ROBOTS IN ASSISTED
LIVING ENVIRONMENTS

UNOBTRUSIVE, EFFICIENT, RELIABLE AND MODULAR SOLUTIONS FOR INDEPENDENT AGEING

Research Innovation Action

Project Number: 643892 Start Date of Project: 01/04/2015 Duration: 36 months

DELIVERABLE 6.15

Medical evaluation report III

Dissemination Level	**Public**
Due Date of Deliverable	Project Month M36, March 2018
Actual Submission Date	10 May 2018
Work Package	WP6, *Piloting and evaluation*
Task	Task 6.5, *Medical evaluation*
Lead Beneficiary	FHAG
Contributing beneficiaries	NCSR-D, FSL
Type	R
Status	Submitted
Version	Final

Project funded by the European Union’s Horizon 2020 Research and Innovation Actions

Abstract

This deliverable reports the Medical Evaluation of the second round of the *Summative Phase* pilot studies using the final integrated RADIO prototype at FHAG premises. Four (I)ADL methods integrated in the robot were used to recognize: bed transfer, chair transfer, 4 meter walk and pill intake and three (I)ADLs were detected by the use of Smart Home sensors: TV watching, meal preparation and going out of the room. Precision, recall and F-score equivalents, were used for the evaluation of the methods. Correct detections were further analyzed as to their fitness.

History and Contributors

Ver	Date	Description	Contributors
00	12 March 2018	Document structure	NCSR-D
01	3 May 2018	First complete draft	NCSR-D
02	4 May 2018	Comments and corrections in all Sections	FHAG
03	7 May 2018	Concluding remarks Section 5	NCSR-D, FHAG
04	09 May 2018	Internal peer review	FSL
05	10 May 2018	Addresses peer review comments	NCSR-D, FHAG
Fin	10 May 2018	Final preparation and submission	NCSR-D

Executive Summary

This deliverable reports the Medical Evaluation of the second round of the *Summative Phase* of pilot study using the final integrated RADIO prototype at FHAG premises. Four (I)ADL methods integrated in the robot were used to recognize: bed transfer, chair transfer, 4 meter walk and pill intake and three (I)ADLs were detected by the use of Smart Home sensors: TV watching, meal preparation and going out of the room. For each ADL, RADIO system and ground truth measurements were collected. Based on the RADIO system detections, an ADL instance could be either not detected (false negative), wrongly detected (false positive) or correctly detected (true positive). Based on these, precision, recall and F-score of each ADL method were calculated. Correct detections were further analyzed using correlation and linear regression methods, complemented by metrics that exposed the deviations from the ideal 1:1 line. This pilot study aimed to facilitate the medical evaluation of the integrated RADIO prototype as a support platform for ADL and IADL assessment. The evaluation revealed that: a) the bed transfer and the 4m walk detection methods performed well, b) the chair transfer method is less robust to both different environments and to changes in the conditions within the environment and c) the event detection methods (pill intake and ADLs inferred by Smart Home sensors' activity) performed reasonably well, with a small number of undetected events.

Abbreviations and Acronyms

ADL	Activities of Daily Living
IADL	Instrumental Activities of Daily Living
TP	True Positive
FP	False Positive
FN	False Negative
TN	True Negative
MSD	Mean Standard Deviation
SB	Squared Bias
NU	Non-Unity slope
LC	Lack of Correlation

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 Purpose and Scope

The purpose of this document is to report the *medical evaluation* of the final integrated RADIO system that was piloted in the second round of the Summative Phase pilot studies. Specifically, the document reports the methods used to evaluate the data collected and their results. It describes the fitness for purpose of the system, or in other terms, the capacity and the accuracy of the RADIO system to monitor and actually detect the chosen ADLs.

For this medical evaluation only data from the FHAG trials were used. These trials were carried out by having participants visit the same RADIO deployment, so that all variables except those relevant to the participant are fixed.

1.2 Approach

RADIO studies are conducted in three phases:

1. Formative phase; first pilot at FSL
2. Intermediate phase; second pilot of RADIO components at FSL
3. Summative phase; final RADIO pilots

This deliverable is prepared using the data collected during the second round of Summative Phase pilot studies using the final integrated RADIO prototype at FHAG premises. During this phase, participants were monitored with the RADIO system. At the same time, ground truth measurements were collected by FHAG research assistant. This dual assessment generated a variety of summary statistics (recall, precision, and the F-measure) that are useful to evaluate the final prototype of the RADIO system in a real setting. This report is public. The procedures followed (without any reference to the particular subjects or deployments) are documented in public deliverable *D6.4 Piloting plan IV*. The execution of the pilot studies and details about piloting, its outcomes and technical details are reported in *D6.8. Pilot report II*. User evaluation results and the lessons learned from piloting are described in *D6.12 User Evaluation IV*.

