J Appl Physiol 114: 1042–1051, 2013.
First published February 21, 2013; doi:10.1152/japplphysiol.00984.2012.

Impact of study design on development and evaluation of an activity-type classifier

Vincent T. van Hees, Rajna Golubic, Ulf Ekelund, and Søren Brage

Medical Research Council Epidemiology Unit, Institute of Metabolic Science, Cambridge, United Kingdom

Submitted 10 August 2012; accepted in final form 15 February 2013

van Hees VT, Golubic R, Ekelund U, Brage S. Impact of study design on development and evaluation of an activity-type classifier. *J Appl Physiol* 114: 1042–1051, 2013. First published February 21, 2013; doi:10.1152/japplphysiol.00984.2012.—Methods to classify activity types are often evaluated with an experimental protocol involving prescribed physical activities under confined (laboratory) conditions, which may not reflect real-life conditions. The present study aims to evaluate how study design may impact on classifier performance in real life. Twenty-eight healthy participants (21–53 yr) were asked to wear nine triaxial accelerometers while performing 58 activity types selected to simulate activities in real life. For each sensor location, logistic classifiers were trained in subsets of up to 8 activities to distinguish between walking and nonwalking activities and were then evaluated in all 58 activities. Different weighting factors were used to convert the resulting confusion matrices into an estimation of the confusion matrix as would apply in the real-life setting by creating four different real-life scenarios, as well as one traditional laboratory scenario. The sensitivity of a classifier estimated with a traditional laboratory protocol is within the range of estimates derived from real-life scenarios for any body location. The specificity, however, was systematically overestimated by the traditional laboratory scenario. Walking time was systematically overestimated, except for lower back sensor data (range: 7–757%). In conclusion, classifier performance under confined conditions may not accurately reflect classifier performance in real life. Future studies that aim to evaluate activity classification methods are warranted to pay special attention to the representativeness of experimental conditions for real-life conditions.

accelerometry; monitor; physical activity; classification

THE ASSESSMENT OF DAILY PHYSICAL activity in epidemiology requires accurate and precise measurement methodology (29). Wearable sensors hold promise for improving the characterization of most subdimensions of physical activity, including activity type (e.g., walking or standing). Modern accelerometers are small enough to be positioned on nearly every body location. Certain locations may be more feasible, while other sensor locations may result in higher classification accuracy. Only a few studies have attempted to make standardized comparisons between sensor locations (3, 17). In addition, the assessment of classification accuracy is made difficult by lacking a gold standard method for the classification of activity types in people's daily life (27). Direct observation is accurate, but not feasible in daily life settings. However, direct observation is considered feasible and accurate when applied within a confined area and over a confined period of time (2, 22). Most published studies on the development of measurement methods to classify activity types have made use of an experimental protocol based on prescribed physical activities under confined conditions and direct observation as a reference (1, 3, 6, 12, 14, 18, 20, 25, 31). Most of these studies report that the combined usage of acceleration sensors and machine-learning techniques allows for accurate classification of activity types (1, 3, 6, 12, 14, 18, 20, 25, 31).

The use of machine-learning techniques is commonly justified by making reference to its successful application in the fields of face and voice recognition. However, face and voice recognition techniques can be trained with example data with a high degree of resemblance to their real-life applications, while activity type classifiers are typically only trained with experimental data collected under slightly artificial conditions (e.g., exercise laboratory, a corridor, or a car park). Experiments under confined conditions usually involve a fairly limited number of activity types, less variety within each activity type (e.g., intensity, movement pattern, and objects involved), and a relative contribution in time for each activity type that most likely does not reflect normal daily physical activity patterns (12, 25). As a result, the evaluation of activity type classifiers under confined conditions may miss out on movement patterns in real life, leading to inaccurate estimates of classifier performance. A couple of studies concluded that the development of classification algorithms should, therefore, take place under conditions highly comparable to free-living conditions (3, 11, 18). However, to the best of our knowledge, there is no literature proposing a study design that allows for standardized classifier development under free-living conditions. Ruch et al. (22) and Annegarn et al. (2) recently reported on the development of activity classifiers in the daily environment of study participants. In both studies, participants were observed during one or more bouts of 1–3 h in a 1-wk period. Despite the novel effort of capturing real-life data, it remains unknown whether the activities performed during those sections are representative of the entire day or each individual's habitual behavior in general and, by that, whether classifier performance is accurately estimated.

The limitations of testing under confined conditions or using a study design as reported by Ruch et al. (22) may be unavoidable. Nevertheless, improved insight in how limitations of study design impact on classifier performance in real life may help us to better evaluate methods for activity classification. One way to investigate the representativeness of accuracy estimates derived from an experimental protocol under confined conditions for real life is by considering the concepts of "real life" and "confined conditions" on a relative scale. This can then be examined by scaling both concepts in an experiment that can be done under confined conditions with direct observation as a reference.

The aim of this study was to investigate how classifier performance under confined conditions relates to classifier

Address for reprint requests and other correspondence: V. T. van Hees, MRC Epidemiology Unit, Institute of Metabolic Science, Box 285, Addenbrooke's Hospital, CB2 0QQ Cambridge, UK (e-mail: vincent.van-hees@newcastle.ac.uk).

Table 1. *Participant characteristics*

	Men	Women
n	13	15
Handedness (right/left)	12/1	14/1
Weight, kg	72.9 ± 6.5	64.4 ± 17.1
Height, cm	179.3 ± 7.7	165.2 ± 6.5
BMI, kg/m^2	22.7 ± 1.2	23.6 ± 6.3
Arm length, cm	59.2 ± 2.6	56.5 ± 6.7
Leg length, cm	88.8 ± 3.8	80.6 ± 8.5
Age, yr	31 ± 8	29 ± 8

Values are means ± SD; *n*, no. of subjects. BMI, body mass index.

performance in four different real-life scenarios for all commonly used sensor locations on the human body. For the purpose of the present investigation, we considered the performance of a simple two-class classifier of walking.

METHOD

Participants. Twenty-eight healthy adult participants (15 women, 13 men) were recruited by flyers and e-mails in and around the Institute of Metabolic Science at Addenbrookes Hospital in Cambridge, UK (Table 1). The objectives and procedures of the study were explained in detail to the participants, after which they provided written and verbal informed consent. Ethical approval was obtained from the Cambridgeshire Research Ethics Committee, Cambridge, UK.

