet in comparison the regression models, across all activities, the nnet had the lowest rmse, 1.90 METs compared to 2.07 METs (Freedson et al. [9]) and 1.97 MET's (Crouter et al. [7]). Both the nnet and the Crouter et al. (7) regression method had slightly positive biases (0.32 and 0.18 METs, respectively) indicating they tend to overestimate MET's on average, while the Freedson et al. (9) regression underestimated METs on average (bias = -0.95 METs). It is not surprising the nnetMET tended to overestimate METs given that no sedentary behaviors were included in the training of the nnetMET. There were three UTenn activities below 1 MET (0.79 – 0.86 METs) where the nnetMET produced a substantial error (% rmse = 149.0-196.2%). For comparison purposes we removed these activities from the analysis and reevaluated the three prediction methods. The rmse was slightly higher (1.99 METs) and bias was reduced to 0.10 METs, respectively (Table 2). These data further illustrate the difficulties of prediction models where activities in the nnet training dataset are not identical to those used in its nnetMET validation.

It is not clear as to why there was only a small improvement in the nnetrmse in comparison the rmse from the regression models. One possible explanation is that there were several activities in the training dataset that were not in the validation datasets. It is also possible that there is a limit to the size of improvements expected, given the finite range of activities performed. To address this knowledge gap, future machine learning model development protocols should include a broad spectrum of activities, across the range of energy expenditure that represent activities performed in daily life.

The Actigraph accelerometers and the current data processing techniques were not designed to measure sedentary behavior. Recently however, researchers have become increasingly interested in understanding the interaction between of sedentary behavior and health. This shift has led to new challenges for the field of PA measurement. The nnets currently available do not identify sedentary behaviors nor accurately estimate sedentary activity METs. Some researchers advocate using an 'inactivity threshold' to identify and assign MET's to sedentarybehaviors . Nonetheless, sedentary behaviors often make up a large portion of an individual's day (16) and thus training the nnet to identify sedentary behaviors is an important next step.

A novel feature of the nnet methodology for measuring PA is the identification of activity type. We categorized activities into household, locomotion and sport activity type categories. Table 4presents a confusion matrix illustrating the percentage of activities the nnetACT correctly classified activities into these categories. The nnetACT trained on UMass correctly classified 80.9% of activities from UTenn. The nnet was successful at identifying household (98.1% correct) and locomotion (89.5% correct) activities in UTenn study. A possible factor contributing to why the classification accuracy was higher for household activities in comparison to locomotion activities was that locomotion activities were treadmill-based in the training dataset and were performed on a track or road in the validation dataset. Nevertheless, classification accuracy was high despite these differences in locomotion protocols. The nnetACT did not perform well for sport activities (23.7% correct) which were often misclassified into the household activities category (69.5% of time). All of the UTenn sports, aerobics, frisbee golf, and tennis were poorly classified. The UTenn aerobics and frisbee golf activities were not included in the UMass study. The third sport in the UTenn study, tennis, was played solo against a wall in the UTenn study while in the UMass study tennis was played with a partner. As the registry of activities used to train nnets expands both in terms of the types and intensities of activities included and the number of samples available for a given activity, improvement in identification of activity type will follow.

A major strength of this study is our use of separate development and validation samples. Although some activities were 'similar', the activity protocols were different between the two sites. Additionally, the overall study procedures and metabolic measurement equipment were different between UMass and UTenn (e.g. at UMass activities were performed for 7-minutes vs. 10-minutes at UTenn, Oxycon Mobile vs. Cosmed K4b^2). By validating the nnet on a completely independent sample we provide researchers with evidence how the nnet will perform when applied to other independent samples.

A second strength of this study is our use of a very large, diverse sample for the training dataset. We also used a wide range of commonly performed locomotion, lifestyle and sport activities. There will always be some level of inter-individual variability in how activities are performed, but training the nnet on a broad range of activity types and intensities and on a sample with a range of physical characteristics, increases the generalizability of the model. Another strength is the use of measured activity energy expenditure to compute METs as the criterion for the nnetMET model validation. Albalini and colleagues (3) used a different approach where raw signals from multiple accelerometers were used in machine learning algorithms to first identify activity type. They then applied the Compendium of Physical Activities (2) to estimate MET levels which produced an underestimate of energy expenditure of 15 – 21 percent.