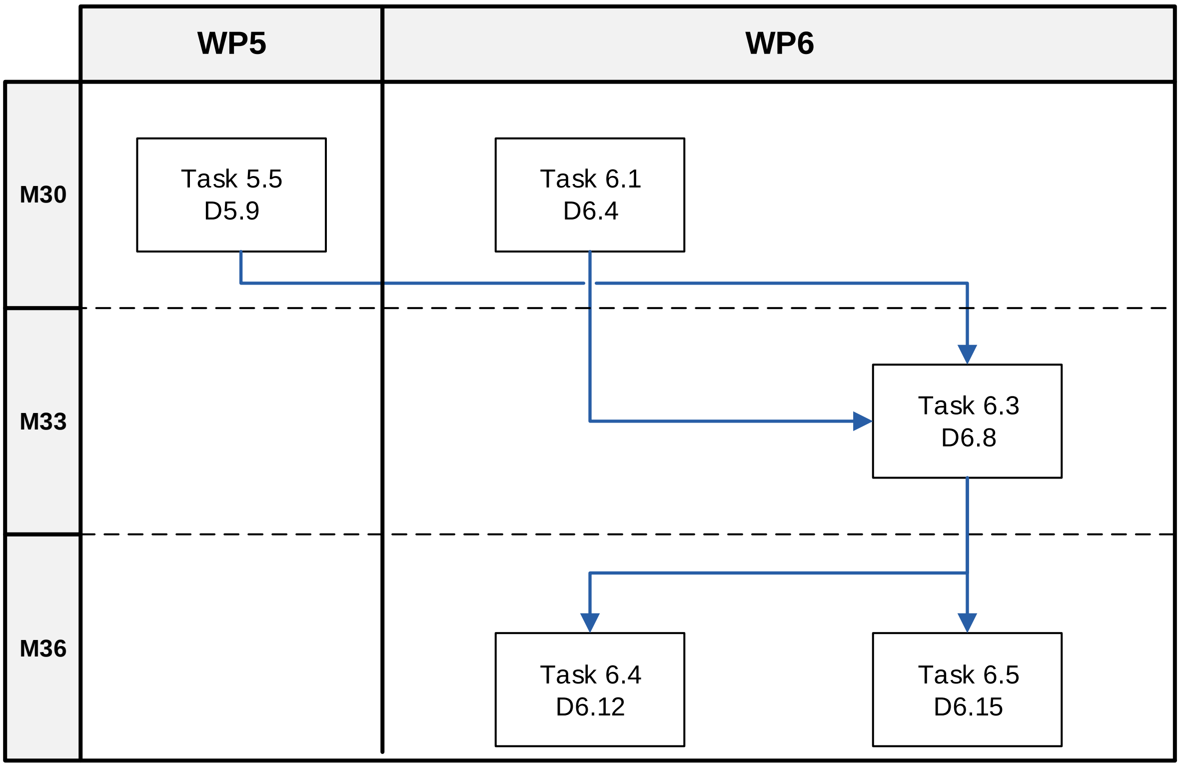


Figure 1. Dependencies between this deliverable and other deliverables.

1.3 Relation to other Work Packages and Deliverables

This document reports the medical evaluation results of the second round of the Summative Phase pilot studies. These trials were executed at FHAG premises during October -December 2017.

The data collected during the second round of the Summative Phase pilot studies were reported in *D6.8 Pilot report II*. These data were analyzed in the context of Task 6.4 and Task 6.5 and were used for user evaluation reported in *D6.12 User Evaluation IV* and for medical evaluation reported in the current document *D6.15 Medical evaluation report III*.

2 METHODS

This section describes the data sets and methods used for the medical evaluation.

2.1 Evaluation dataset

As described in D6.8, sixteen (16) participants completed a 3-day study scenario. During this, each one completed the following repetitions for each of the four ADLs monitored by the RADIO robot:

- Bed transfer (Lying to Standing): ***8 repetitions***
- Chair transfer (Sitting to Standing) ***12 repetitions***
- 4-meter walk: ***16 repetitions***
- Pill intake: ***10 repetitions***

At the end of each participant's scenario, an email was sent to clinical staff informing them about the duration of each detected activity. All sixteen emails were successfully received. However, the **first four participants are excluded** from further analysis due to technical failures of the RADIO system. Moreover, in the *bed transfer* ADL *only*, we excluded from further analysis **five more participants** (both ground truth and RADIO robot data), again due to technical issues unrelated to the bed transfer recording method. The total number of data used for evaluation for each ADL is reported in Section 3.

Together with the RADIO system (robotic platform and smart home sensors) the occurrence of the events or their duration was also collected by FHAG researchers (ground truth). Details about how the ground truth was collected can be found in *D6.4 Pilot Plan IV*.

In summary, the evaluation reported in this document includes data from the RADIO system and their ground truth. For all ADLs, except pill intake and those recorded through Smart Home sensors, there are two kinds of information: detection of the activity and duration of the activity.

2.2 ADL detection analysis

Overall, we characterize monitored ADLs as **detected** when the RADIO system returned an entry for this ADL. In any other case, we refer to them as ***no detections*.**

From the detected instances we will further discriminate between ***correct detections*** and ***wrong (erroneous) detections***. In order to discriminate between correct detections and erroneous ones we assess ***if a RADIO measurement could be overall a realistic measurement for that ADL.*** The exact rule for each case is presented in Table 1.

So overall, in reference to detection we can discriminate three different cases:

- **Correct detection:** the event was successfully recognized compared to researchers' ground truth. Events correctly detected constitute the ***true positives*** in further analysis.
- **Wrong detection:** the event was not successfully recognized compared to the ground truth. In this case, we included instances where an ADL was actually detected but the duration reported implies 'erroneous' detection. The rules based on which we characterized detections as wrong are presented in Table 1. Events wrongly detected constitute the ***false positives*** in further analysis.
- **No detection:** the system failed to recognize the event. Events not detected constitute the ***false negatives*** in further analysis.

Based on these definitions of True Positive (TP), False Positive (FP), and False Negative (FN) values, and consistently with *D2.1 Early Detection methods and relevant system requirements I*, Precision, Sensitivity and F-measure indices were calculated as reported below.

Table 1. ADL data categorization based on detection.

<table>
<tr><th>ADL</th><th>Correct detection</th><th>Wrong detection</th><th>No detection</th></tr>
<tr><td>Bed</td><td rowspan="3">The robot detected an actual event and the value reported is <u>not</u> lower than the min value of ground truth or higher than the max value of ground truth.

min (GT measurement) < RADIO measurement < max (GT measurement)</td><td rowspan="3">The robot detected an actual event and the value reported is lower than the min value of ground truth or higher than the max value of ground truth.

RADIO measurement < min (GT measurement)

AND

RADIO measurement> max (GT measurement)</td><td rowspan="7">The RADIO system did not detect an actually occurring event (no email entry).</td></tr>
<tr><td>Chair</td></tr>
<tr><td>4-meter walk</td></tr>
<tr><td>Medication intake</td><td>The robot detected an actual event.</td><td>N/A</td></tr>
<tr><td>TV watching</td><td>Smart home sensors detected an actual event.</td><td>N/A</td></tr>
<tr><td>Meal Preparation</td><td>Smart home sensors detected an actual event.</td><td>N/A</td></tr>
<tr><td>Going out of the room</td><td>Smart home sensors detected an actual event.</td><td>N/A</td></tr>
</table>

Importantly, no True Negatives (TN) are defined in our case as the calculation of this index implies counting the number of no-events correctly rejected as no-events. Considered the nature of our study, this kind of measure is inapplicable, thus not allowing the calculation of the Accuracy index, being (TP + TN)/(TP + FP + TN + FN).

