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 6.14

Medical evaluation report II

Dissemination Level	**Public**
Due Date of Deliverable	Project Month M30, September 2017
Actual Submission Date	6 February 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 first round of the *Summative Phase* pilot studies using first integrated RADIO prototype at FHAG premises and at FZ clients' private residences. 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. Based on the results of this, evaluation methods and the design of the pilot studies have been improved for the next round of pilot studies.

History and Contributors

Ver	Date	Description	Contributors
00	04 Sept 2017	Document structure	NCSR-D
01	29 Sept 2017	Input in Sections 2.1 and 3	FHAG
02	3 Oct 2017	Input in Sections 4.1	FSL
03	5 Oct 2017	Input in Section 4.2	FSL, FHAG
04	23 Nov 2017	Updated analysis	FHAG
05	8 Dec 2017	Updated analysis	NCSR-D
06	11 Dec 2017	Review of the whole document, minor corrections in all sections.	FSL
07	14 Dec 2017	Review of the whole document, minor corrections in all sections	FHAG
08	18 Dec 2017	Updated analysis in Sections 3.3	FHAG.
09	21 Dec 2017	Updated analysis in Section 3 and revised Section 5.	FHAG
10	20 Jan 2018	Review of the whole document, minor corrections in all sections	FHAG, FSL
11	23 Jan 2018	Review of the whole document, minor corrections in all sections	FSL
12	27 Jan 2018	Internal peer review	TWG
12	31 Jan 2018	Addresses peer review comments	NCSR-D, FSL, FHAG
Fin	6 Feb 2018	Final preparation and submission	NCSR-D

Executive Summary

This deliverable reports the Medical Evaluation of the first round of the *Summative Phase* of pilot study using first integrated RADIO prototype at FHAG premises and at FZ clients' private residences. 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 pilot study revealed the existence of a very high percentage of missing observations with a range between 6 and 56% of the records due to both methodological and network issues. Currently, the only ADL acceptably detected by the RADIO system would be the 4 meters walking. The results, have been used to improve the methods that will be used in the next round of pilot studies.

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 methods and the analysis according to these methods of the data collected during the first round of Summative Phase pilot studies.

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 first round of Summative Phase pilot studies using first integrated RADIO prototype at FHAG premises and at FZ clients' private residences. During this phase, patients were monitored with RADIO system and ground truth assessment was recorded as well. This dual assessment generated a variety of summary statistics (recall, precision, and the F-measure) that are useful to evaluate the first 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 *D6.3 Piloting plan III*. The execution of trials and details about piloting, its outcomes and technical details are reported in *D6.7. Pilot report I*. User evaluation results and the technical lessons learned from piloting are described in *D6.11 User Evaluation III*.

1.3 Relation to other Work Packages and Deliverables

This document reports the medical evaluation results of the first round of the Summative Phase pilot studies. These trials were executed at FGA premises and at FZ clients' private residences during May – June 2017.

The data collected during the trials reported were reported in *D6.7. Pilot report I*. These data were analyzed in the context of Task 6.4 and Task 6.5 and were used for user evaluation reported in *D6.11 User Evaluation III* and for medical evaluation reported in the current document *D6.14 Medical evaluation report II*. The evaluation results also include points to be considered in the design of the next piloting plan (D6.4).

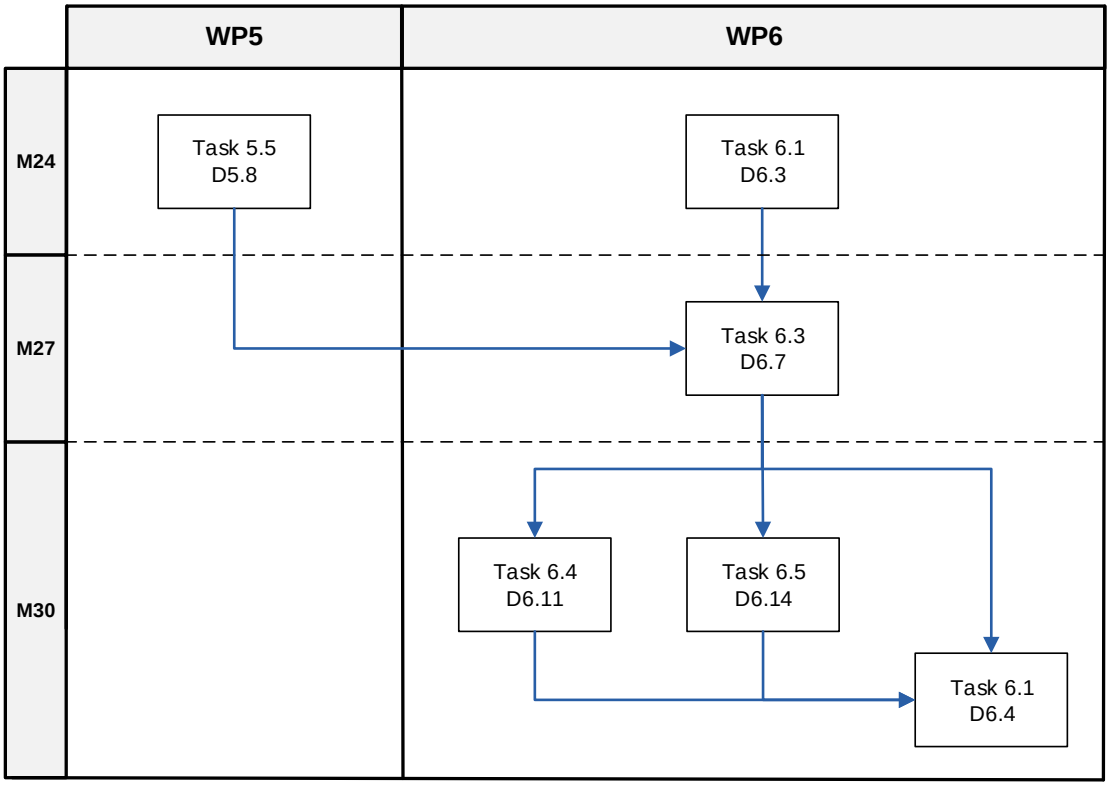


Figure 1. Dependencies between this deliverable and other deliverables.

2 METHODS

This section 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 four ADLs.

2.1 Evaluation dataset at FHAG

As described in D6.7, eight (8) participants completed a 3-days 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 informing clinical staff about the duration of each detected activity. All eight emails were successfully received. However, the **first participant is excluded** from further analysis due to both pilot and technical failures. Moreover, in the *bed transfer* ADL *only*, we excluded from further analysis the **first three participants** (both ground truth and RADIO robot data), as there were no items recorded 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 robotic platform, the occurrence of the events, as well as their duration, was also collected by FHAG researchers (ground truth). Details about how the ground truth was collected can be found in *D6.3 Pilot Plan III*.