Our methodology has several limitations. The nnet cannot identify sedentary behaviors. Moving forward, inclusion of sedentary behaviors in the calibration and nnet training process should be a priority. A second limitation is that the results apply only to experimental conditions in a highly controlled laboratory data collection setting. Thus, differences in protocol and criterion measures may alter nnet error estimates. Additionally, the nnet produces PAestimates on a minute-by-minute basis. Free-living behavior does not take place in minute increments; thus in order to apply the nnet to free-living settings, methodology advances need to include analytic procedures for identifying the end of one activity type and the beginning of the next activity type. One possible solution to this problem is to train the nnet to identify individual activity bouts and to then produce PA estimates for specific activities.It should also be noted that the nnet algorithms may only be applied to adults 20 to 60 yrs of age. Future investigations should develop specific nnet algorithms for children and older adults using activities that are relevant in model development and validation for these age groups. We used the fixed denominator of 3.5 ml $\cdot$ $kg^{-1} \cdot min^{-1}$ to compute activity METs. Although baseline RMR is known to be influenced by such factors as age and fat-free mass, we used the standard of 1 MET = 3.5 ml $\cdot kg^{-1} \cdot min^{-1}$ to comply with recommendations for MET computation (2). A limitation of using the constant 3.5 ml $\cdot kg^{-1} \cdot min^{-1}$in the denominator is evident in the UTenn validation dataset with MET values for selected sedentary behaviors (driving, tv viewing, and reading)falling below 1.0 (see Table 2a). As shown in the current study, use of this constant is particularly problematic and may lead to underestimates for computing METs for sedentary behaviors. Although the advantage of using the 3.5 ml $\cdot$ $kg^{-1} \cdot min^{-1}$ constant standardizes the expression of METs, future studies should consider this limitation in light of individual differences in RMR.

Finally, this analysis uses derived activity countsto produce the nnet prediction models. Future studies should employ raw acceleration features as nnet input variables to provide a universal metric for accelerometer sensors output. However, given that currently there is pervasive use of accelerometers employing integrated outputs (e.g. counts$\cdot min^{-1}$),nnets developed from integrated accelerometer signals remain useful.

In summary, we developed and trained nnets to estimate METs, classify activity intensity and identify activity type. We validated these nnets on an independent sample, performing activities that were not identical to the training dataset, and we compared the nnetMET results to regression models. Our nnet produced a lower bias and rmse than the regression models in estimating METs. The intensity classification from the nnetMET was reasonably accurate and we were successful in identifying activity type using the nnetACT for household and locomotion activities. Further advancement of these techniques will require algorithm modification to estimate sedentary behaviors and to identify specific activity bouts under free-living conditions. The nnetMET models only predict absolute intensity prediction and further work is warranted to extend this approach to address relative intensity predictions. We also recommend the development of an open access physical activity registry where accelerometer and metabolic data from a broad array of activities is created. This will facilitate refinement and improvement ofmachine learning algorithms for prediction of activity energy expenditure and activity type identification.

References

1. ActiGraph. *Actisoft analysis software 3.2 User's Manual*. Fort Walton Beach (FL): MTI Health Services, 2005, p.17.

2. Ainsworth BE, Haskell WL, Whitt MC, Irwin ML, Swartz AM, Strath SJ, O'Brien WL, Bassett DR, Schmitz KH, Emplaincourt PO, Jacobs DR, Leon AS. Compendium of physical activities: An update of activity codes and MET intensities. *Med Sci Sports Exerc. Suppl* 32: S498 – S516, 2000.

3. Albinali S,Intille S, Haskell, W, Rosenberger, M. Using wearable activity type detection to improve physical activity energy expenditure estimation, in *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, New York, NY: ACM Press, pp 311-320, 2010.

4. Bassett, DR Jr, Ainsworth BE, Swartz AM, Strath SJ, O'Brien WL, King GA. Validity of four motion sensors in measuring moderate intensity physical activity. *Med Sci Sports Exerc.* 32, *Suppl*: S471-S480, 2000.

5. Butte, NE, Wong, WW, Adolph, AL, Puyau, MR, Vohra, FA, Zakari, IF. Validation of cross-sectional time seres and multivariate adaptive regression splines models for the prediction of energy expenditure in children and adolescents using doubly labeled water. *J Nutr.* 140: 1516-1523, 2010.