As for the other indices, these were calculated as follows:

Precision, also known as Positive Predictive Value (PPV), measures the likelihood that a detected event corresponds to an actually occurred event, thus answering the question 'How likely is it that this event occurred given that the test result is positive?' Precision is calculated as follows:

$$\frac{TP}{TP + FP}$$

Sensitivity, also known as recall or true positive rate, measures the percentage of positives that are correctly identified as such (i.e., the percentage of occurred ADLs detected as occurred). It is calculated by the following formula:

$$\frac{TP}{TP + FN}$$

F-measure is defined as the weighted harmonic mean of precision and sensitivity as it combines the precision and recall rates into a single measure of performance, thus resulting in a compromise between the two measures. It is high only when both precision and sensitivity are high. The F-measure assumes values in the interval [0,1]: it is 0 when no actually occurred events have been detected, and is 1 if all detected events are actually occurred and all actually occurred events have been detected.

$$2* \frac{Precision*Sensitivity}{Precision+Sensitivity}$$

2.3 ADL duration measurements

The second part of the evaluation takes into account ***ADLs recognized correctly*** (as defined above) and compares them to ground truth. Ideally, RADIO methods should give identical or almost identical measurements to ground truth. In order to compare ground truth measurements X_n and RADIO measurements Y_n, we produce the scatterplots for each ADL and if correlation is identified we proceed in calculating the linear regression and metrics that inform us about the sources of deviation from the 1:1 line.[1]

Specifically, we calculate:

- the mean standard deviation (MSD) between the ground truth measurements and RADIO
    - $MSD = \frac{\sum(X_n - Y_n)^2}{N}$, where N is the number of correct detections.
- the squared bias (SB) – indicative of translation compare to 1:1 line,
    - SB= $SB = (\bar{X} - \bar{Y})^2$, where $\bar{X}$ and $\bar{Y}$ are the mean values of ground truth measurements and RADIO accordingly.
- non-unity slope (NU) – indicative of rotation compare to 1:1 line,
    - $NU = (1 - b)^2 * \frac{\sum x_n^2}{N}$, where b is the slope of the calculated linear regression and $\frac{\sum x_n^2}{N}$ is the variance of the ground truth measurements.
- lack of correlation (LC) – indicative of scattering, where r is the correlation of the samples and $\frac{\sum y_n^2}{N}$ is the variance of the RADIO measurements.
    - $LC = (1 - r^2) * \frac{\sum y^2}{N}$

[1] Gauch HG, Hwang JT, Fick GW. "Model evaluation by comparison of model-based predictions and measured values." *Agronomy Journal* 95(6):1442-6, 1 Nov 2003.

3 ADL DETECTION

3.1 Bed Transfer

For the bed transfer ADL, we analyzed in total **56 sessions** (7 participants x 8 repetitions *–c.f. Section 2.1*). Of these sessions, the RADIO system *did not detect* the ADL in 15 instances. The robot data for the rest 41 *detected bed transfers* and the corresponding ground truth data are presented in Figure 2. Out of the 41 actually detected bed transfers, 38 can be classified as correct detections (true positives), while 3 are classified as wrong detections (false positives) falling out of the ground truth measurements' interval: min =4.46 and max = 54.52.

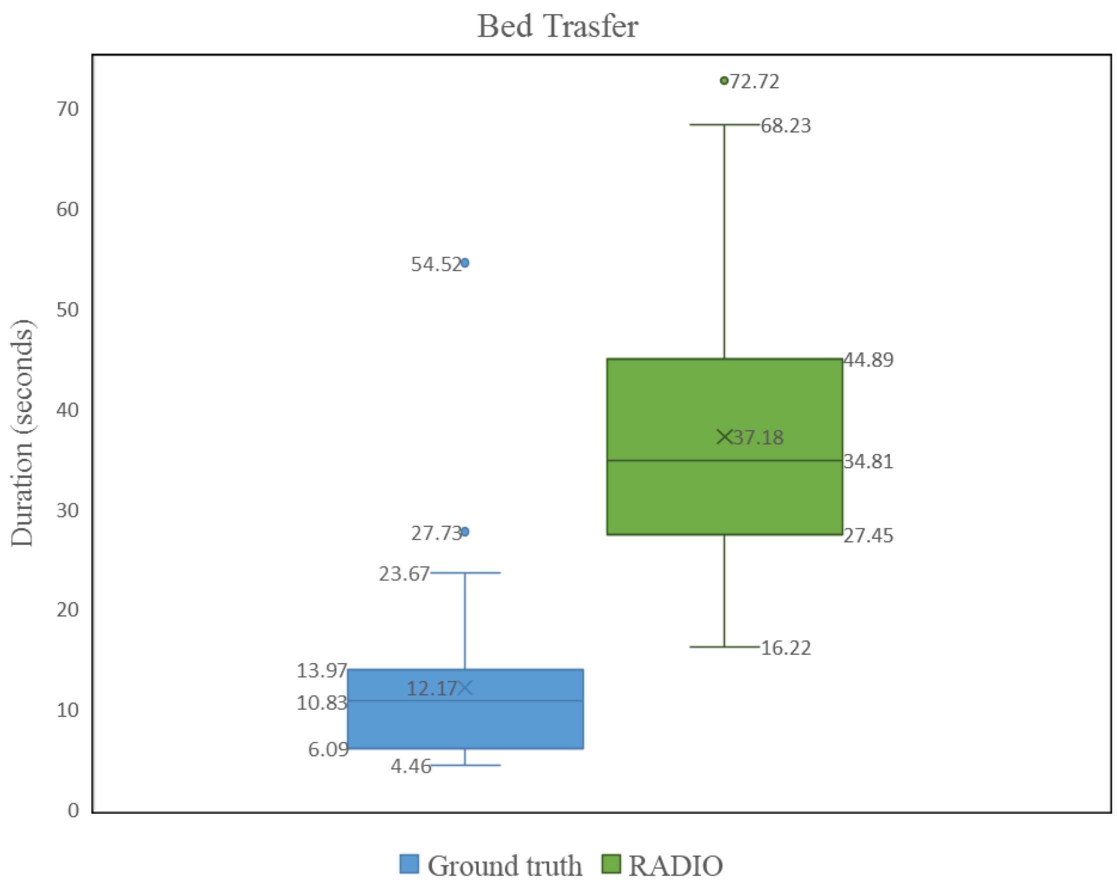


Figure 2. Box plots of detected bed transfers.

3.2 Chair transfer

For the chair transfer ADL, we analyzed in total **144 sessions** (12 repetitions x 12 participants –*c.f. Section 2.1*). Out of these sessions, the RADIO system *did not detect* the ADL 64 instances. The robot data for the rest 80 *detected chair transfers* and the corresponding ground truth data are presented in Figure 3. As can be seen in Figure 3, most of the RADIO values fall outside the min to max range of ground truth values (min = 0.78 and max = 6.77). Out of 80 actually detected chair transfers, only 8 can be classified as correct detections (true positives), while 72 are classified as wrong detections (false positives).