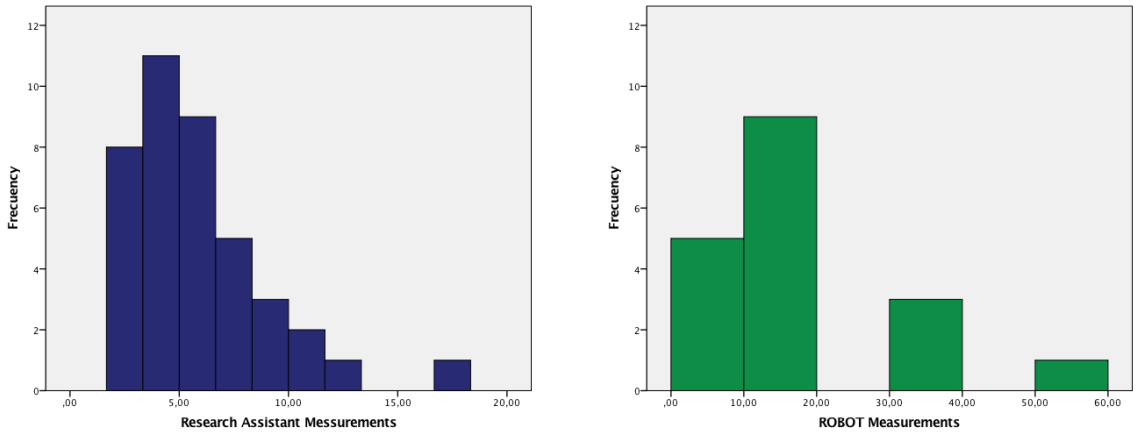
In summary, the evaluation reported in this document includes data from the RADIO system and their ground truth. For all ADLs, except pill intake and those recorded through Smart Home sensors, there are two kinds of information: detection of the activity and duration of the activity. Figures 2a, 2b and 2c show the comparison of the distributions between measures of ADL duration as recorded by the robot and the ground truth. This comparison is available for bed/walking/chair ADL, not for medication intake ADL as for this activity no measure of duration was performed but only the occurrence detection.

2.2 Evaluation dataset at FZ

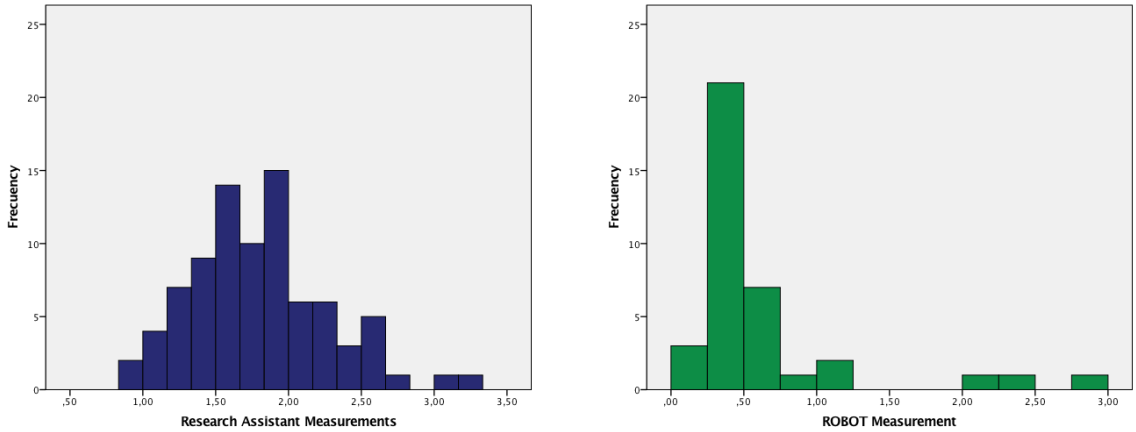
As described in D6.7, two (2) participants completed a 3-days study scenario. During this, each one completed the following repetitions for each of the four ADLs monitored by the RADIO robot:

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

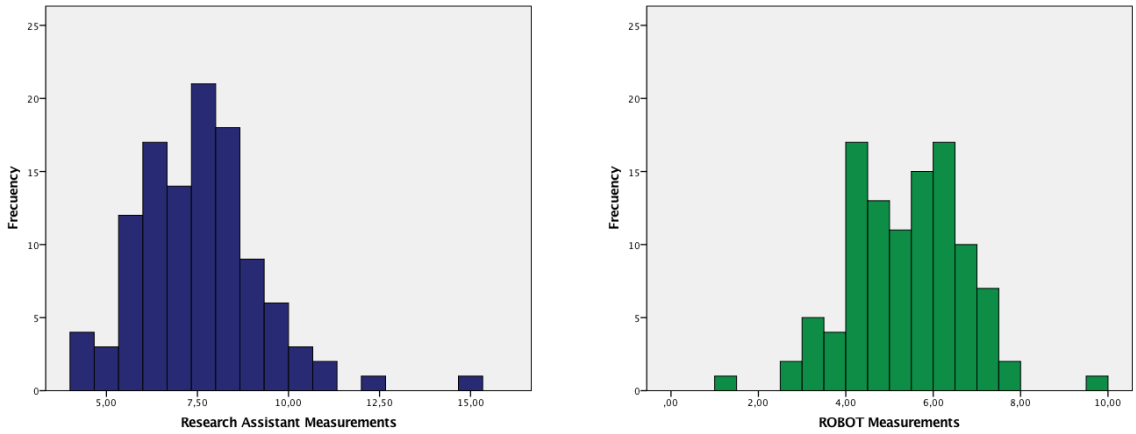
At the end of each participant's scenario, an email was sent informing clinical staff about the duration of each detected activity. The total number of data used for evaluation for each ADL is reported in Section 3.



a) Comparison of bed transfer ground truth (left) and RADIO measurements (right).



b) Comparison of chair transfer ground truth (left) and RADIO measurements (right).



c) Comparison of 4 meters walk ground truth (left) and RADIO measurements (right).

Figure 2. Comparison of ground truth and RADIO measurements for the FHAG pilot study.

2.3 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 a 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.

Table 1. ADL data categorization based on detection.