6. Chang KH, Chen MY, Canny J. Tracking free-weight exercises. In : *Ubicomp 2007: Ubiquitous Computing*, edited by Krumm J, Abowd GD, Seneviratne A, Strang T. Innesbook, Austria: Springer, 2007, p. 19-37

7. Crouter SE, Clowers KG, Bassett DR Jr. A novel method for using accelerometer data to predict energy expenditure. *J ApplPhysiol* 100:1324-1331, 2006.

8. de Vries SI, Garre FG, Engbers LH, Hildebrandt VH, Van Buuren S. Evaluation of neural networks to identify types of activity using accelerometers. *Med Sci SportsExerc.* 43: 101 – 107, 2011.

9. Freedson P, Melanson E, Sirard J. Calibration of the Computer Sciencee and Applications, Inc. accelerometer. *Med Sci Sports Exerc*30: 777-781, 1998.

10. Kozey SL, Staudenmayer JW, Troiano RP, Freedson PS. Comparison of the Actigraph 7164 and the Actigraph GT1M during self-paced locomotion. *Med Sci Sports Exerc*. 42:971-976, 2010.

11. KozeySL, Lyden K, Howe CA, Staudenmayer JW, Freedson PS. Accelerometer output and MET values of common physical activities. *Med Sci Sports Exerc*. 42: 1776-1784, 2010.

12. Kozey-Keadle SL, Libertine A, Lyden K, Staudenmayer J, Freedson P. Validation of wearable monitors for assessing sedentary behavior. Published ahead of print *Med Sci Sports Exerc*. Jan 12, 2011.

13. Lester J, Choudhury T, Kern N, Borriello G, Hanneford B. A hybrid discriminative/generative approach to recognizing physical activities. In: Proceedings of the 19th International Joint Conferences on Artificial Intelligence, Edinburgh, UK: IJCAI, 2005, p. 766-772.

14. Lyden K, Kozey SL, Staudenmayer JW, Freedson PS. A comprehensive evaluation of commonly used accelerometer energy expenditure and MET prediction equations. *Eur J ApplPhysiol* 111:187-201, 2011.

15. Matthews, CE. Calibration of accelerometer output for adults.*Med Sci Sports and Exerc* 37*Suppl*:S512-S522, 2005.

16. Matthews CE, Chen KY, Freedson PS, Buchowski MS, Beech BM, Pate RR, Troiano RP. Amount of time in sedentary behaviors in the United States, 2003-2004. *Am JEpidemiol* 167: 875-881, 2008.

17. Perret C, Mueller G. Validation of a new portable ergospirometric device (Oxycon Mobile) during exercise. *Int J Sports Med* 27:363 -367, 2006.

18. Pober DM, Staudenmayer J, Raphael C, Freedson PS. Development of novel techniques to classify physical activity mode using accelerometers. *Med Sci SportsExerc* 38: 1626-1634, 2006.

19. Preece SJ, Goulermas JY, Kenney LPJ, Howard D, Meijer K, Crompton R. Activity identification using body-mounted sensors – a review of classification techniques. *PhysiolMeas* 30: R1 – R33, 2009.

20. R Core Development Team. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing, 2009

21. Rothney MP, Neumann M, Beziat A, Chen KY. An artificial neural network model of energy expenditure using nonintegrated acceleration signals. *J ApplPhysiol* 103:1419-1427, 2007.

22. Rothney MP, Schaefer EV, Neumann MM,ChoiL, Chen KY. Validity of physical activity intensity predictions by Actigraph, Actical and RT3. *Obesity* 16: 1946-1952, 2008.

23. Staudenmayer J, Pober, D Crouter S, Bassett D, Freedson P. An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer. *J ApplPhysiol* 107:1300-1307, 2009.

24. Swartz AM, Strath SJ, Bassett DR Jr, O'Brien WL, King GA, Ainsworth BE. Estimation of energy expenditure using CSA acceleromters at hip and wrist sites. *Med Sci Sports Exerc* 32 Suppl: S450-S456, 2000.

Acknowledgements

The authors thank the graduate and undergraduate students for their assistance with data collection and the subjects for their participation.

The authors thank Dr. David Bassett Jr. for providing the University of Tenneessee data for independent sample validation.