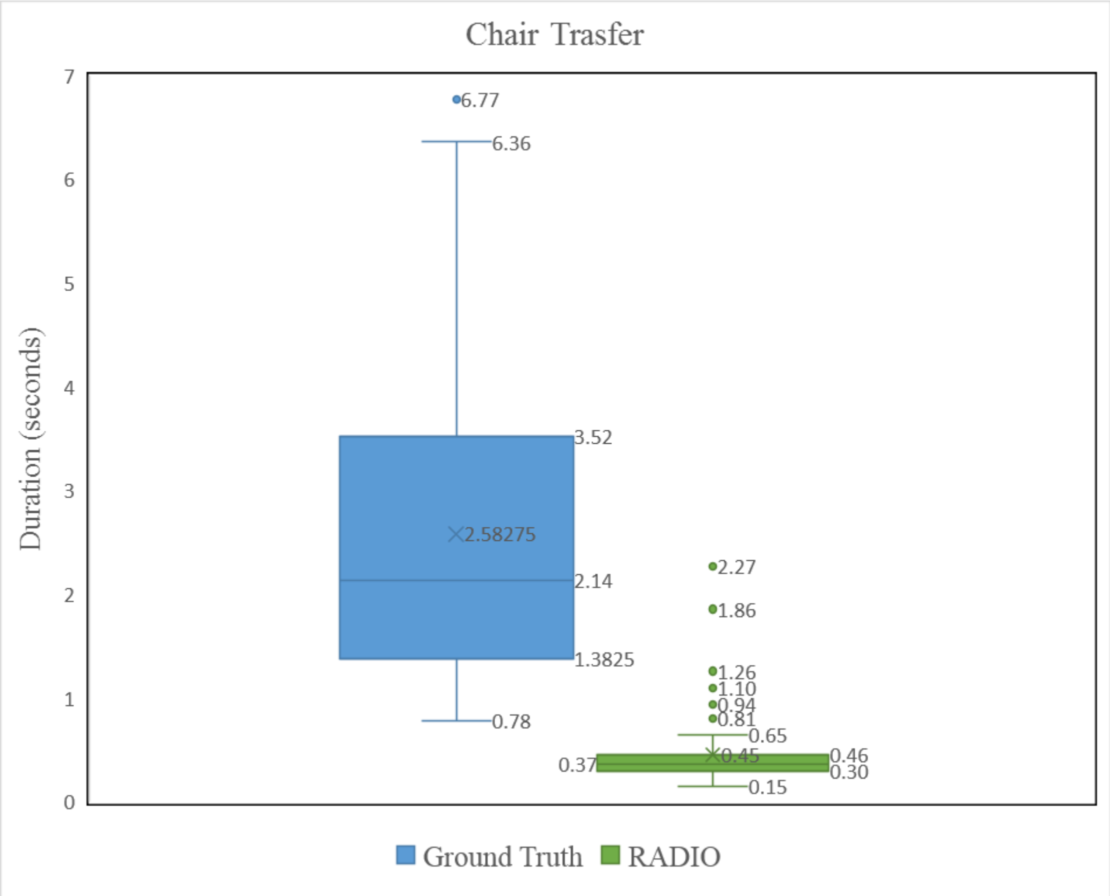


Figure 3. Box plots of detected chair transfers.

3.3 4-meter walk

For the 4-meter walk ADL, we analyzed in total **192 sessions** (16 repetitions x 12 participants –*c.f. Section 2.1*). Of these sessions, the RADIO system *did not detect* the ADLon 19 instances. The robot data for the rest 173 *detected bed transfers* and the corresponding ground truth data are presented in Figure 4. As can be seen, some of the RADIO values fall outside the min to max range of ground truth values (min=3.95 and max=32.68). Out of the 173 detected 4-meter walks, 152 can be classified as correct detections (true positives), while 21 are classified as wrong detections (false positives).

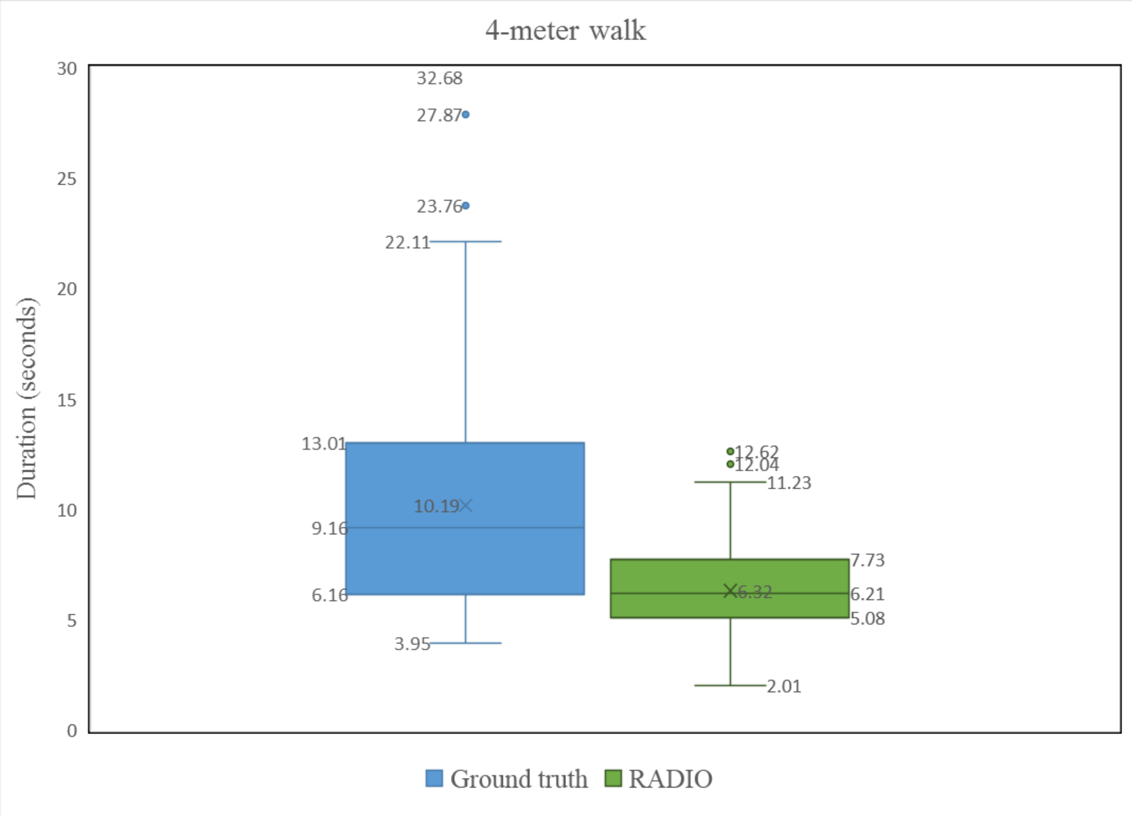


Figure 4. Box plots of detected 4-meter walks.

3.4 Pill intake

For the *pill intake* ADL, we analyzed in total 120 sessions (10 repetitions x 12 participants). Of these sessions, the RADIO system *did not detect* the ADL in 17 instances and *detected* 103 pill intakes.