ADL	Correct detection	Wrong detection	No detection
Bed	The robot detected an actual event and the value reported is **not** lower that the min value of ground truth or higher that the max value of ground truth. min (GT measurement) < RADIO measurement < max (GT measurement)	The robot detected an actual event and the value reported is lower that the min value of ground truth or higher that the max value of ground truth. RADIO measurement < min (GT measurement) AND RADIO measurement> max (GT measurement)	The RADIO system **did not detect** an actually occurring event (no email entry).
Chair			
4-meter walk			
Medication intake	The robot detected an actual event.	N/A	
TV watching	Smart home sensors detected an actual event.	N/A	
Meal Preparation	Smart home sensors detected an actual event.	N/A	
Going out of the room	Smart home sensors detected an actual event.	N/A	

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 and are reported in Section 3.

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.

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

2.4 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.

[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.

- 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.
    - $LC = (1 - r^2) * \frac{\sum y^2}{N}$

3 RESULTS AT FHAG

3.1 ADL Detection

3.1.1 Bed Transfer

For the bed transfer ADL, we analyzed in total 40 sessions (5 participants x 8 repetitions –*c.f. Section 2.1*). Of these sessions, the RADIO system *did not detect* the ADL in 20 instances. Moreover, 2 more instances are further considered as no detections due to unreasonably high values (1530.6 and 155.76). The rest of the data, as recorded by both the RADIO system and ground truth are presented in Figure 3. Out of the 18 actually detected bed transfers, 11 can be classified as correct detections (true positives), while 7 are classified as wrong detections (false positives – falling out of the ground truth measurements' interval: min (2.50) and max (17.10))

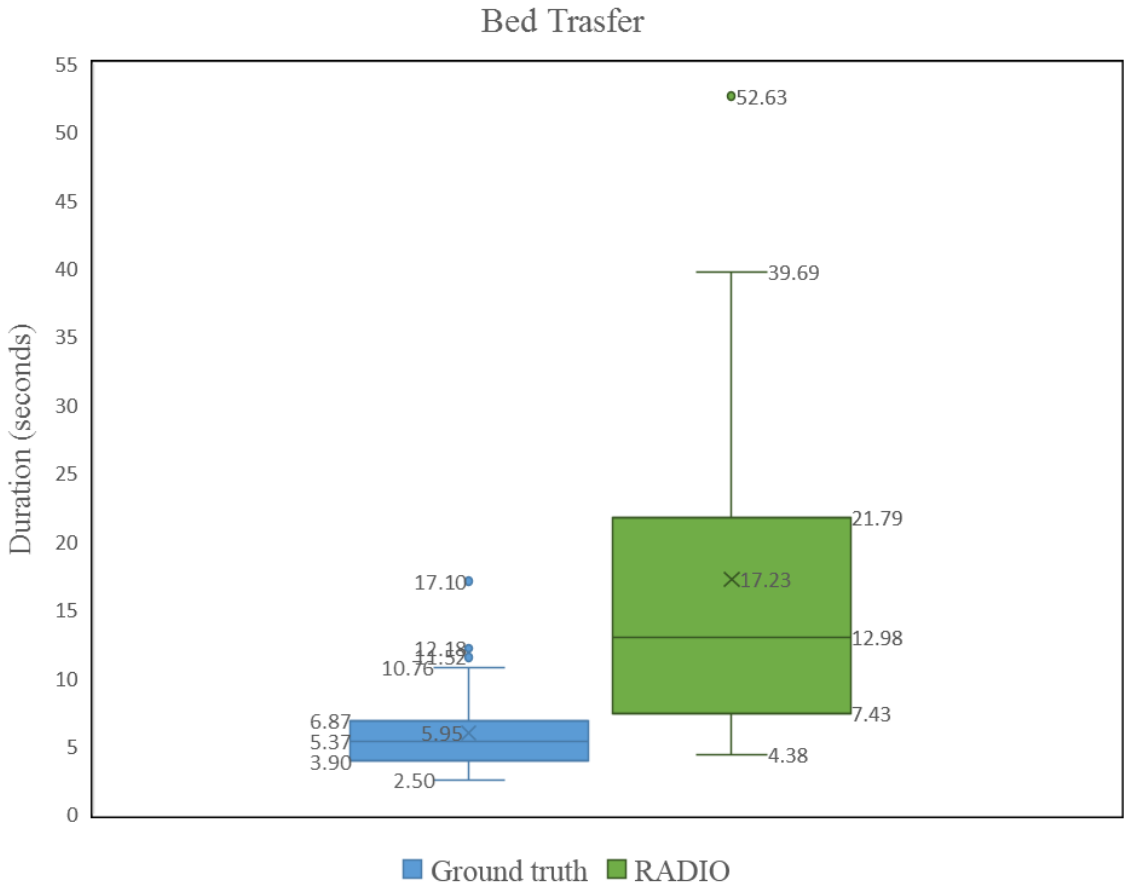


Figure 3. Box plots of detected bed transfers.

3.1.2 Chair transfer

For the chair transfer ADL, we analyzed in total **84 sessions** (12 repetitions x 7 participants). Out of these sessions, the RADIO system *did not detect* the ADL 47 instances. The rest of the data, as recorded by both the RADIO system and ground truth are presented in Figure 4. As can be seen in Figure 4, most of the RADIO values fall outside the min to max range of ground truth values (min = 0.84 and max = 3.28). Out of 37 actually detected chair transfers, only 6 can be classified as correct detections (true positives), while 31 are classified as wrong detections (false positives).