Grants

Supported by NIH RO1 CA121005

Figure Legends

Figure 1. Measured METs vs METs predicted from nnetMET. The nnnetMET was developed on UMass dataset (n = 277) and applied to UTenn (n = 65) dataset. The bias was 0.32 METs and the rmse was 1.90 METs.

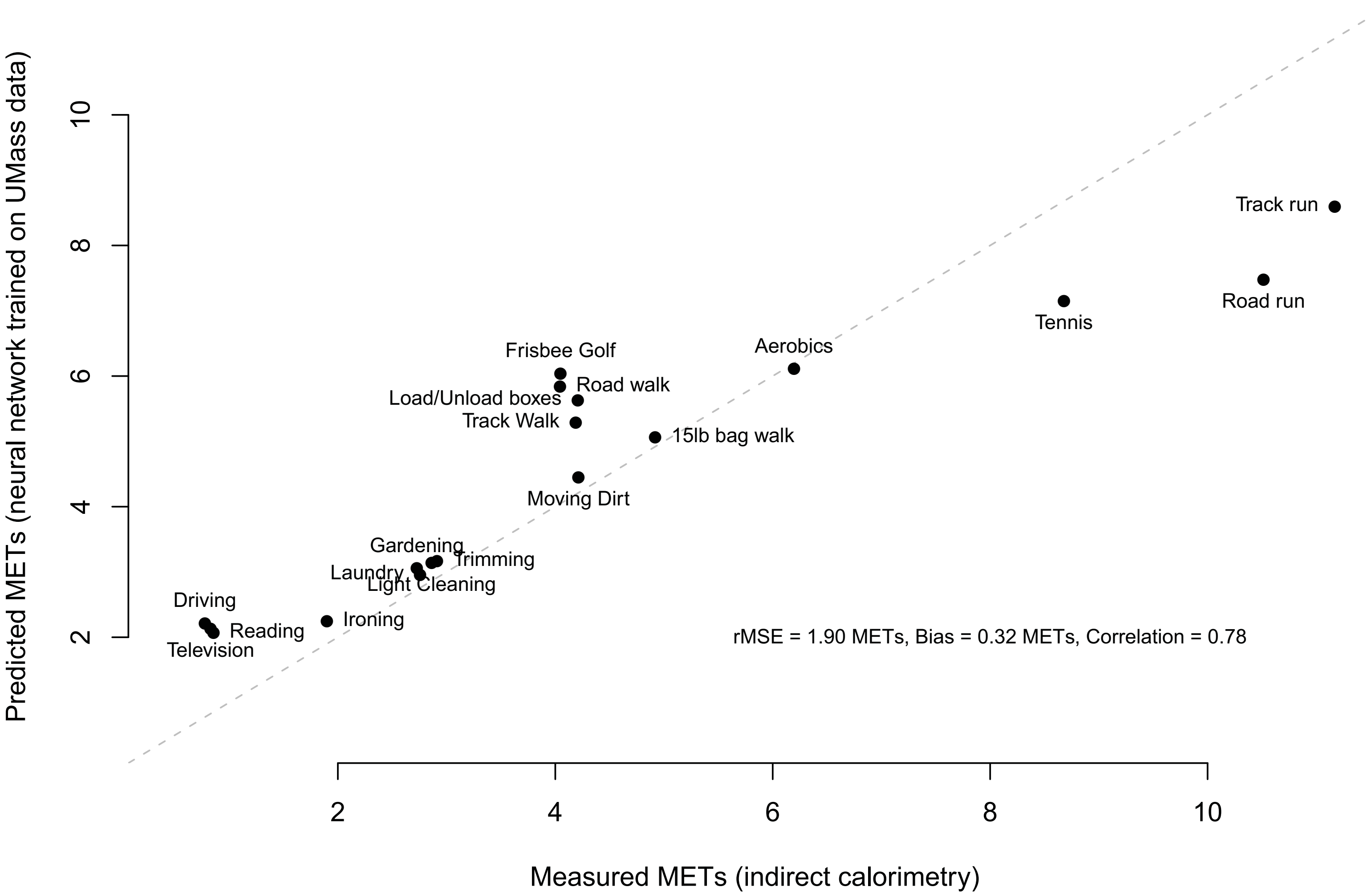
Predicted METs (neural network trained on UMass data)
Measured METs (indirect calorimetry)
Track run
Road run
Tennis
Aerobics
Frisbee Golf
Road walk
Load/Unload boxes
Track Walk
15lb bag walk
Moving Dirt
Gardening
Trimming
Laundry
Light Cleaning
Driving
Ireading
Ironing
Television
rMSE = 1.90 METs, Bias = 0.32 METs, Correlation = 0.78
2
4
6
8
10