3.5 Meal preparation

For the *meal preparation* ADL, we analyzed in total 48 sessions. Of these sessions, the RADIO system *did not detect* the ADL in 37 instances and *detected* 11 meal preparation events.

3.6 TV watching

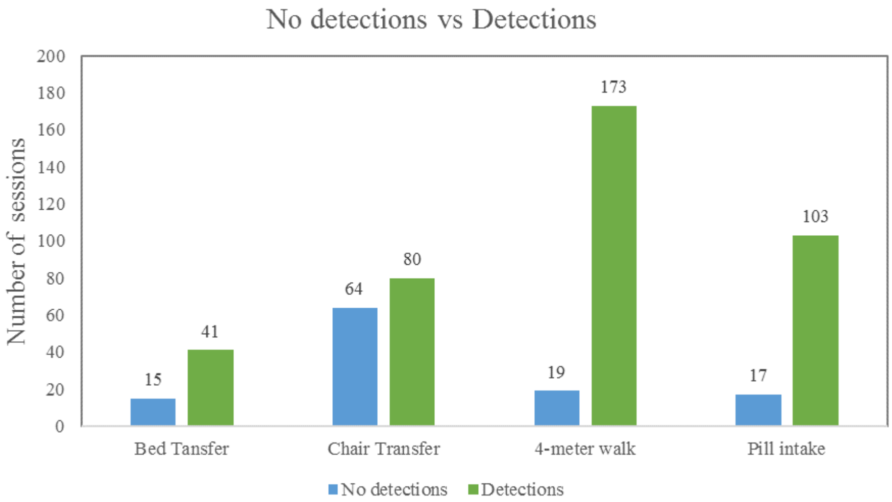
For the *TV watching* ADL, we analyzed in total 24 sessions. Of these sessions, the RADIO *system did not detect* the ADL in 12 instances and *detected* 12 events.

3.7 Going out of the room

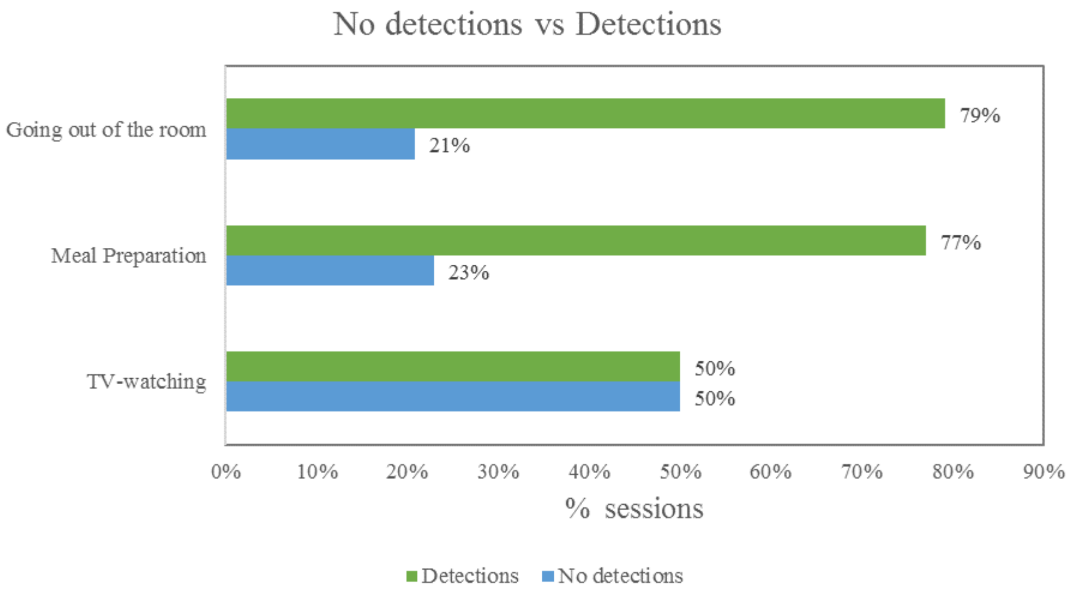
For the *going out of the room* ADL, we analyzed in total 24 sessions. Of these sessions, the RADIO system *did not detect* the ADL in 5 instances and *detected* 19 events.

3.8 Overall detection evaluation of the ADL methods

Figure 5a presents the bar charts of detection vs no detection sessions across all methods monitored by the RADIO robot and Figure 5b the percentage of events detected by the RADIO Smart Home. Figure 6 presents the correct vs wrong detections again across all methods that monitored an actual value and not just the occurrence or not of the event (i.e. bed transfer, chair transfer and 4 meter walk).



(a) ADLs monitored by the RADIO robot



(b) ADLs monitored by the RADIO Smart Home

Figure 5. RADIO system's no detections and detections.

Table 2 presents the overall results of the detection sessions performed by the RADIO system divided into correct detections, wrong detections, and no detections. The variables in Table 2 are used to calculate the fitness for purpose of the system as defined by the Precision, Sensitivity and F-measure indices. The results of this analysis are presented in Table 3.

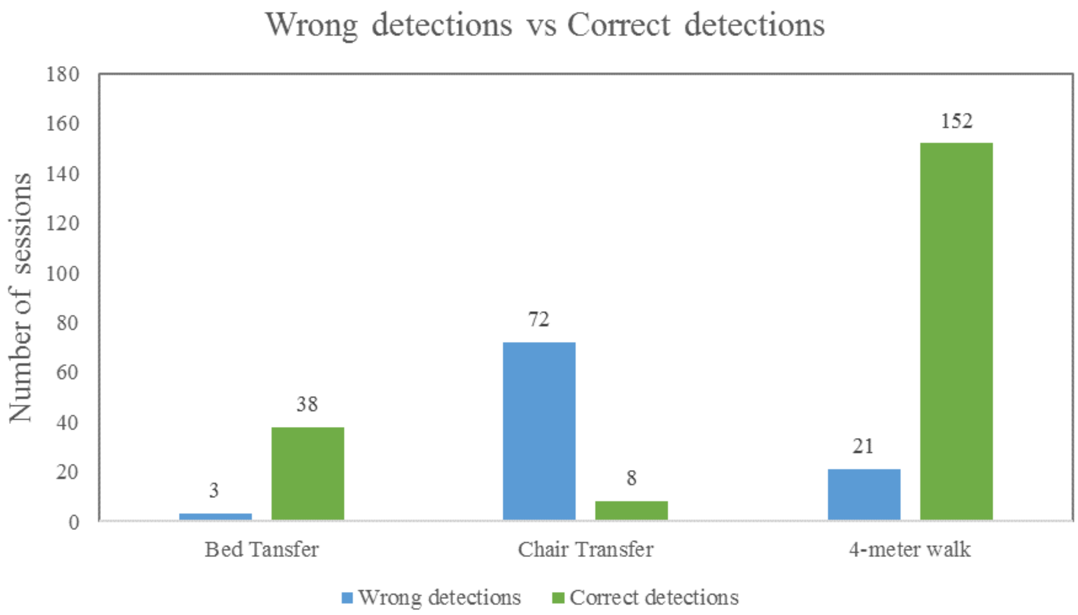


Figure 6. RADIO system's wrong and correct detections.