Experimental design. After the subjects' arrival at the Institute in the early afternoon, the following anthropometric characteristics were assessed: body weight using weighting scale (Tanita, model BC-418, MA; Tanita, Tokyo, Japan), body height using a stadiometer, and leg and arm length using a measurement tape. Arm length and leg length are usually not captured in studies on activity type classification, but may be a useful descriptive when interpreting wrist and ankle acceleration, which, in theory, depend on segment length. These data are not used for the study, as described in this paper. Arm length was assessed from the acromion of the scapulae to the styloid process of the ulna based on stretched arms, where the palm of the hand touched the leg. Leg length was assessed from the top edge of greater trochanter of the femur to the floor (no footwear). Body weight was assessed to the nearest kilogram, while body height, leg length, and arm length were assessed to the nearest 0.5 cm. Handedness was assessed using the Edinburgh handedness questionnaire (16). Next, participants were equipped with nine triaxial raw accelerometers (GENEA, Unilever Discover, Sharnbrook Bedfordshire, UK), as described below. Additionally, four other sensors were positioned on the participant's body, but these data are not included in the present analyses. The sensors included one Actigraph GT3X (Actigraph, Pensacola, FL), one Sensewear (Bodymedia, Pittsburgh, PA), and two ActiWave Cardio monitors (CamNtech, Papworth, UK). None of the monitors obstructed body movement.

The activity protocol included 60 physical activities, which are listed in Table 2, along with their duration (if fixed). The activities were selected aiming for capturing the majority of the activities done by a typical office worker. Occupational activity contributes a large part to total physical activity (9), and, owing to technological advances, most jobs have become sedentary (8). Therefore, an office worker was chosen for this study because he/she is considered representative of the majority of occupationally active with respect to physical activity at work. Activities were structured in nine sections: office work, personal care, shopping and street life, rest and conver-

Table 2. *List of activity types and their duration*

No.	Description	*T*, s	No.	Description	*T*, s
A1	Filling in electronic questionnaire (RPAQ)	*	E31	Standing + drinking	20
A2	Standing	10	E32	Standing + putting butter on slice of bread	30
A3	Sitting on office chair	40	E33	Standing + eat bread with hands	20
A4	Sitting on office chair + mouse task	50	E34	Standing + eat with knife and fork	20
A5	Sitting on office chair + keyboard task	50	E35	Standing + wipe mouth with tissue	15
A6	Sitting on office chair + writing task	50	E36	Walking to lounge area + standing	20
A7	Standing	40	E37	Take a seat and relax	60
A8	Sitting on office chair	40	E38	Sitting + read newspaper	60
A9	Switch chair - sitting on lounge chair	40	E39	Standing	20
A10	Standing	10	F40	Standing	10
B11	Standing	10	F41	Slow walking	40
B12	Standing + combing hair	40	F42	Standing	30
B13	Standing	10	F43	Fast walking	40
B14	Standing + brushing teeth	60	F44	Standing	30
B15	Standing	10	F45	Walking normal speed with phone near head	40
B16	Standing + washing hands	40	F46	Standing	30
B17	Standing + drying hands	40	F47	Walking normal speed	240
B18	Standing	10	F48	Standing	90
B19	Standing or sitting + changing clothes	145	G49	Cycling on ergometer part 1	180
B20	Standing	10	G50	Cycling on ergometer part 2	180
C21	Pushing trolley	*	H51	Ascending stair without holding railing	*
C22	Walking + carrying bag	*	H52	Descending stair without holding railing	*
C23	Walking + unloading bag	*	H53	Ascending stair with holding railing	*
C24	Walking + loading bag	*	H54	Descending stair with holding railing	*
C25	Standing + step through door twice	*	I55	Cycling on pavement slow speed - trial 1	*
D26	Lying quietly on the back	480	I56	Cycling on pavement slow speed - trial 2	*
D27	Sitting + having conversation	240	I57	Cycling on pavement normal speed - trial 1	*
D28	Standing + having conversation	240	I58	Cycling on pavement normal speed - trial 2	*
E29	Standing	10	I59	Cycling on pavement fast speed - trial 1	*
E30	Standing + opening bottle	20	I60	Cycling on pavement fast speed - trial 2	*

T, duration; RPAQ, Recent Physical Activity Questionnaire; A, office work; B, personal care; C, shopping/street life; D, rest and conversations; E, activities in the home; F, indoor walking; G, indoor cycling (ergometer); H, stair walking; I, outdoor cycling. *Duration defined by speed with which activity was performed.

sations, activities in the home, indoor walking, indoor cycling (cycle ergometer), stair walking, and outdoor cycling.

Accelerometers. The triaxial accelerometer (GENEA, Unilever Discover) has been described in detail elsewhere (10, 28). The device uses the STMicroelectronics three-axis LIS3LV02DL acceleration sensor with a dynamic range of ±6 g. The accelerometer weights 17 g (batteries included), and dimensions are 12 × 29 × 37 mm. The acceleration was sampled at 80 Hz, and data for each axis were stored in gram units for offline analyses. Additionally, real-time stamps were stored. The accelerometer was attached to the dorsal side of the wrists and lateral side of the ankles with nylon straps, while an elastic band was used for the waist (hips and lower back), the ventral side of the right upper leg, and the ventral side of the right upper arm. All nine sensors were attached such that they did not obstruct movement and were worn continuously for the entire duration of the experiment.

Procedures. An audio recording was used to provide participants with activity instructions for the sections: office work, personal care, activities in the home, and indoor walking. The audio recording was used to ensure that all participants did the physical activities for the same duration. At the start of each activity, a beep was sounded, and the researcher made a note of the clock time, from which audio instructions and accelerometer data were synchronized. For all other sections of the experiment (cycling, street life, and stair walking), the researcher annotated the time of activity onset and end manually. The duration of these activities was defined by the speed with which they were done.

The questionnaire as used for activity A1 was the electronic version of the Recent Physical Activity Questionnaire (RPAQ) and filled in by computer (4). Questions were predominantly multiple choice and required the participant to use the computer mouse, although a few questions required the participant to use the keyboard. The walking tasks within the section "shopping and street life" and "indoor walking" were done in a corridor (level ground) on a straight trajectory of 20 and 100 m, respectively. Outdoor cycling was done with the participant's own bike at slow, normal, and fast speed. Each speed was done by cycling on a straight 100-m section of pavement (asphalt) outside the institute. The participant was asked to start at one end and cycle to the other end, stop, turn, and cycle back, resulting in two 100-m parts for each speed. The turning points were not standardized and removed from the final analysis. Stair walking involved two flights of stairs with 180° turn in the middle (counterclockwise when going up), with 12 steps per flight (step height 17 cm, step depth 28 cm) and railing positioned at a height of 98 cm on the inner side. To improve future data interpretation and understanding of the study protocol, a 10-min demonstration video was made to visualize all activity types. A copy of the audio files (.mp3) and the demonstration video are available from our website (15).

Feature extraction. Fifty signal features were extracted for each sensor, as described in Table 3. Features were selected to capture as many of the features that have been used in other studies, covering both frequency domain and basic descriptive statistical characteristics. Feature selection was constrained to those features that could be easily calculated and interpreted. Features were calculated for each non-overlapping 2-s window. Signal features relating to the frequency content of the signals (e.g., dominant frequency and entropy) were derived from 4-s time windows with 50% overlap. Window size was chosen to be both long enough to capture the signal signature of walking and short enough to allow for the detection of short-lasting activities. All unlabeled samples and the first and last 3 s of each activity were omitted from further analysis.