Table 1. Descriptive summary for accelerometer output from a) University of Massachusetts and b) University of Tennessee

Table 1a

Activity	Counts per Minute	Coefficient of Variation (per minute)	Percentiles (counts per second over the course of a minute) 10th	25th	50th	75th	90th	Lag One Auto-correlation
Washing Dishes	7	157.8	0	0	0	0	0	0.06
Laundry	144	194.9	0	0	0	1	7	0.41
Dusting	353	174.4	0	0	0	5	18	0.53
Painting	687	99.9	0	0	2	14	38	0.44
Sweeping	548	134.9	0	0	3	10	26	0.51
Trimming	267	144.7	0	0	1	5	13	0.43
Vacuuming	632	83.9	0	1	6	14	26	0.41
Mopping	676	80.5	0	1	7	17	28	0.43
Gardening	1234	98.9	0	1	9	29	58	0.49
Walk 1.34 m/s, 0%	3000	4.4	44	47	50	53	56	0.15
Descending Stairs	3476	9.4	26	46	62	72	80	0.11
Raking	602	74.7	1	2	5	14	26	0.39
Moving Boxes	2149	36.6	2	18	35	45	72	0.5
Walk 1.56 m/s, 0%	3905	4.2	59	62	65	68	71	0.26
Walk 1.34 m/s, 3%	3145	5	46	49	52	56	59	0.18
Cleaning Room	3396	39.1	7	22	48	84	117	0.38
Mowing	2016	26.1	13	23	34	43	51	0.5
Walk 1.56 m/s, 3%	4003	4.4	60	63	67	70	73	0.28
Basketball	4642	23.3	20	40	68	101	143	0.09
Run 2.23 m/s, 0%	7496	5.3	112	118	125	132	138	0.24
Tennis	3412	33.1	15	29	50	77	106	0.36
Ascending Stairs	3041	11	27	40	52	62	71	0.22
Run 2.23 m/s, 3%	7757	4.5	117	123	129	136	142	0.2

Table 1b

Activity	Counts per Minute	Coefficient of Variation (per minute)	Percentiles (counts per second over the course of a minute) 10th	25th	50th	75th	90th	Lag One Auto-correlation
Driving	26	141.3	0	0	0	0	0	0.29
Television	43	46.6	0	0	0	1	2	0.18

Reading	33	27.5	0	0	0	0	2	0.08
Ironing	59	100.9	0	0	0	0	2	0.31
Gardening	690	109.2	0	0	1	12	38	0.55
Laundry	512	131	0	0	1	8	27	0.59
Light cleaning	572	111.2	0	0	1	11	30	0.56
Trimming	470	102.4	0	0	3	10	21	0.49
Road walk	3933	17.2	47	60	68	73	79	0.77
Frisbee golf	3060	36.2	9	29	49	64	89	0.49
Track walk	3987	14.6	56	63	67	72	76	0.57
Load/unload boxes	2380	37.3	9	24	38	49	69	0.42
Moving dirt	1707	53.8	1	7	22	42	64	0.49
15 lb bag walk	3817	22.1	51	55	62	71	79	0.5
Aerobics	2781	58.5	8	17	34	61	106	0.65
Tennis	4385	42.7	15	35	71	102	132	0.57
Road run	6286	21.5	82	98	108	116	122	0.59
Track run	7618	13.2	112	119	129	138	145	0.41

METs

mean	SD
1.88	0.36
2.27	0.36
2.57	0.51
2.9	0.73
3.1	0.62
3.16	0.63
3.24	0.56
3.55	0.76
3.63	1.09
3.8	0.46
3.88	0.78
4.08	1.07
4.52	0.93
4.52	0.55
4.7	0.51
4.79	1.1
5.33	1.02
5.58	0.6
8.33	2.35
8.45	0.92
9.01	1.85
9.62	1.65
9.75	1