Table 2. Overall detection results of the RADIO system

Detection	Bed Transfer	Chair Transfer	4-meter walk	Pill intake	Meal Prep	TV watching	Going out
Correct – True Positives	38	8	152	103	37	12	19
Wrong – False positives	3	72	21	N/A	N/A	N/A	N/A
No detection – False Negatives	15	64	19	17	11	12	5
Total	56	144	192	120	48	24	24

Table 3. Measures of fitness for purpose of the ADL recognition methods

Measure	Bed Transfer	Chair Transfer	4-meter walk	Pill intake	Meal Prep	TV watching	Going out
Precision	0.93	0.10	0.88	1.00	1.00	1.00	1.00
Sensitivity	0.72	0.11	0.89	0.86	0.77	0.50	0.79
F-measure	0.81	0.11	0.88	0.92	0.87	0.67	0.88

4 ADL DURATION MEASUREMENT

4.1 Bed Transfer

Figure 7 presents ground truth measurements against RADIO ones. The points presented in Figure 7 refer to the ***38 sessions*** where RADIO measurements were ***classified as correct detections***. Kolmogorov-Smirnov test for *ground truth* and *RADIO robot measurements* was p=0.094 and p=0.2 respectively. The correlation analysis indicated that there is a statistically significant moderate correlation between the two groups of measurements (Pearson r =0.597, $p<0.001$).

The linear regression between the two groups of measurements is given by:

Robot-data = 23.5+ (1.04 * Ground-Truth), R^2 = 0.357

The mean standard deviation (MSD) between the RADIO and ground truth measurements is 634.93. This is partitioned in squared bias (SB) of 572.14 (translation of unity slope), non-unity slope (NU) of 0.05 (rotation of unity slope) and lack of correlation (LC) of 62.77 (representative of scatter). In other words, the deviation of the data set from the 1:1 line can mainly be primarily explained by a bias which was expected based on the linear regression factors. This bias is a systematic error that is caused by the specifics of the definition of the transfer start and transfer end between the method and the human observer.

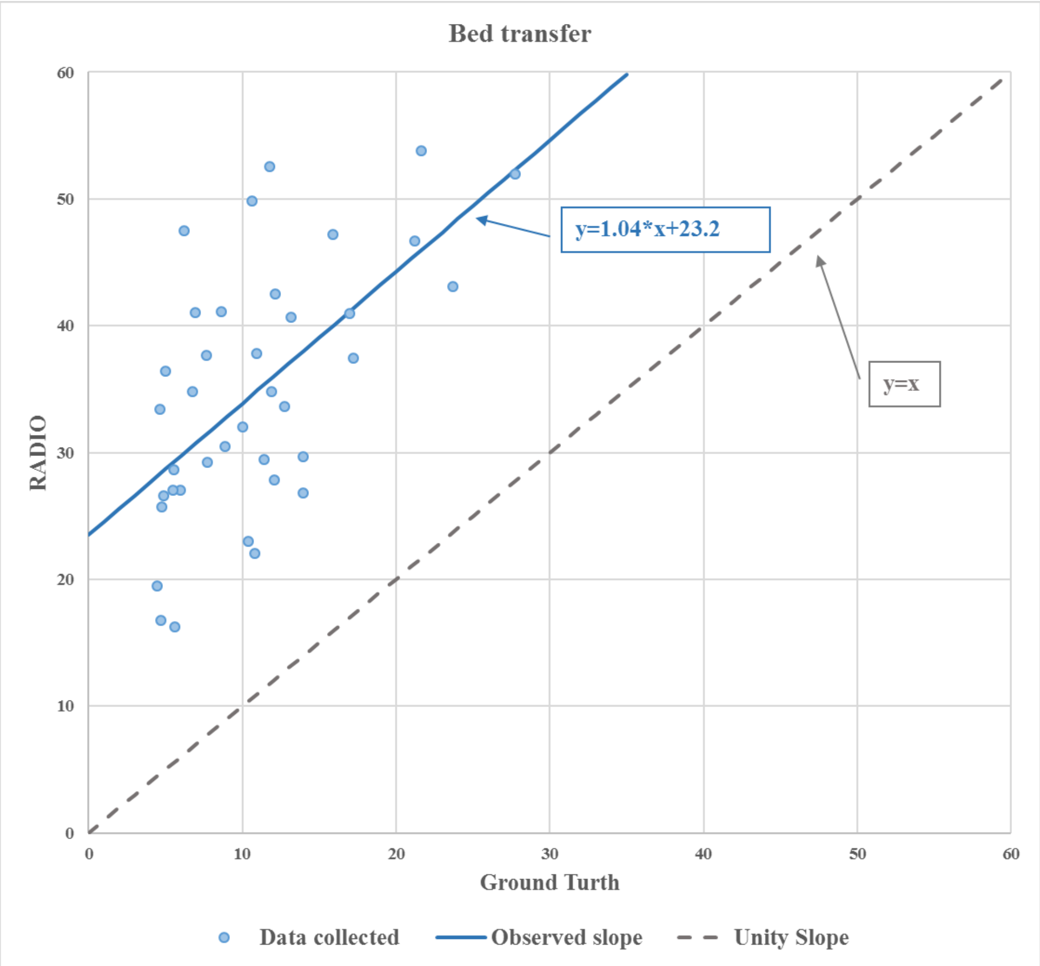


Figure 7. Bed transfer ground truth data versus RADIO measurements.

The difference in the results between the first and the second pilot is attributed to the higher number of correct detections collected in the second round of pilot studies, which was achieved by improvements in the method carried out by AVN making it more robust to the different lighting conditions observed at different times during the day.

4.2 Chair Transfer

Figure 8 presents ground truth measurements against RADIO ones for the chair transfer. The points presented in Figure 8 refer to the ***8 sessions*** where RADIO measurements were classified as ***correct detections***. Kolmogorov-Smirnov test for *ground truth* and *RADIO robot measurements* was p=0.088 and p=0.12 respectively. There was a weak negative non statistically significant correlation between the two groups of measurements (Pearson r=-0.41and p=0.924).