Principal component analysis was used to identify the signal features that are most likely valuable for the classifier. Principal component analysis is often used for dimensionality reduction by extraction of the so-called principal components that represent agglomerations of measured data. However, these principal components have no meaning to researchers, and relying solely on principal components would still require the extraction of all signal features; this is undesirable from a computational perspective. To address these limitations, principal component analysis was used to identify the six signal features that contribute most to the principal components. Features were normalized before entering the principal component analysis.

Principal components were derived based on the criterion that the standard deviation of each new principal component needed to be >10% of the standard deviation of the first principle component (parameter "tol" = 0.1 in function "prcom" in R-package "stats"). After derivation of the principal components, weighting factors for all signal features in PC1 (first principal component) were expressed as a ratio of the sum of all weighting factors for PC1 and ordered in decreasing contribution to PC1. Next, a number of signal features were selected to comprise at least 90% of PC1, with a maximum of six signal features. If less than six signal features were needed, the

Table 3. *List of features*

Feature Name	Description
MEAN_BFEN	Mean Euclidean norm of the bandpass-filtered acceleration signals (ω0: 0.2–18, 4th order Butterworth filter)
SD_BFEN	SD of Euclidean norm of the bandpass-filtered acceleration signals (ω0: 0.2–18, 4th order Butterworth filter)
MEAN_LFEN	Mean Euclidean norm of the low-pass-filtered acceleration signals (ω0: 0.5, 4th order Butterworth filter)
MEAN_ENMO	Mean of the Euclidean norm minus 1 g, with negative values replaced by zeros
SD_ENMO	SD of the Euclidean norm minus 1 g, with negative values replaced by zeros
MEAN*x*, MEAN*y*, MEAN*z*	Mean acceleration for each axis
SD*x*, SD*y*, SD*z*	SD for each axis
SK*x*, SK*y*, SK*z*	Skewness for each axis
KU*x*, KU*y*, KU*z*	Kurtosis for each axis
COR*xy*, COR*yz*, COR*xz*	Pairwise correlation between axes; *x-y*, *y-z*, and *x-z*
dfx, dfy, dfz	Dominant frequency using function dfreq (R-package Seewave) with amplitude threshold for signal detection set to 5 (%) and the bandpass frequency filter set to 0.25–10 Hz
ex, ey, ez	Entropy of each axis
fx1, fy1, fz1, fsvm1	Dominant frequency for each axis and the signal vector magnitude (svm)
px1, py1, pz1, psvm1	Power of dominant frequency for each axis and the signal vector magnitude (svm)
fx2, fy2, fz2, fsvm2	Second dominant frequency defined as the frequency with the highest power deviating ≥0.4 Hz from the dominant frequency
px2, py2, pz2, psvm2	Power of second dominant frequency
tpx, tpy, tpz, tpsvm	Total power for the frequencies between 0.3 and 15 Hz
relpx1, relpy1, relpz1, relpsvm1	Power of dominant frequency divided by total power between 0.3 and 15 Hz

svm, Signal magnitude vector; *ex*, *ey*, and *ez* were estimated by using function "csh" from R-package "Seewave"; *dfx*, *dfy*, *dfz* were estimated using function "dfreq" from R-package Seewave; all other frequency domain features were derived from function "spectro" from R-package Seewave.

Table 4. *Contribution of activity types to training data per training model*

		Training Data per Model							
Activity Type	Code	m1	m2	m3	m4	m5	m6	m7	m8
Sitting	3	1/4	1/4	1/4	1/6	1/6	1/6	1/6	1/8
Standing	7	1/4	1/4	1/4	1/6	1/6	1/6	1/6	1/8
Keyboard task	5				1/6				1/8
Washing hands	16					1/6			1/8
Cycling	57 & 58						1/6		
Walking normal speed	47	2/4	1/4	1/4	1/4	1/4	1/4	1/4	1/8
Walking slow speed	41		1/4		1/4	1/4	1/4	1/4	1/8
Walking fast speed	43			1/4					1/8
Stair walking	51 & 52							1/6	1/8

Code = activity code as specified in Table 6. m1–m8, *models 1–8*.

procedure was repeated for PC2 (second principal component) until six signal features were selected. If more than six signal features were needed to comprise the required 90% of PC1, only the six strongest were used.

Classifier training. Logistic regression analysis, based on the selected six signal features, was used to develop binary classifiers to discriminate walking from nonwalking activities. To account for dependence of observations, we used a two-level model with participant ID as repeated-measures indicator (function "lmer" in R-package "lme4").

To simulate the discrepancy between commonly used training data sets and real-life data sets, only a small number of activity types were used for classifier training. It is acknowledged that the selection of these limited number of activity types may affect classifier performance. Therefore, we evaluated eight different configurations of the training data set, including three to eight activity types (see Table 2). This results in the derivation of eight models for every sensor location on the body, each of which was evaluated. To facilitate standardized model comparisons, the amount of data used to train each model was limited to 2 min of data, composed of 50% walking and 50% nonwalking. The activity types within each half had equal contribution to that half (see Table 4).

Classifier evaluation. All 60 activity types were labeled as walking or nonwalking. "Stair walking" was assigned to the category walking. The two activity types, "stepping through a door" (activity code 25) and "walking to the lounging area" (activity code 36), were omitted as they comprised a mixture of both categories with unknown relative contributions. The accuracy of the eight classification models for each body location were evaluated, and the resulting (2 × 2) confusion matrices per activity type were used to derive overall confusion matrices for four hypothetical daily life scenarios (waking hours only) and one traditional laboratory experiment scenario. The conversion of the activity type-specific confusion matrices to real-life scenario-specific confusion matrices was done by the following steps: *1*) summation of activity type-specific confusion matrices across participants; *2*) normalizing each activity type-specific confusion matrix; *3*) multiplication of confusion matrices by weight factors, representing the relative contribution of each activity type to a scenario; and *4*) summation of all matrices to one (2 × 2) confusion matrix per scenario. Weighting factors for all scenarios were chosen based on hypothetical distribution of activity types during waking hours (15 h out of bed + 1 h of lying awake in bed). Sleeping time, assumed to be 8 h, was not included in the scenarios. Time-use surveys from North America and Europe were used as inspiration for creating four hypothetical daily life scenarios (Fig. 1); the surveys cannot be applied directly, as they lack sufficient detail within broad categories, e.g., no distinction between different activity types within the office environment (7, 23, 24, 26). Specifically for this study, scenarios combined a sedentary lifestyle or an active lifestyle with leisure time or office time. The most sedentary lifestyle included 2.3% (22 min) of total walking time, and the most active lifestyle included 13.9% (133 min) of total walking time (Fig. 1). An overview of all weighting factors is provided in Tables 5 and 6. Table 5 shows the contribution of activity types to the four real-life scenarios described in normal language. Table 6 shows the contribution of activity types as measured in the study to simulate the real-life scenarios as shown in Table 5. A fifth scenario represented a traditional experimental protocol for classifier evaluation involving the following 11 activities: sitting (4 min), mouse task (2 min), keyboard task (2 min), writing task (2 min), standing (4 min), lying (6 min), slow walking (4 min), moderate walking (4 min), fast walking (4 min), ascending stairs (2 min), and descending stairs (2 min), amounting to 44% of data being walking.