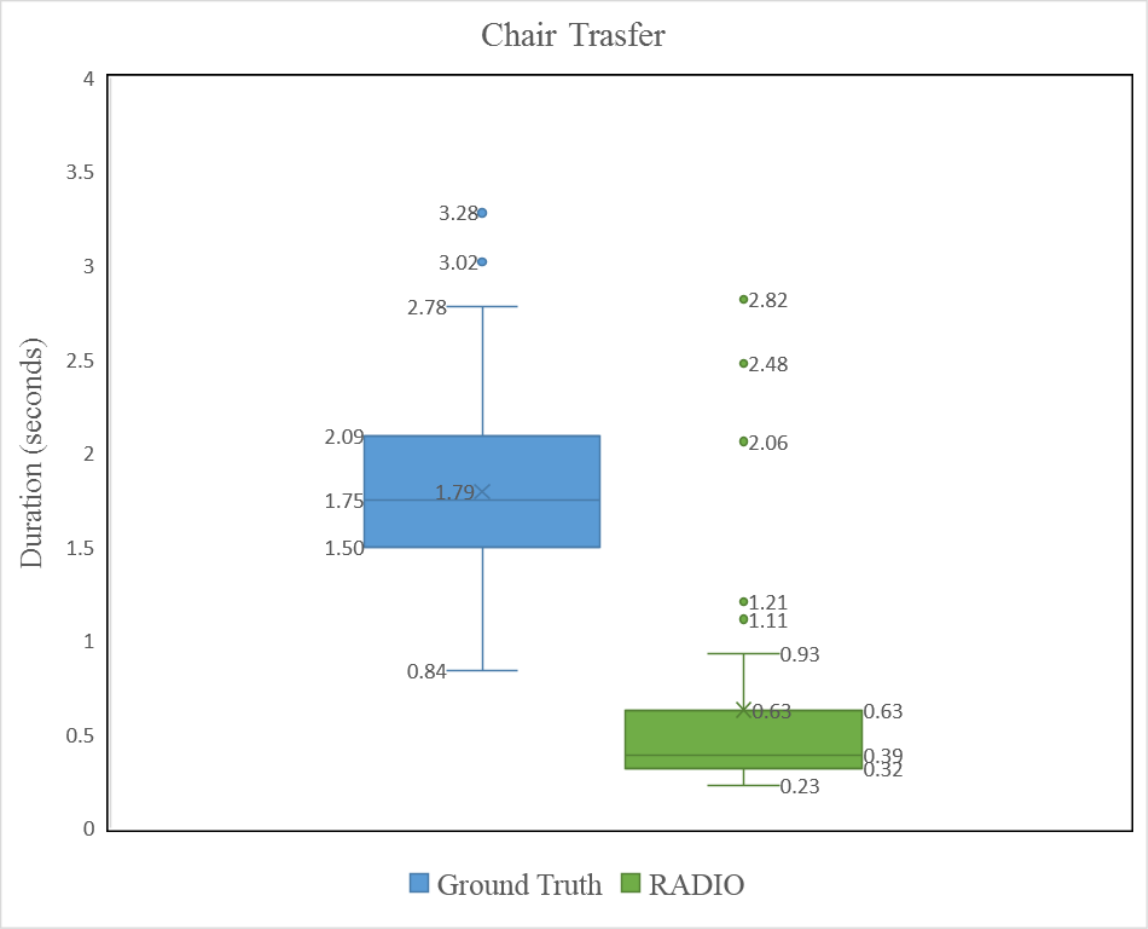


Figure 4. Box plots of detected chair transfers.

3.1.3 4-meter walk

For the 4-meter walk ADL, we analyzed in total **112 sessions** (16 repetitions x 7 participants). Of these sessions, there was one instance of ground truth lost due to human error and the RADIO system *did not detect* the ADL 7 instances. The rest of the data, as recorded by both the RADIO system and ground truth are presented in Figure 5. As can be seen, some of the RADIO values fall outside the min to max range of ground truth values (min=4.28 and max=14.90). Out of 105 detected 4-meter walks, 86 can be classified as correct detections (true positives), while 19 are classified as wrong detections (false positives).

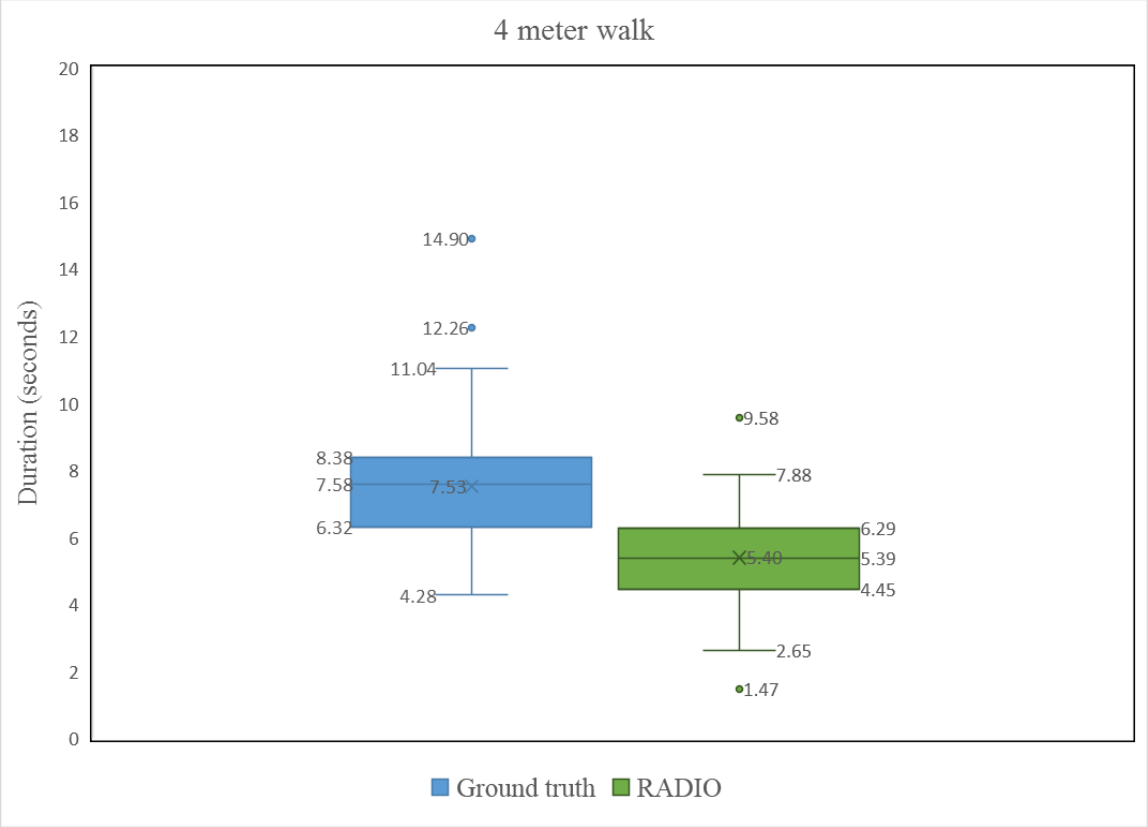


Figure 5. Box plots of detected 4-meter walks.

3.1.4 Pill intake

For the *pill intake* ADL, we analyzed in total 70 sessions (10 repetitions x 7 participants). Of these sessions, the RADIO system *did not detect* the ADL in 25 instances and *detected* 45 pill intakes.

3.1.5 Meal preparation

For the *meal preparation* ADL, we analyzed in total 28 sessions. Of these sessions, the RADIO system *did not detect* the ADL 7 instance and *detected* 21 meal preparation events.

3.1.6 TV watching

For the *TV watching* ADL, we analyzed in total 28 sessions. Of these sessions, the RADIO *detected* all 28 events.

3.1.7 Going out of the room

For the *going out of the room* ADL, we analyzed in total 16 sessions. Of these sessions, the RADIO system *detected* all 16 events.