METs

mean	SD
0.78	0.17
0.83	0.25

0.86	0.26
1.9	0.45
2.73	0.69
2.76	0.84
2.86	0.68
2.91	0.76
4.04	0.58
4.05	0.49
4.19	0.67
4.21	0.54
4.21	0.93
4.92	1.26
6.2	1.39
8.68	1.56
10.51	2.64
11.17	2.6

Table 2. Measured METs, METs predicted from nnetMET, and nnetMET biases and rmses for independent sample validation (UTenn) and cross-validation

UTN Study

Activity	n (# subjects)	Measured METs	**Biases**: Neural Network trained on MA study data	**Biases**: Freedson Linear Regression	**Biases**: Crouter Two-Equation Method	**Root Mean Squared Errors**: Neural Network trained on MA study data	**Root Mean Squared Errors**: Freedson Linear Regression	**Root Mean Squared Errors**: Crouter Two-Equation Method
Driving (sed)	22	0.78	1.50 $^{+}$	0.68 $^{*+}$	0.42 $^{*+}$	1.52	0.70 *	0.71 *
Television (sed)	23	0.83	1.28 $^{+}$	0.64 $^{*+}$	0.35 $^{*+}$	1.34	0.70 *	0.67 *
Reading (sed)	22	0.86	1.23 $^{+}$	0.61 $^{*+}$	0.25 *	1.28	0.67 *	0.63 *
Ironing	23	1.90	0.38 $^{+}$	-0.41 $^{*+}$	-0.41 $^{*+}$	0.59	0.58	0.79
Gardening	22	2.73	0.40	-0.74 $^{*+}$	0.67 $^{*+}$	0.80	0.98	0.97
Laundry	22	2.76	0.25	-0.91 $^{*+}$	0.31	0.83	1.22	0.96
Light Cleaning	22	2.86	0.27	-0.97 $^{*+}$	0.36	0.76	1.18	0.83
Trimming	21	2.91	0.27	-1.10 $^{*+}$	0.08	0.90	1.29	0.71
Road walk	17	4.04	1.41 $^{+}$	0.52 $^{*+}$	2.12 $^{+}$	1.75	0.84	2.62 *
Frisbee Golf	22	4.05	2.11 $^{+}$	-0.18 *	2.26 $^{+}$	2.38	0.59 *	2.34
Track Walk	17	4.19	0.94 $^{+}$	0.42	1.62 $^{+}$	1.23	0.85	2.19
Load/Unload boxes	22	4.21	1.44 $^{+}$	-0.88 $^{*+}$	1.36 $^{+}$	1.71	1.04	1.54
Moving Dirt	22	4.21	0.35	-1.42 $^{*+}$	0.55	0.86	1.65 *	1.07
15lb bag walk	17	4.92	0.10	-0.45	0.48	2.08	1.90	2.13
Aerobics	20	6.20	-0.46	-2.55 $^{*+}$	-0.29	2.09	2.83	1.28
Tennis	18	8.68	-1.35	-3.75 $^{*+}$	1.60 $^{+}$	2.50	4.07 *	2.54
Road run	18	10.51	-2.94 $^{+}$	-4.08 $^{*+}$	-3.00 $^{+}$	4.45	5.11 *	4.68
Track run	18	11.17	-2.68 $^{+}$	-3.67 $^{*+}$	-3.03 $^{+}$	3.55	4.16 *	4.15
Overall		**4.32**	**0.32**	**-0.95** $^{*+}$	**0.18**	**1.90**	**2.07**	**1.97**
Overall without sedentary activities		**5.02**	**0.10**	**-1.31** $^{*+}$	**0.14**	**1.99**	**2.26**	**2.15** *

(sed)=sedentary activities

*=p<0.05 in a paired comparison with Neural Network.

$^{+}$=<0.05 in comparison with zero (statistically significantly biased). (Activities are Bonferonni corrected.)