Classifier performance was evaluated based on sensitivity, specificity, and accuracy, as described by Zhang et al. (31). Additionally, the ratio between estimated time spent in walking and actual time spent in walking was calculated.

RESULTS

Two out of 252 (0.8%) accelerometer recordings failed as a result of initialization faults by the researcher. Principal component analyses resulted in the selection of varying feature sets to be used in the eight training models and for each sensor location (see Table 7). Twenty-nine signal features were not used in any of the models.

Classifier sensitivity estimated with a traditional laboratory scenario (93% on average) falls within the range of sensitivity estimated with the four real-life scenarios for every combination of training model and body location (Fig. 2). Classifier

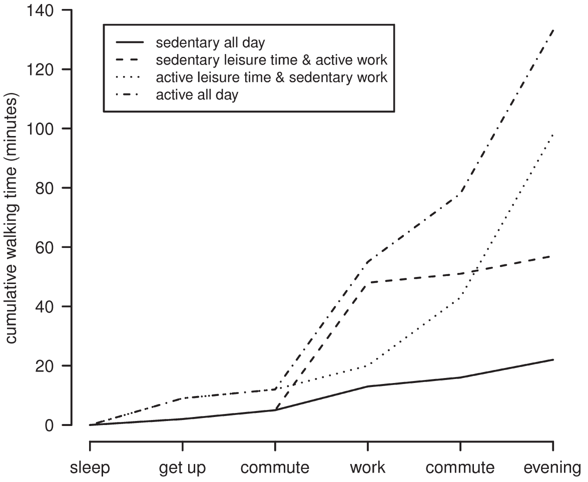

Fig. 1. Simplified representation of hypothetical real-life scenarios with which activity classifiers were evaluated based on laboratory data. See Tables 5 and 6 for full overview.

Table 5. *Contribution of activity types to scenarios in normal language*

	Real-life Scenarios			
Leisure Time + Commuting	S	S	A	A
Work	S	A	S	A
Getting up				
Personal care combing hair	3	3	3	3
Personal care brushing teeth	2	2	2	2
Personal care washing hands	5	5	3	3
Personal care drying hands	5	5	3	3
Personal care changing clothes	10	10	6	6
Personal care standing	5	5	3	3
Breakfast open bottle	0.5	0.5	0.5	0.5
Breakfast drink	5	5	5	5
Breakfast eat put butter on sandwich	4	4	4	4
Breakfast eat sandwich with hand	6	6	6	6
Breakfast wipe mouth with tissue	0.5	0.5	0.5	0.5
Breakfast sit	5	5	5	5
Breakfast sit and read newspaper	0	0	5	5
Walking in the house	2	2	4	4
Stair walking in the house without railing	0	0	2.5	2.5
Stair walking in the house with railing	0	0	2.5	2.5
Computer - questionnaire	4	4	2	2
Computer - mouse	6	6	0	0
Computer - keyboard	6	6	0	0
Computer - write	2	2	0	0
Television	25	25	15	15
Commuting morning				
Walk slow	0.5	0.5	0.5	0.5
Walk fast	0.5	0.5	0.5	0.5
Walk normal	1	1	1	1
Walk with phone	1	1	1	1
Stand	1	1	1	1
Cycle slow	0	0	6	6
Cycle fast	0	0	2	2
Cycle normal	0	0	12	12
Work				
Desk work sitting	140	40	140	40
Desk work mouse	60	30	60	30
Desk work keyboard	123	64	123	64
Desk work writing	20	10	20	10
Desk work questionnaire	60	30	60	30
Meetings sitting	5	80	5	80
Meetings sitting and talking	5	140	5	140
Meetings sitting and writing	4	10	4	10
Meetings in corridor talk	10	10	10	10
Meetings in corridor stand	11	9	11	9
(Lunch)break open bottle	2	2	2	2
(Lunch)break drink	15	10	15	10
(Lunch)break eat sandwich with hand	10	8	10	8
(Lunch)break wipe mouth with tissue	1	1	1	1
(Lunch)break sit	12	6	12	6
(Lunch)break sit and have conversation	15	10	15	10
(Lunch)break sit and read newspaper	7	5	7	5
(Lunch)break walk	2	2	2	2
Walk with phone	1	15	1	15
Walk normal	1	15	1	15
Walk fast	0	3	0	3
Walk slow	1	3	1	3
Washing hands	1	1	1	1
Drying hands	1	1	1	1
Stair walking, holding railing	3	0	3	0
Stair walking, not holding railing	0	5	0	5
Commuting evening				
Walk slow	0.5	0.5	0.5	0.5
Walk fast	0.5	0.5	0.5	0.5
Walk normal	1	1	1	1
Walk with phone	1	1	1	1
Stand	1	1	1	1
Cycle slow	0	0	6	6
Cycle fast	0	0	2	2
Cycle normal	0	0	12	12
Shopping carrying back	0	0	5	5
Shopping loading back	0	0	7	7
Shopping unloading back	0	0	3	3
Shopping pushing trolley	0	0	5	5
Evening				
Meal preparation	0	0	30	30
Diner	30	30	24	24
Diner wipe mouth with tissue	1	1	1	1
Computer - questionnaire	20	20	5	5
Computer - mouse	25	25	15	15
Computer - keyboard	20	20	5	5
Computer - write	5	5	5	5
Computer - sit	55	55	30	30
Television	75	75	40	40
Conversation standing	5	5	5	5
Conversation sitting	28	28	20	20
Personal care brushing teeth	2	2	2	2
Personal care washing hands	1	1	1	1
Personal care drying hands	1	1	1	1
Personal care changing cloths	8	8	8	8
Personal care standing	4	4	3	3
Leisure time walk slow	2	2	5	5
Leisure time walk fast	0	0	1	1
Leisure time walk normal	2	2	39	39
Leisure time walk normal + phone	1	1	5	5
Stair walking, not holding railing	0	0	3	3
Stair walking, holding railing	1	1	2	2
Bed time				
Lying in bed	60	60	60	60
Sleeping	480	480	480	480

Values are in min. S, sedentary; A, active.

specificity, however, was systematically overestimated by the laboratory scenario (98 vs. 87–97% on average). The percentage of time being correctly classified (accuracy) as estimated in a traditional laboratory scenario does not fall within the range of the same percentage when derived from the four real-life scenarios for most combinations of training models and sensor locations (Fig. 2). In some cases, the laboratory scenario overestimated the accuracy, especially for ankle and wrist attachment. The laboratory scenario underestimated the accuracy for a few training models for lower back and upper arm attachment (Fig. 2).