3.2 Overall detection evaluation of the ADL methods

Figure 6 presents the bar charts of detection vs no detection sessions across all methods. Figure 7 presents the correct vs wrong detections again across all methods (besides pill intake as this classification is not applicable in this case).

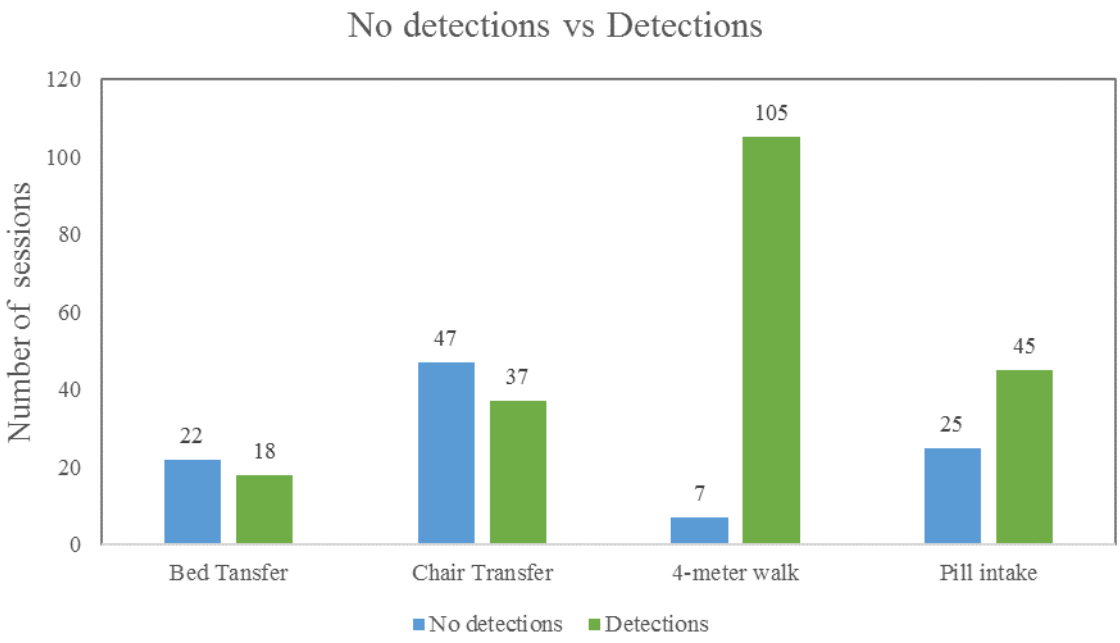


Figure 6. RADIO system's no detections and detections.

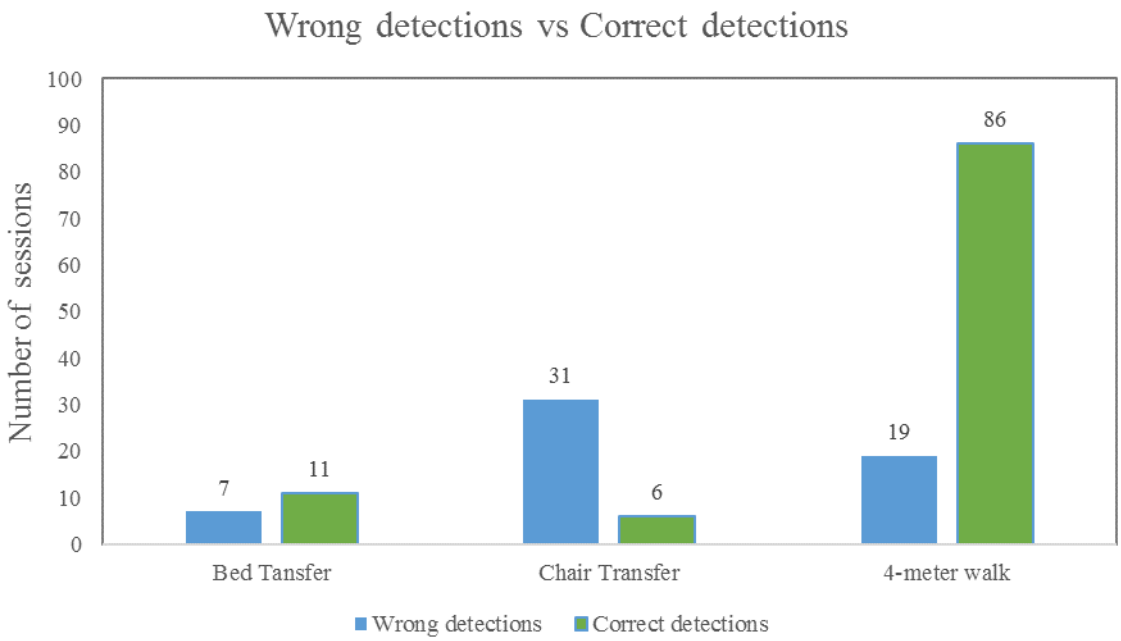


Figure 7. RADIO system's wrong and correct detections.

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.

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	11	6	86	45	21	28	16
Wrong – False positives	7	31	19	N/A	0	0	0
No detection – False Negatives	22	47	7	25	7	0	0
Total	40	84	112	70	28	28	16

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	0.61	0.16	0.82	1.00	1	1	1
Sensitivity	0.33	0.11	0.92	0.64	0.75	1	1
F-measure	0.43	0.13	0.87	0.78	0.86	1	1

3.3 ADL Duration Measurement

3.3.1 Bed Transfer

Figure 8 presents ground truth measurements against RADIO ones. The points presented in Figure 8 refer to the ***11 sessions*** where RADIO measurements were ***classified as correct detections***. Kolmogorov-Smirnov test for Ground Truth and Robot measurements were p=0.015 and p=0.141 respectively. The correlation analysis indicated that there is a non-statistically significant moderate correlation between the two groups of measurements (Spearmam r=0.345, p=0.298).