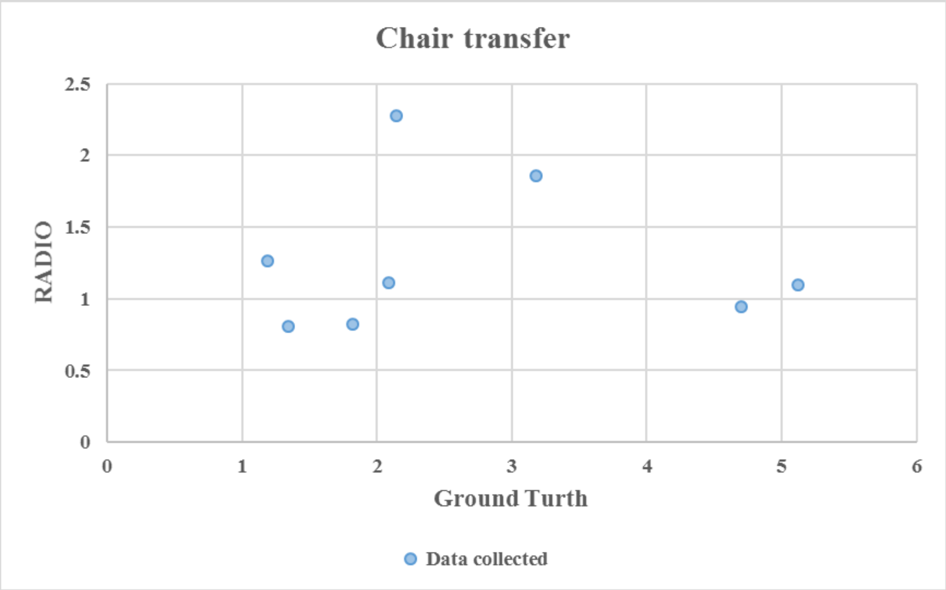


Figure 8. Chair transfer ground truth data versus RADIO measurements.

4.3 4-meter walk

Figure 9 presents ground truth measurements against RADIO ones for the 4m walk ADL. The points presented in Figure 9 refer to the ***152 sessions*** where RADIO measurements were classified as ***correct detections***. Kolmogorov-Smirnov test for *ground truth* and *RADIO robot measurements* was p=0.004 and p=0.059 respectively. In this case there was moderate positive, statistically significant correlation between the two groups of measurements (Spearman's rho: r=0.5, p<0.001).

The linear regression between the two groups of measurements is significant (p<0.001) and it is given by:

Robot-data = 5.16 + (0.15 * Ground-Truth).

The mean standard deviation (MSD) between the RADIO and ground truth measurements is 31.85. This is partitioned in squared bias (SB) of 13.65 (translation of unity slope), non-unity slope (NU) of 16.23 (rotation of unity slope) and lack of correlation (LC) of 1.97 (representative of scatter). In other words, the deviation of the data set from the 1:1 line can be explained mainly by a bias and a rotation of the dataset and in a lesser degree from a scatter of the collected points.

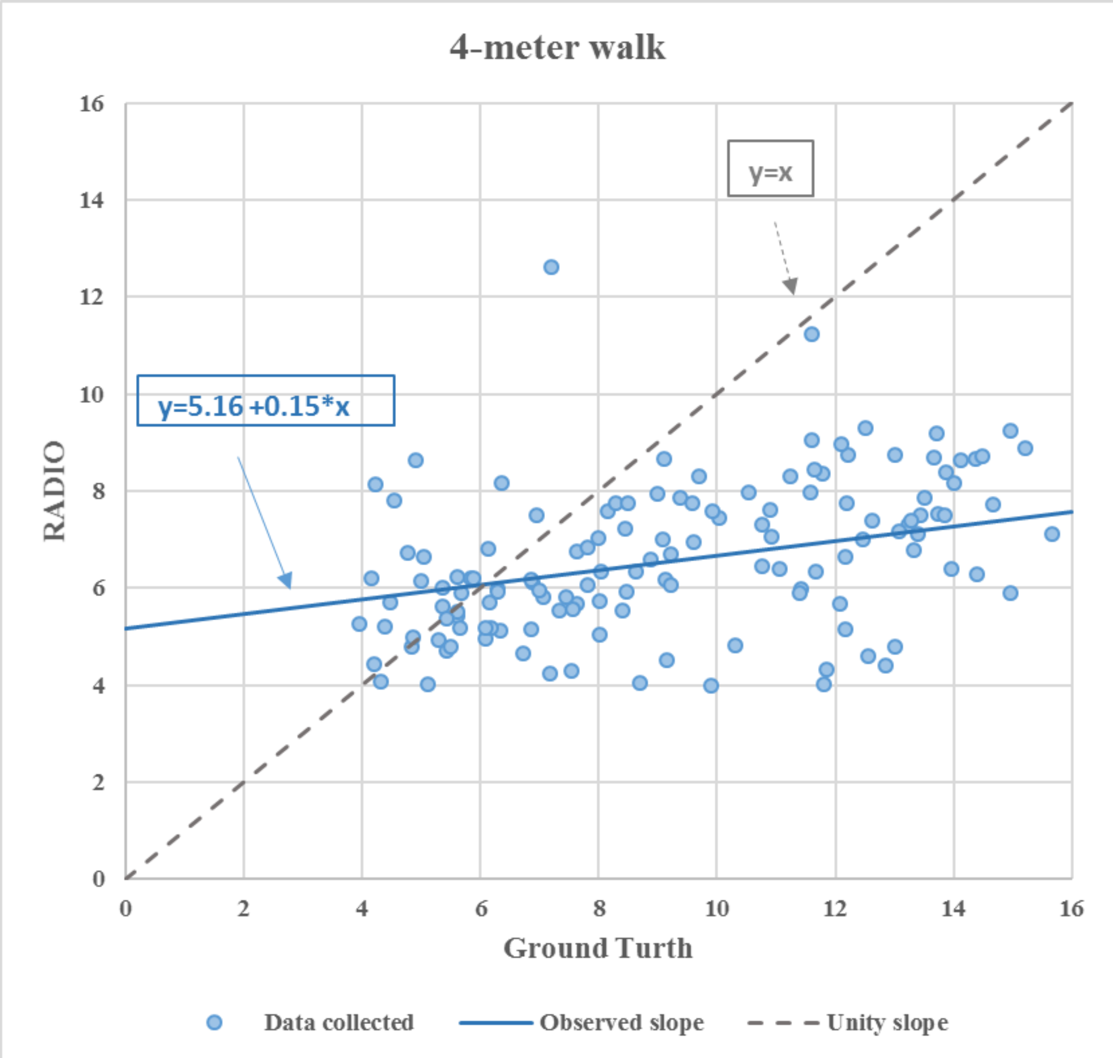


Figure 9. 4-meter walk ground truth data versus RADIO measurements.