The ratio between estimated and actual time spent walking evaluated with a traditional laboratory scenario was systematically closer to 1, compared with the range derived from the four real-life scenarios (Fig. 3). Time spent in walking was overestimated in all four real-life scenarios, for all models, and for all body locations (range in ratio: 1.07–8.57), except for lower back attachment (range in ratio: 0.84–3.32). The median of all ratios across the four real-life scenarios and eight training models ranges from 2.33 (mean: 2.87) for the left wrist to 1.30 (mean: 1.40) for the lower back.

When comparing training models, no consistent pattern was observed between body locations and classifier performance indicators: sensitivity, specificity, accuracy, and the ratio between estimated and reference values (Figs. 2 and 3).

Wrist position models were found with at least 82% accuracy (according to real-life scenarios); upper arm position

Table 6. *Contribution of activity types as used for the study protocol to the five scenarios*

		Scenarios				
	Leisure Time + Commuting	S	S	A	A	
	Work	S	A	S	A	Lab
1	Filling in electronic RPAQ questionnaire	84	54	67	37	0
2	Standing	0	0	0	0	0
3	Sitting on office chair	65.5	20.5	65.5	20.5	4
4	Sitting on office chair + mouse task	91	61	75	45	2
5	Sitting on office chair + keyboard task	149	90	128	69	2
6	Sitting on office chair + writing task	31	27	29	25	2
7	Standing	2.75	2.25	2.75	2.25	4
8	Sitting on office chair	142	129.5	114.6	102.1	0
9	Switch chair–sitting on lounge chair	9.5	42	9.5	42	0
10	Standing	0	0	0	0	0
11	Standing	2.25	2.25	1.5	1.5	0
12	Standing + combing hair	3	3	3	3	0
13	Standing	2.25	2.25	1.5	1.5	0
14	Standing + brushing teeth	4	4	4	4	0
15	Standing	2.25	2.25	1.5	1.5	0
16	Standing + washing hands	7	7	5	5	0
17	Standing + drying hands	7	7	5	5	0
18	Standing	2.25	2.25	1.5	1.5	0
19	Standing or sitting + changing clothes	18	18	14	14	0
20	Standing	0	0	0	0	0
21	Pushing trolley	0	0	5	5	0
22	Walking + carrying bag	0	0	5	5	0
23	Walking + unloading bag	0	0	3	3	0
24	Walking + loading bag	0	0	7	7	0
25	Standing + step through door twice	0	0	0	0	0
26	Lying quietly on the back	60	60	60	60	6
27	Sitting + having conversation	48	178	40	170	0
28	Standing + having conversation	15	15	15	15	0
29	Standing	0	0	0	0	0
30	Standing + opening bottle	2.5	2.5	8.5	8.5	0
31	Standing + drinking	23	18	22.4	17.4	0
32	Standing + putting butter on slice of bread	4	4	16	16	0
33	Standing + eat bread with hands	16	14	16	14	0
34	Standing + eat with knife and fork	15	15	12	12	0
35	Standing + wipe mouth with tissue	2.5	2.5	2.5	2.5	0
36	Walking to lounge area + standing	0	0	0	0	0
37	Take a seat and relax	112	106	67	61	0
38	Sitting + read newspaper	7	5	12	10	0
39	Standing	0	0	0	0	0
40	Standing	0	0	0	0	0
41	Slow walking	6	8	10	12	4
42	Standing	3.75	3.25	15.75	15.25	0
43	Fast walking	1	4	2	5	4
44	Standing	3.75	3.25	3.75	3.25	0
45	Walking normal speed with phone near head	4	18	8	22	0
46	Standing	2.75	2.25	2.75	2.25	0
47	Walking normal speed	7	21	45	59	4
48	Standing	0	0	0	0	0
49	Cycling on ergometer part 1	0	0	0	0	0
50	Cycling on ergometer part 2	0	0	0	0	0
51	Stair walking up, not holding railing	1.5	0	4.25	2.75	2
52	Stair walking down, not holding railing	1.5	0	4.25	2.75	2
53	Stair walking up, holding railing	0.5	3	2.25	4.75	0
54	Stair walking down, holding railing	0.5	3	2.25	4.75	0
55	Cycling on pavement slow speed - trial 1	0	0	6	6	0
56	Cycling on pavement slow speed - trial 2	0	0	6	6	0
57	Cycling on pavement normal speed - trial 1	0	0	12	12	0
58	Cycling on pavement normal speed - trial 2	0	0	12	12	0
59	Cycling on pavement fast speed - trial 1	0	0	2	2	0
60	Cycling on pavement fast speed - trial 2	0	0	2	2	0

models were found with at least 80% accuracy; for all other sensor locations, models were found with at least 89% accuracy. Average difference in accuracy between body side across the four scenarios ranged from 4.5% lower for the right body side (*model 7*, ankles, scenario: sedentary leisure time, active work) to 7.2% lower accuracy for the left body side (*model 6*, wrists, most sedentary scenario). The direction of the difference between body sides varied between

Table 7. *Number of times signal features were used across the eight training models stratified per body location*

	Wrist (R)	Wrist (L)	Ankle (R)	Ankle (L)	Hip (R)	Hip (L)	Lower Back	Upper Arm	Upper Leg
MEAN_BFEN	8	8	1	2	6	8	7	8	1
SD_BFEN	3	5	0	0	3	3	4	5	4
MEAN_ENMO	0	0	0	0	0	0	0	1	0
SD_ENMO	5	6	0	0	0	0	0	5	0
SD*x*	1	6	0	0	4	8	4	3	0
SD*y*	0	0	0	0	0	0	1	4	0
ex	0	0	5	4	2	2	0	0	3
ey	1	0	5	5	8	8	8	8	8
ez	0	0	0	2	0	0	0	0	0
tpx	8	8	8	8	8	8	8	6	5
tpy	8	8	0	0	0	0	4	4	0
tpz	8	2	0	0	3	0	2	2	0
tpsvm	6	2	0	0	0	0	0	0	0
px1	0	0	2	2	0	0	0	0	0
py1	0	0	1	0	0	0	0	0	0
pz1	0	0	1	0	0	0	0	0	0
psvm1	0	0	1	2	0	0	0	0	0
px2	0	0	6	4	1	0	4	0	6
py2	0	0	6	6	5	1	0	1	8
pz2	0	1	6	7	1	5	4	0	5
psvm2	0	2	6	6	7	5	2	1	8

R, right; L, left. Signal features that are not listed were not used for any training model.

training models and body locations; no consistent pattern was observed (Figs. 2 and 3).