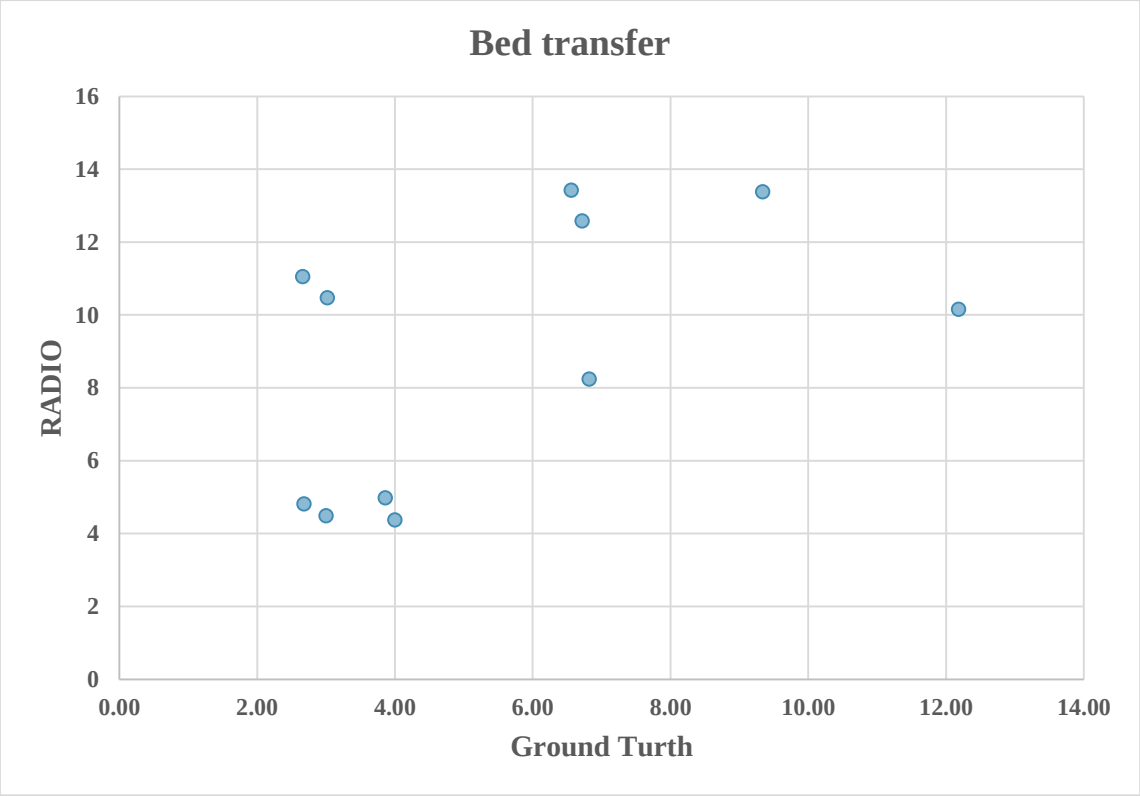


Figure 8. Bed transfer ground truth data versus RADIO measurements.

3.3.2 Chair Transfer

Figure 9 presents ground truth measurements against RADIO ones for the chair transfer. The points presented in Figure 9 refer to the ***6 sessions*** where RADIO measurements were classified as ***correct detections***. Kolmogorov-Smirnov test for Ground Truth and Robot measurements were $p<0.001$ and $p=0.2$ respectively. There was a weak negative non statistically significant correlation between the two groups of measurements (Spearman $r=-0.257$ and $p=0.623$).

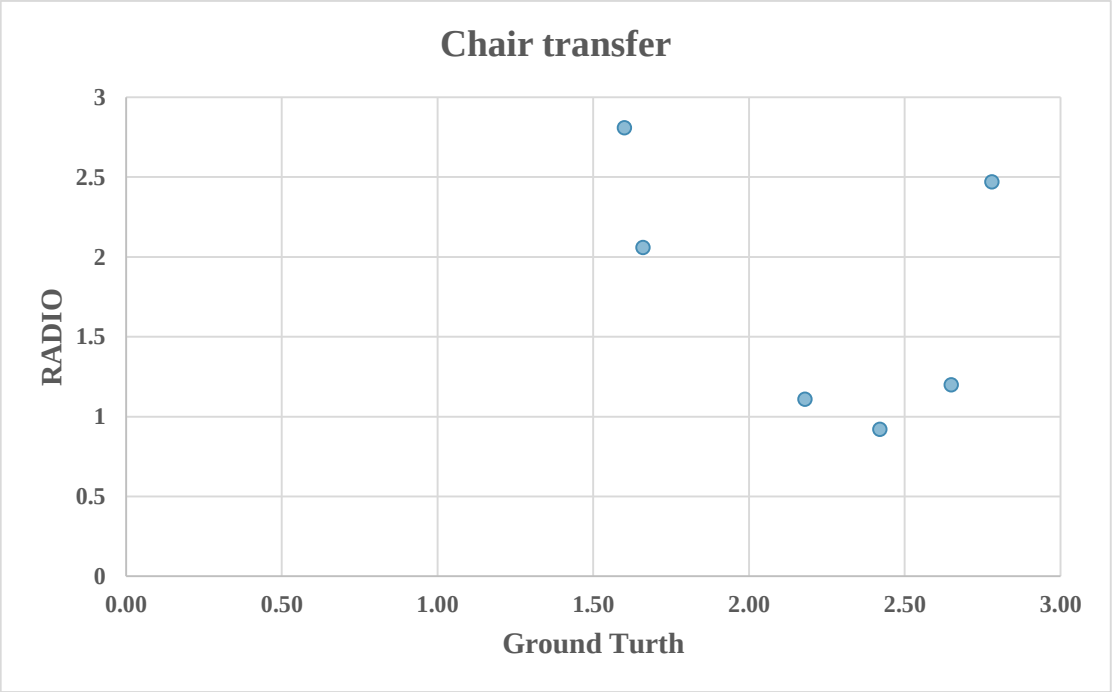


Figure 9. Chair transfer ground truth data versus RADIO measurements.

3.3.3 4-meter walk

Figure 10 presents ground truth measurements against RADIO ones for the 4m walk ADL. The points presented in Figure 10 refer to the ***86 sessions*** where RADIO measurements were classified as ***correct detections***. Kolmogorov-Smirnov test for Ground Truth and Robot measurements were $p=0.179$ and $p=0.2$ respectively. In this case there was moderate positive, statistically significant correlation between the two groups of measurements (Pearson $r=0.323$, $p=0.002$).

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

Robot-data = 4.37 + (0.19 * Ground-Truth)

The mean standard deviation (MSD) between the RADIO and ground truth measurements is 6.3. This is partitioned in squared bias (SB) of 3.2 (translation of unity slope), non-unity slope (NU) of 2.2 (rotation of unity slope) and lack of correlation (LC) of 0.9 (representative of scatter). In other words, the deviation of the data set from the 1:1 line can be explained by a bias and also rotation of the dataset and scatter of collected points.