UMass Study

Activity	n (# subjects)	Mean Measured METs	**Biases**: Neural Network trained on MA study data (cross-validated)	**Biases**: Freedson Linear Regression	**Biases**: Crouter Two-Equation Method	**Root Mean Squared Errors**: Neural Network trained on MA study data (cross-validated)	**Root Mean Squared Errors**: Freedson Linear Regression	**Root Mean Squared Errors**: Crouter Two-Equation Method
Washing Dishes	42	1.88	0.16	-0.43 $^{*+}$	-0.84 $^{*+}$	0.36	0.56	0.92 *
Doing Laundry	39	2.27	0.24 $^{+}$	-0.72 $^{*+}$	-0.09 *	0.50	0.80 *	0.78
Dusting	36	2.57	0.14	-0.85 $^{*+}$	0.32 $^{*+}$	0.40	0.94 *	0.51
Painting	37	2.90	0.26	-0.92 $^{*+}$	0.47 $^{*+}$	0.73	1.11	0.85
Sweeping	39	3.10	0.06	-1.22 $^{*+}$	-0.03	0.68	1.37 *	0.88
Trimming Bushes	39	3.16	-0.26	-1.51 $^{*+}$	0.69 $^{*+}$	0.61	1.60 *	1.04 *
Vacuuming	35	3.24	0.37 $^{+}$	-1.30 $^{*+}$	0.07 *	0.64	1.40 *	0.57
Mopping	39	3.55	0.06	-1.57 $^{*+}$	-0.20 *	0.77	1.73 *	0.90
Gardening	38	3.63	-0.10	-1.21 $^{*+}$	0.50 $^{*+}$	0.90	1.46 *	0.99
Treadmill 1.34 m/s, 0%	255	3.80	0.69 $^{+}$	0.02 *	-0.19 $^{*+}$	0.93	0.67 *	0.61 *
Descending Stairs	152	3.88	2.12 $^{+}$	0.33 $^{*+}$	0.88 $^{*+}$	2.76	1.10 *	1.80 *
Raking	40	4.08	-0.41 $^{+}$	-2.17 $^{*+}$	-0.83 $^{*+}$	1.01	2.35 *	1.22 *
Moving Boxes	267	4.52	0.08	-1.37 $^{*+}$	0.78 $^{*+}$	0.93	1.62 *	1.24 *
Treadmill 1.56 m/s, 0%	255	4.52	0.35 $^{+}$	0.01 *	-0.42 $^{*+}$	0.88	0.85	0.85
Treadmill 1.34 m/s, 3%	224	4.70	-0.02	-0.76 $^{*+}$	-0.97 $^{*+}$	0.75	1.05 *	1.16 *
Cleaning Room	38	4.79	1.37 $^{+}$	-0.66 $^{*+}$	1.74 $^{+}$	2.14	1.52	2.25
Mowing	38	5.33	0.48	-2.27 $^{*+}$	-0.17 *	1.19	2.43 *	0.95
Treadmill 1.56 m/s, 3%	239	5.58	-0.55 $^{+}$	-0.96 $^{*+}$	-1.39 $^{*+}$	1.06	1.30 *	1.61 *
Basketball	38	8.33	-0.69	-3.20 $^{*+}$	-0.80 $^{+}$	2.27	3.55 *	1.76
Treadmill 2.23 m/s, 0%	213	8.45	0.30 $^{+}$	-1.07 $^{*+}$	-1.52 $^{*+}$	1.21	1.72 *	2.22 *
Tennis	38	9.01	-1.68 $^{+}$	-4.87 $^{*+}$	-2.44 $^{*+}$	2.35	5.08 *	2.83 *
Ascending Stairs	166	9.62	-2.25 $^{+}$	-5.74 $^{*+}$	-5.04 $^{*+}$	2.91	5.92 *	5.34 *
Treadmill 2.23 m/s, 3%	165	9.75	-0.98 $^{+}$	-2.18 $^{*+}$	-2.71 $^{*+}$	1.46	2.52 *	3.12 *
Overall		**4.90**	**0.00**	**-1.26** $^{*+}$	**-0.84** $^{*+}$	**1.43**	**2.18** *	**2.05** *

*=p<0.05 in a paired comparison with Neural Network.

$^{+}$=<0.05 in comparison with zero (statistically significantly biased). (Activities are Bonferonni corrected.)

Table 3. Confusion matrices for intensity category classification comparing criterion measure (measured METs) and a) Freedson et al. (9), b) Crouter et al., (7) and c) nnetMET, (n = 365 subject/activity combinations for validation group [UTenn]).