DISCUSSION

The study presented in this paper suggests that experimental protocols performed under confined conditions (laboratory), which are traditionally used for the evaluation of activity type classification methods, provide only limited insight into the validity of classifiers under real-life conditions. To examine this in a standardized fashion, "real life" was simulated under confined conditions, which was necessary as no gold standard exists for activity-type classification in real life. Classifiers were trained on a fraction (≤8/58) of the activities in the aforementioned simulation of "real life" to mimic the discrepancy between traditional laboratory experiments and real-life data. Classifiers were evaluated in four real-life scenarios and one laboratory scenario simulated from all activities included in this investigation but by different weightings. The laboratory scenario was more complex compared with the training data, which was deemed necessary to improve resemblance to the laboratory scenarios, as often used in the literature.

Eight training models were evaluated, each with different configurations of activity types. The best choice of training data configuration seems to depend on sensor location (see Figs. 2 and 3).

The choice of sensor position is determined by its feasibility and the performance of the classifier in relation to the population and lifestyles one wants to assess. When classifier performance is expressed as accuracy, all body locations tested seem to provide reasonable classification accuracy of walking, however, with most accurate classification achieved with sensor information from hip, lower back, and upper limb (Fig. 2). However, when classifier performance was expressed as the ratio between estimated walking time and actual walking time, none of the models performed particularly well (Fig. 3). A median of 30% overestimation of walking time in the best case (lower back) and a median of 133% overestimation in the worst case (wrist) may be considered unacceptable (Fig. 3). Using the ratio as a measure of classifier performance is only meaningful when calculated over data with realistic proportions of walking time and nonwalking time. Other studies have used receiver operator characteristic curves to assess classifier performance (21, 31). Unweighted receiver operator characteristic curves, like unweighted sensitivity and specificity estimates, do not take into account the proportions of time spent in activity types and the possible effect this has on quantifying the uncertainty in estimating time spent in activity types.

We used a relatively simple classifier in the present analyses. The use of a binary classifier has the advantage that it makes classifier performance indicators easier to interpret, which was an important objective of the present investigation. However, decision trees, neural networks, hidden Markov models, and other classification techniques may well provide more accurate classification (3, 5, 19, 31), but it is generally harder to interpret these more complex classification techniques, e.g., pinpoint why a certain combination of features produces a certain result. We acknowledge, however, that both classifier design and study design are two different challenges and require further attention in future studies. Nonetheless, the relation between classifier sensitivity, classifier specificity, and the overestimation of time spent in short-lasting activities is independent of the underlying classifier design. For example, if an activity is truly performed for 5% of the time, and if the classifier has 98% sensitivity and 98% specificity, then that particular activity will, on average, be estimated to have lasted for 6.8% of the time (5% × 0.98 + 95% × 0.02).

In contrast to some other studies on activity classifier development, the present study included also short-duration activities, which were deemed necessary to capture a wide range of activity types representative of real life and without exhausting the participant. An additional reason for the relative short duration of the activity types was that this study was not designed to include assessment of energy expenditure. The assessment of energy expenditure, as done in other studies (25,

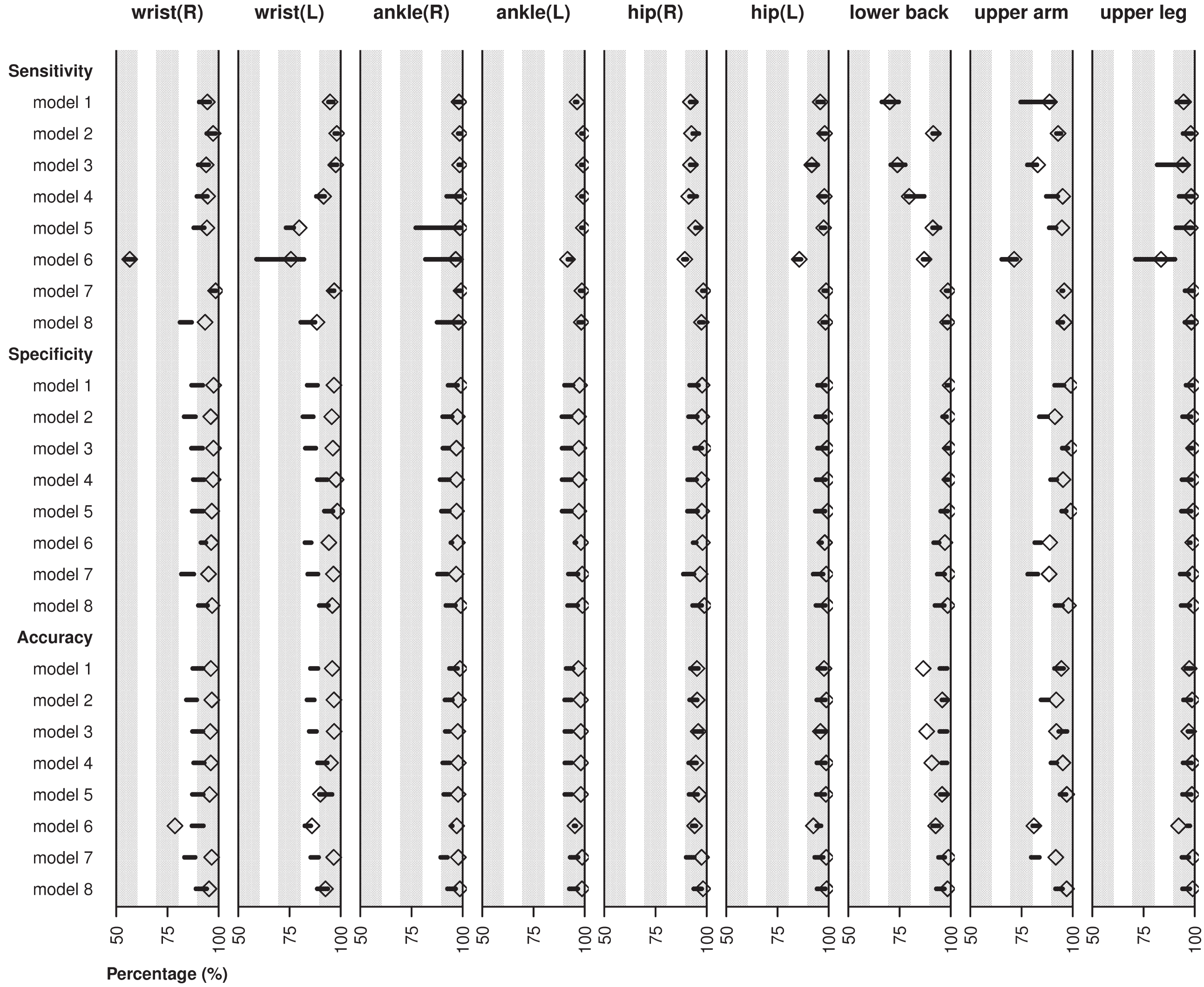

Fig. 2. Classifier performance across training models and body locations. ◇ represent traditional laboratory scenario; horizontal lines reflect range in simulated real-life scenarios. R, right; L, left.