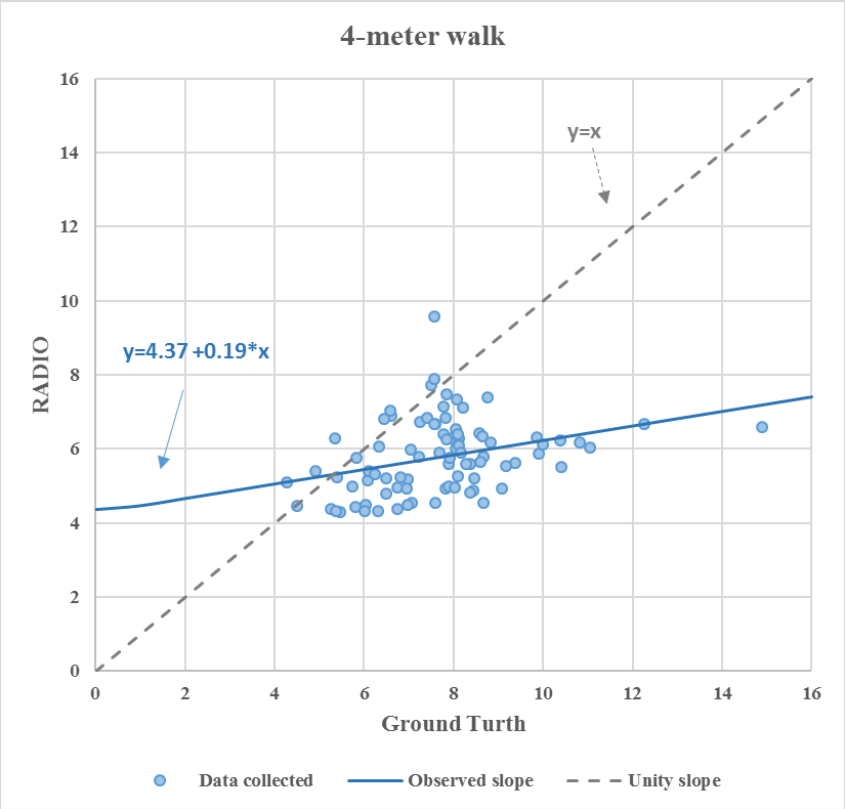


Figure 10. 4 meter walk ground truth data versus RADIO measurements.

4 RESULTS AT FZ

4.1 ADL Detection

4.1.1 Bed Transfer

For the bed transfer ADL, we analyzed in total 16 sessions (8 repetitions x 2 participants). Of these sessions, the RADIO system *did not detect* the ADL 6 instances. The rest of the data, as recorded by both the RADIO system and ground truth are presented in Figure 11. As can be seen in Figure 11, none of the RADIO values (ALL values recorded by RADIO appear in the right-hand side boxplot) falls inside the min to max range of ground truth values: min and max values are 3.4 and 8.16 accordingly, although the boxplots appear to have an overlapping area. Thus, out of 10 actually detected bed transfers, none can be classified as correct detections (true positives), and as a consequence, no further analysis is conducted for this ADL.

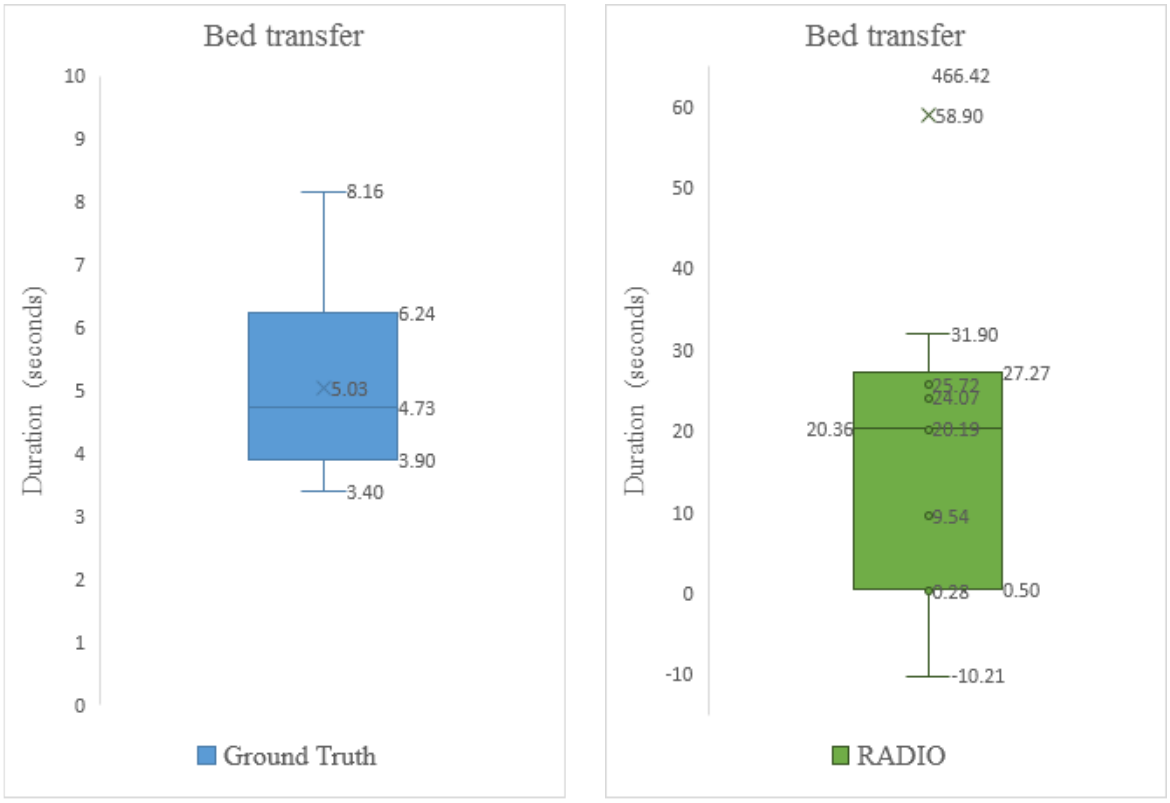


Figure 11. Box plots of detected bed transfers.

4.1.2 Chair transfer

For the chair transfer ADL, we analyzed in total 24 sessions (12 repetitions x 2 participants). Of these sessions, the RADIO system *did not detect* the ADL 21 instances. Figure 12 presents the data as recorded by the nurse (ground truth). The three values detected from RADIO are (measured in seconds): 1.34, 1.96, 0.68. As can be seen in Figure 12, 2 out of the 3 RADIO values fall outside the min to max range of ground truth values; min and max values are 1.90 and 5.60 accordingly. Out of 3 actually detected chair transfers, 1 can be classified as correct detection (true positive), while 2 are classified as wrong detections (false positives). Thus, no further analysis is conducted for this ADL.