The regression lines between the first and second round of pilot studies expose a relatively consistent behavior of the 4 - meter walking method (the regression line of the first round of pilot studies was Robot-data = 4.37 + (0.19 * Ground-Truth). However, ideally an identical behavior would have been observed. A noticeable difference between the two pilot studies was the observed range of the participants' behavioural data. Figure 10 shows the distribution of ground truth values in the two pilot studies.

To further explore the behavior of the 4m walk monitoring method we explored the correlation and regression lines in two subgroups of the 2nd round pilot study. The first subgroup contained data below the group average (mean = 10.49) and the second group above average. The results are presented in the following table:

Table 4. Statistical tests for the two 4-meter walking subgroups.

	K-S	Correlation	Regression
Subgroup 1 < mean	Ground truth: p=0.077, Radio: p=0.065	Pearson's r= 0.286, p=0.01	Robot-data = 4.7 + (0.19 * Ground-Truth) (significant, p=0.01)
Subgroup 2 > mean	Ground truth: p <0.001, Radio: p=0.2	Spearman's rho, r=.167, p= 0.057	Robot-data = 6.2 + (0.01 * Ground-Truth) (non-significant, p=0.075)

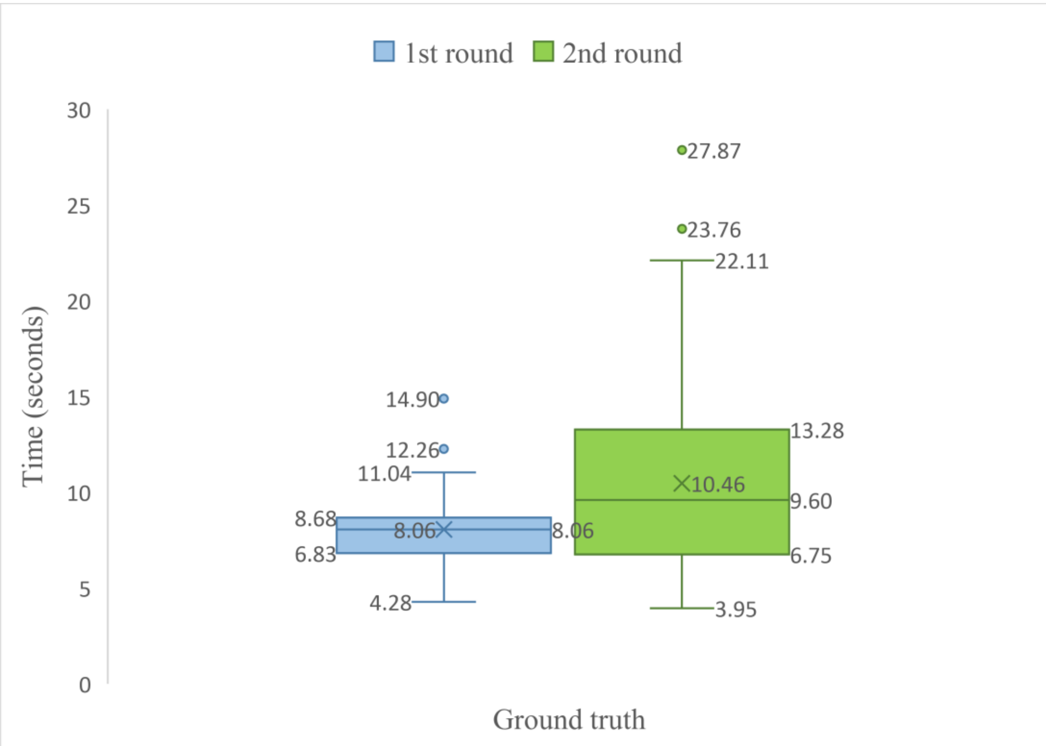


Figure 10. 4-meter walk ground truth data (correct detections) in the 1st & 2nd round of pilot studies.

As can be seen, the method gives a significantly smaller measurement than the human observer, which is more pronounced for instances where the human observation is over 10.5sec. This is due to the fact that the method separates walking from non-walking and the measurement is made over walking only, ignoring the parts of the scene where no walking occurs. Measurements over 10.5sec for 4m are a clear indication that stops were made during the scene and the 4m were not completed without stopping. Given this, it is expected that the method provides a smaller measurement for those instances where the human measurement is larger.

5 CONCLUSION

This pilot study aimed to facilitate the medical evaluation of the integrated RADIO prototype as a support platform for ADL and IADL assessment.

The overall observation regarding measurements is that:

- Besides from a systematic bias, the manually configured visual motion detection method (bed transfer) and the laser scan method (4m walk) correlate closely with the manually recorded values. That is to say, if the evolution of these values was used to assess ADL ability, the RADIO measurements would evolve in the same way as the manually obtained values. These methods performed well, as expected from the outcomes of the technical validation carried out in WP3.
- The configuration-free visual motion classification method (used in chair transfer) is less robust to both different environments and to changes in the conditions within the environment (lighting, changes in background due to changes in furniture position). This method performed worse than what was observed during the technical validation carried out in WP3.
- The methods that inferred ADLs based on Smart Home sensors' activity, performed reasonably well, with a small number of undetected events. Unlike the measurement methods above, these methods are not direct observations but rely on assumptions (encoded as rules) that infer event occurrence from multiple elementary events (such as opening and closing of cupboards followed by turning on the stove infers "meal preparation"). Undetected instances are sometimes due to undetected elementary events, violating the assumptions encoded in the rules.

Table 5. Summary of ADL recognition results

Measure	Bed Transfer	Chair Transfer	4-meter walk	Pill intake	Meal Prep	TV watching	Going out
Recognition F-measure	0.81	0.11	0.88	0.92	0.87	0.67	0.88
Measurement consistence (Pearson's)	$r = 0.597$, $p < 0.001$	$r = -0.41$, $p = 0.924$	$r = 0.286$, $p = 0.01$	N/A	N/A	N/A	N/A

From these we conclude that:

- The *range data walking detection method* is a robust and configuration-free method. If the measurement that is needed is possible to be obtained from planar range data (like walking speed measurements), then it is an *adequate ADL monitoring solution*.
- The manually configured *visual motion detection method* is a robust method, but it requires configuration for each particular deployment. This is also an *adequate ADL monitoring solution*, but creates a *usability* question from the perspective of the installation technician.
- For the configuration-free *visual motion classification method* to become adequate, *further development* in the upstream machine vision technologies is needed.
- Inferring complex events from elementary events (instead of directly observing them) gains from the robustness of simple sensor activation by comparison to machine vision, but then suffers from the lack of robustness of the inferences that combine the elementary events. This approach is nevertheless an *adequate ADL monitoring solution*.