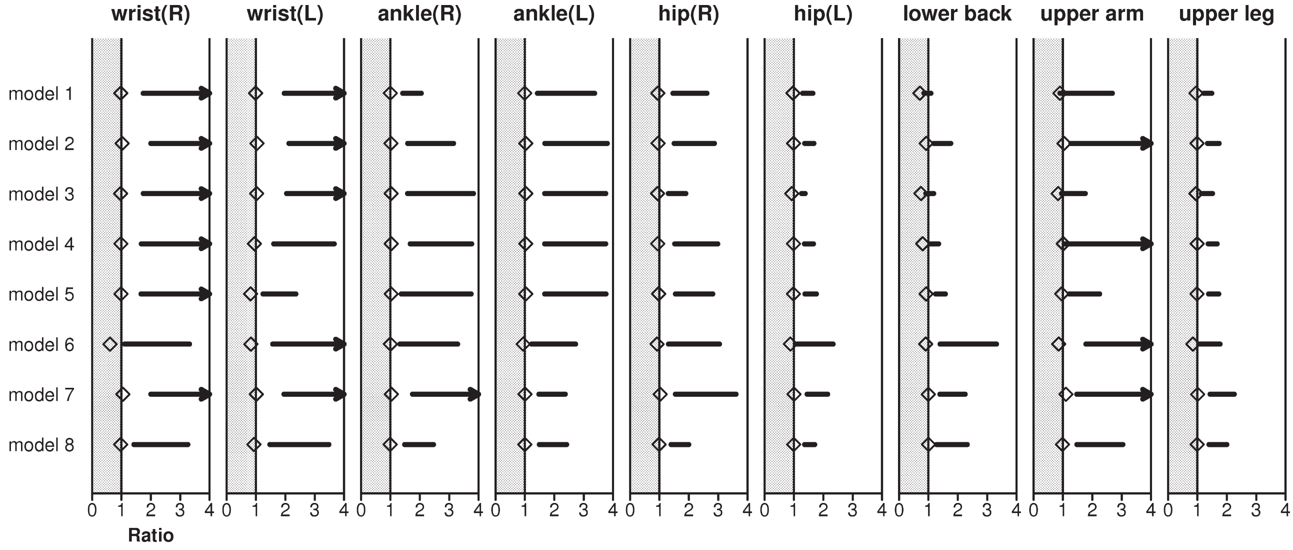

Fig. 3. Ratio between estimated time spent in walking and actual time spent in walking per training model and body location. ◇ represent traditional laboratory scenario; horizontal lines reflect range in simulated real-life scenarios. A range that reaches beyond a ratio of 4 is indicated with an arrow.

31), requires a steady state in oxygen consumption, which can only be achieved when activities are maintained for a certain period of time.

Confusion matrices were averaged across participants to focus on the evaluation of the study design. Between-participant variation would have added another layer to the analysis and complicated the interpretation.

Zhang et al. (31) used a similar accelerometer for classifier development and used a window size of 12.8-s time to capture the frequency content of the signal. Under real-life conditions, however, many walking periods are likely to be shorter than 12.8 s, by which a 4-s time window as applied in the present study seems more realistic, even if this may limit the accuracy with which frequency content can be assessed. Principal component analyses assigned at least one frequency content-based signal feature to each of the models, indicating that a 4-s time window still has value.

Our findings suggest that the future evaluation of activity classifiers should aim for improved representativeness of experimental data sets to data collected in real life, but it remains unclear what makes an experiment more realistic. It may depend on the type of population and their lifestyles (e.g., job type) for which no universal standards exist. An alternative to direct observation as criterion method is asking participants to annotate their own activity types in their daily life, but this may still affect the participant's physical activity and may only work for the less frequent long-duration activity types (e.g., longer walking periods, exercise bouts, and sleep). Yet another alternative is to use another, preferably more accurate, classification system as reference. Gyllensten and Bonomi (13) evaluated their accelerometer-based classifier under real-life conditions by using the IDEEA system as a reference. The IDEEA is a system based on five movement sensors connected by wires, and its accuracy has previously been evaluated under the confined conditions of a room calorimeter involving bouts of treadmill exercise and other prescribed activities using direct observation as a reference (30). It remains unknown whether the IDEEA is valid for real-life applications, as it would depend on the representativeness of the evaluation data used, which is inherent to the problem discussed in the present study.

One way to address this is by establishing multicenter studies to share resources in establishing a repository of annotated accelerometer data involving a wide range of activity types, variations within activity types (e.g., intensity, movement pattern, and objects involved), and variations between individuals. For evaluation of developed classification methods, it would be informative to manipulate the relative weighting of activity types, according to a broad range of real-life scenarios to simulate real-life conditions in a similar manner as the current investigation. Mapping human movement patterns in an exhaustive way and using weighting factors to determine ranges of classifier performances as proposed above may be essential to more thoroughly evaluate the activity classifiers for real-life application. From a practical perspective, we think that this can be achieved by collecting training and evaluation data for more activity types. Here, we suggest a focus on those activity types that, in theory, could pose a challenge to the classifier performance. Second, the use of weighting factors to simulate different activity-type contribution profiles (scenarios) of real life, as done in the present study, is recommended to make results more representative for the diversity in lifestyles across the general population. Scenarios should be seen as a means of identifying the potential weaknesses of the classifier and not per se as proof of validity under real-life conditions: real hard proof of validity under real-life conditions may be an unrealistic goal. Third, we recommend future studies to include analytics like the ratio between estimated and true time spent in an activity type to enhance the description of classifier performance.

No definitive conclusions can be made regarding the best sensor position. The choice of sensor location was only evaluated with one classifier technique and one classification challenge: logistic regression and the detection of walking, respectively. Under these conditions, lower back or hip positioning performed relatively well. Classifier performance would depend on the classification challenge, the sensor position, sensor quality, and the classification technique. Future research is needed to systematically evaluate which combination of components results in best classifier performance.

In conclusion, present approaches to activity type classifier development may be limited in their translation to real-life conditions. Future studies are warranted to quantify the potential impact of study design representativeness on classifier performance in the populations where classifiers are ultimately applied.

ACKNOWLEDGMENTS

We are grateful to study participants for their time and cooperation. We thank Neil Grant from Medical Research Council Visual Aids department for the production of the demonstration video.

GRANTS

The Medical Research Council funded the study (MC_U106179473). V. van Hees was funded by a Biotechnology and Biological Sciences Research Council industry Collaborative Awards in Science and Engineering studentship (Unilever, UK).

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the author(s).