Rows: Intensity Category from Criterion Measure. Columns: Intensity Category from Prediction Models.

a.

	light	moderate	vigorous	percent correct
light	145	2	0	98.6%
moderate	60	94	0	61.0%
vigorous	5	32	27	42.2%
			overall	**72.9%**

b.

	light	moderate	vigorous	percent correct
light	114	32	1	77.6%
moderate	6	97	51	63.0%
vigorous	4	7	53	82.8%
			overall	**72.3%**

c.

	light	moderate	vigorous	percent correct
light	116	30	1	78.9%
moderate	8	119	27	77.3%
vigorous	4	14	46	71.9%
			overall	**77.0%**

Note: Light intensity is <3 METs, moderate 3-5.99 METs and vigorous ≥ 6 METs.

Table 4. Activities and activity type assignment

UTenn Activities

Activity	Type
Driving	Sedentary
Television	Sedentary
Reading	Sedentary
Ironing	Household
Gardening	Household
Laundry	Household
Light Cleaning	Household
Trimming	Household
Road walk	Locomotion
Frisbee Golf	Sports
Track Walk	Locomotion
Load/Unload boxes	Household
Moving Dirt	Household
15lb bag walk	Locomotion
Aerobics	Sports
Tennis	Sports
Road run	Locomotion
Track run	Locomotion

UMass Activities

Activity	Type
Washing Dishes	Household
Laundry	Household
Dusting	Household
Painting	Household
Sweeping	Household
Trimming	Household
Vacuuming	Household
Mopping	Household
Gardening	Household
Walk 1.34 m/s, 0%	Locomotion
Descending Stairs	Locomotion
Raking	Household
Moving Boxes	Household
Walk 1.56 m/s, 0%	Locomotion
Walk 1.34 m/s, 3%	Locomotion
Cleaning Room	Household
Mowing	Locomotion
Walk 1.56 m/s, 3%	Locomotion
Basketball	Sports
Run 2.23 m/s, 0%	Locomotion
Tennis	Sports
Ascending Stairs	Locomotion
Run 2.23 m/s, 3%	Locomotion

Note: Type - criterion activity type classification

Table 5. Confusion matrices illustrating accuracy of activity type classification

Predicted Activity (nnetACT)

Actual Activity

a.

	Household	**Locomotion**	**Sports**	**Percent correct**
Household	151	2	1	98.1%
Locomotion	9	77	0	89.5%
Sports	41	4	14	23.7%
			Overall	**80.9%**

b.

	Household	**Locomotion**	**Sports**	**Percent correct**
Household	689	23	9	95.6%
Locomotion	11	1640	5	99.0%
Sports	15	3	58	76.3%
			Overall	**97.3%**

Table 5a) Accuracy of activity type identification predicted from the nnetACT trained on UMass and validated on the independent data set from UTenn, (n = 299 subject/activity combinations) 5b) Cross-validation results (trained and validated on UMass using hold-one-out validation), (n =2453 subject/activity combinations) Note: List of activity type assignments is in Table 4.

Table 6: Intensity classification accuracy by activity for nnet, Freedson et al. (9) and Crouter et al. (7).

Activity	nnetMET	Freedson	Crouter
Frisbee Golf	48.0%	95.0%	24.0%
Gardening	55.0%	73.0%	45.0%
Light Cleaning	55.0%	73.0%	55.0%
Aerobics	55.0%	20.0%	70.0%
Trimming	67.0%	48.0%	71.0%
Laundry	68.0%	59.0%	73.0%
Track Walk	71.0%	94.0%	53.0%
Road walk	76.0%	94.0%	41.0%
Tennis	78.0%	17.0%	89.0%
Load/Unload boxes	82.0%	68.0%	64.0%
15lb bag walk	82.0%	94.0%	59.0%
Track run	82.0%	82.0%	94.0%
Road run	83.0%	72.0%	78.0%
Moving Dirt	91.0%	23.0%	86.0%
Ironing	96.0%	100.0%	100.0%
Television	96.0%	100.0%	96.0%
Driving	100.0%	100.0%	100.0%
Reading	100.0%	100.0%	95.0%

Note: Shading identifies activities that were different between training (UMass) and validation (UTenn) datasets. Activities were classified as light (<3METs), moderate (3-5.99 METs) or vigorous (≥6 METs) intensity.