AUTHOR CONTRIBUTIONS

Author contributions: V.T.v.H. and S.B. conception and design of research; V.T.v.H. performed experiments; V.T.v.H. analyzed data; V.T.v.H. interpreted results of experiments; V.T.v.H. prepared figures; V.T.v.H. drafted manuscript; V.T.v.H., R.G., U.E., and S.B. edited and revised manuscript; V.T.v.H., R.G., U.E., and S.B. approved final version of manuscript.

REFERENCES

1. **Allen FR, Ambikairajah E, Lovell NH, Celler BG.** Classification of a known sequence of motions and postures from accelerometry data using adapted Gaussian mixture models. *Physiol Meas* 27: 935–951, 2006.
2. **Annegarn J, Spruit MA, Uszko-Lencer NH, Vanbelle S, Savelberg HH, Schols AM, Wouters EF, Meijer K.** Objective physical activity assessment in patients with chronic organ failure: a validation study of a new single-unit activity monitor. *Arch Phys Med Rehabil* 92: 1852–1857 e1851, 2011.
3. **Bao L, Intille SS.** Activity recognition from user-annotated acceleration data. In: *Pervasive Computing, Second International Conference, PERVASIVE*, edited by Ferscha A and Mattern F. Vienna, Austria: Springer, 2004.
4. **Besson H, Brage S, Jakes RW, Ekelund U, Wareham NJ.** Estimating physical activity energy expenditure, sedentary time, and physical activity intensity by self-report in adults. *Am J Clin Nutr* 91: 106–114, 2010.
5. **Bonomi AG, Goris AH, Yin B, Westerterp KR.** Detection of type, duration, and intensity of physical activity using an accelerometer. *Med Sci Sports Exerc* 41: 1770–1777, 2009.

6. **Bonomi AG, Plasqui G, Goris AH, Westerterp KR.** Improving the assessment of daily energy expenditure by identifying types of physical activity using a single accelerometer. *J Appl Physiol* 107: 655–661, 2009.
7. **Bureau of Labor Statistics, US Dept. of Labor.** *American Time Use Survey* (Online). http://www.bls.gov/tus/current/household.htm [18 December 2011].
8. **Church TS, Thomas DM, Tudor-Locke C, Katzmarzyk PT, Earnest CP, Rodarte RQ, Martin CK, Blair SN, Bouchard C.** Trends over 5 decades in US occupation-related physical activity and their associations with obesity. *PLos One* 6: e19657, 2011.
9. **Csizmadi I, Lo Siou G, Friedenreich CM, Owen N, Robson PJ.** Hours spent and energy expended in physical activity domains: results from the Tomorrow Project cohort in Alberta, Canada. *Int J Behav Nutr Phys Act* 8: 110, 2011.
10. **Esliger DW, Rowlands AV, Hurst TL, Catt M, Murray P, Eston RG.** Validation of the GENEA Accelerometer. *Med Sci Sports Exerc* 43: 1085–1093, 2011.
11. **Foerster F, Fahrenberg J.** Motion pattern and posture: correctly assessed by calibrated accelerometers. *Behav Res Methods Instrum Comput* 32: 450–457, 2000.
12. **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. *J Appl Physiol* 111: 1804–1812, 2011.
13. **Gyllensten IC, Bonomi AG.** Identifying types of physical activity with a single accelerometer: evaluating laboratory-trained algorithms in daily life. *IEEE Trans Biomed Eng* 58: 2656–2663, 2011.
14. **Mathie MJ, Coster ACF, Lovell NH, Celler BG.** Detection of daily physical activities using a triaxial accelerometer. *Med Biol Eng Comput* 41: 296–301, 2003.
15. **Medical Research Council. Epidemiology Unit.** *Physical Activity Epidemiology* (Online). http://www.mrc-epid.cam.ac.uk/Research/Programmes/Programme_5/InDepth/index.html [8 August 2012].
16. **Oldfield RC.** The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* 9: 97–113, 1971.
17. **Parkka J, Ermes M, Antila K, van Gils M, Manttari A, Nieminen H.** Estimating intensity of physical activity: a comparison of wearable accelerometer and gyro sensors and 3 sensor locations. *Conf Proc IEEE Eng Med Biol Soc* 2007: 1511–1514, 2007.
18. **Pober DM, Staudenmayer J, Raphael C, Freedson PS.** Development of novel techniques to classify physical activity mode using accelerometers. *Med Sci Sports Exerc* 38: 1626–1634, 2006.
19. **Preece SJ, Goulermas JY, Kenney LP, Howard D.** A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data. *IEEE Trans Biomed Eng* 56: 871–879, 2009.
20. **Preece SJ, Goulermas JY, Kenney LP, Howard D, Meijer K, Crompton R.** Activity identification using body-mounted sensors–a review of classification techniques. *Physiol Meas* 30: R1–R33, 2009.
21. **Ridgers ND, Salmon J, Ridley K, O'Connell E, Arundell L, Timperio A.** Agreement between activPAL and ActiGraph for assessing children's sedentary time. *Int J Behav Nutr Phys Act* 9: 15, 2012.
22. **Ruch N, Rumo M, Mader U.** Recognition of activities in children by two uniaxial accelerometers in free-living conditions. *Eur J Appl Physiol* 111: 1917–1927, 2011.
23. **Statistics Canada. Government of Canada.** *Overview of the Time Use of Canadians* (Online). http://publications.gc.ca/site/eng/393062/publication.html [18 December 2011].
24. **Statistics Sweden.** *The Swedish Time Use Survey* (Online). http://www.scb.se/Pages/Product____12211.aspx [18 December 2011].
25. **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. *J Appl Physiol* 107: 1300–1307, 2009.
26. **Tudor-Locke C, Leonardi C, Johnson WD, Katzmarzyk PT.** Time spent in physical activity and sedentary behaviors on the working day: the American time use survey. *J Occup Environ Med* 53: 1382–1387, 2011.
27. **van Hees V.** The challenge of assessing physical activity in populations. *Lancet* 380: 1555; author reply 1555–1556, 2012.
28. **van Hees VT, Pias M, Taherian S, Brage S, Ekelund U.** A method to compare new and traditional accelerometry data in physical activity monitoring. In: *2nd IEEE International WoWMoM Workshop on Interdisciplinary Research on E-Health Services and Systems (IREHSS)*. Montreal, Canada: 2010, p. 1–6.
29. **Wareham NJ, Rennie KL.** The assessment of physical activity in individuals and populations: why try to be more precise about how physical activity is assessed? *Int J Obes Relat Metab Disord* 22, *Suppl* 2: S30–S38, 1998.
30. **Zhang K, Werner P, Sun M, Pi-Sunyer FX, Boozer CN.** Measurement of human daily physical activity. *Obes Res* 11: 33–40, 2003.
31. **Zhang S, Rowlands AV, Murray P, Hurst T.** Physical activity classification using the GENEA wrist worn accelerometer. *Med Sci Sports Exerc* 44: 742–748, 2012.