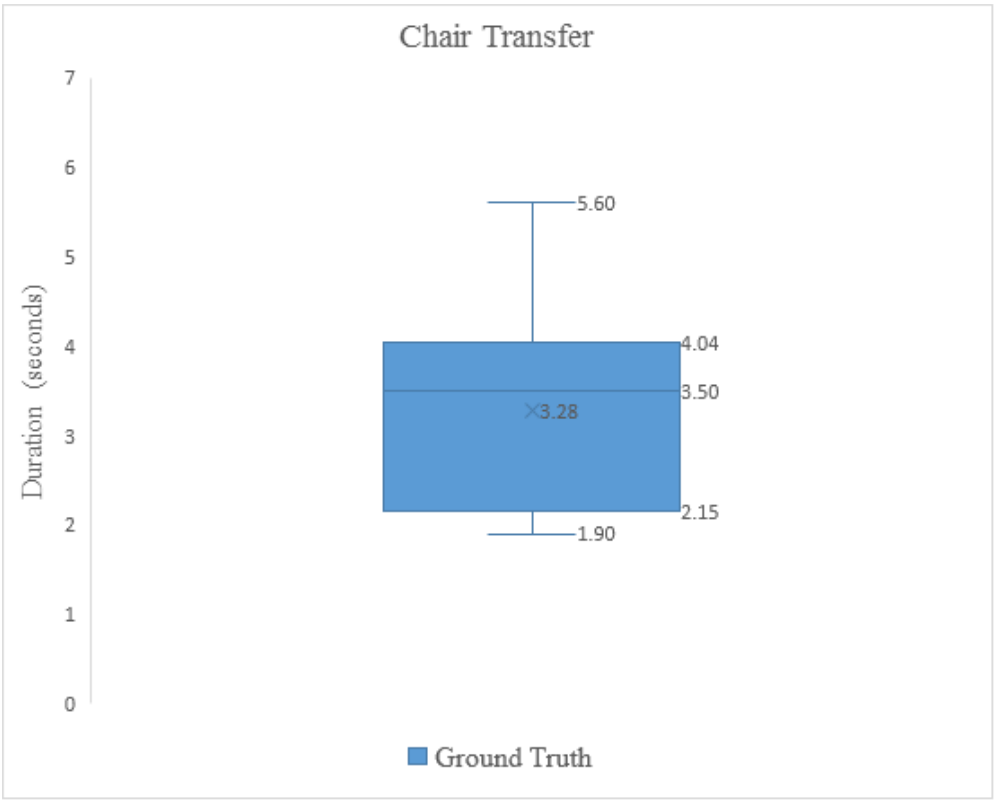


Figure 12. Box plots of detected chair transfers.

4.1.3 4-meter walk

For the 4-meter walk ADL, we analyzed in total 24 sessions (12 repetitions x 2 participants). Of these sessions, the RADIO system *did not detect* the ADL 10 instances. The rest of the data, as recorded by both the RADIO system and ground truth are presented in Figure 13. As can be seen, most of RADIO values fall inside the min to max range of ground truth values: min and max values are 4.50 and 9.00 accordingly. Out of 14 actually detected 4-meter walks, 13 can be classified as correct detections (true positives), while 1 is classified as wrong detection (false positives).

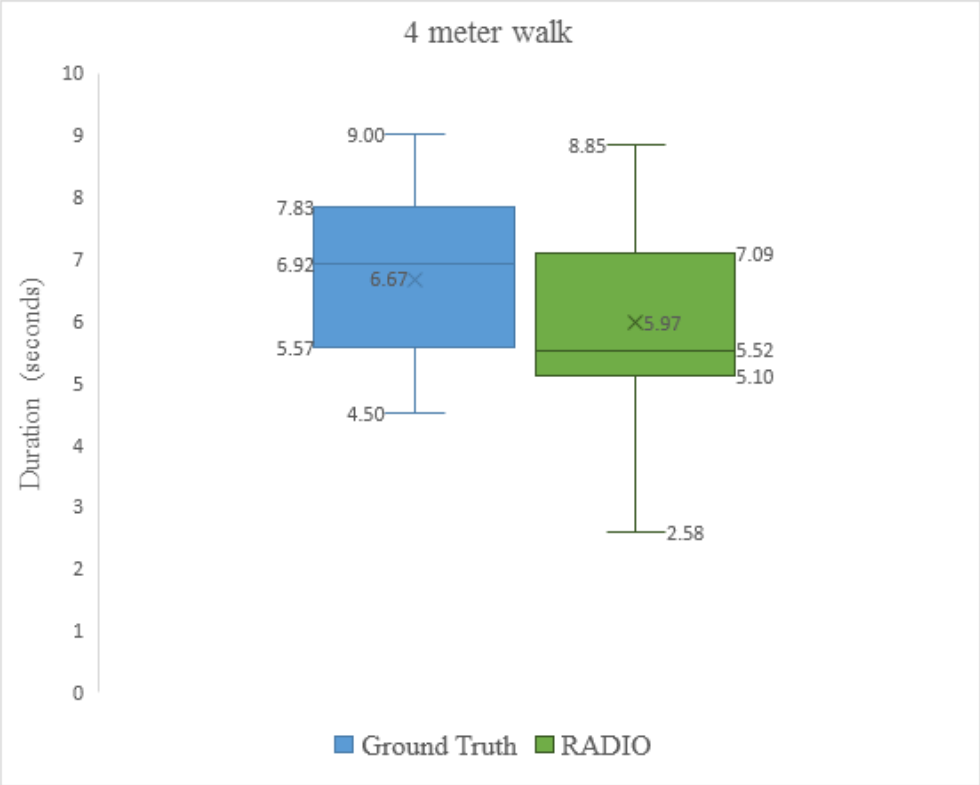


Figure 13. Box plots of detected 4-meter walks.

4.1.4 Pill intake

For the *pill intake* ADL, we analyzed in total 20 sessions (10 repetitions x 2 participants). Of these sessions, the RADIO system *did not detect* the ADL 16 instances and *detected* 4 pill intakes.

4.1.5 TV watching

Regarding Smart Home event detection, TV- watching was part of private residences pilot methods. Although smart – plugs to detect this event were installed, no events were detected due to Internet connectivity issues that made it impossible to access the S&C rule engine from the main controller.

4.1.6 Overall detection evaluation of the ADL methods

Figure 14 presents the bar charts of detection vs no detection sessions across all methods. Figure 15 presents the correct vs wrong detections again across all methods (besides pill intake as this classification is not applicable in this case).

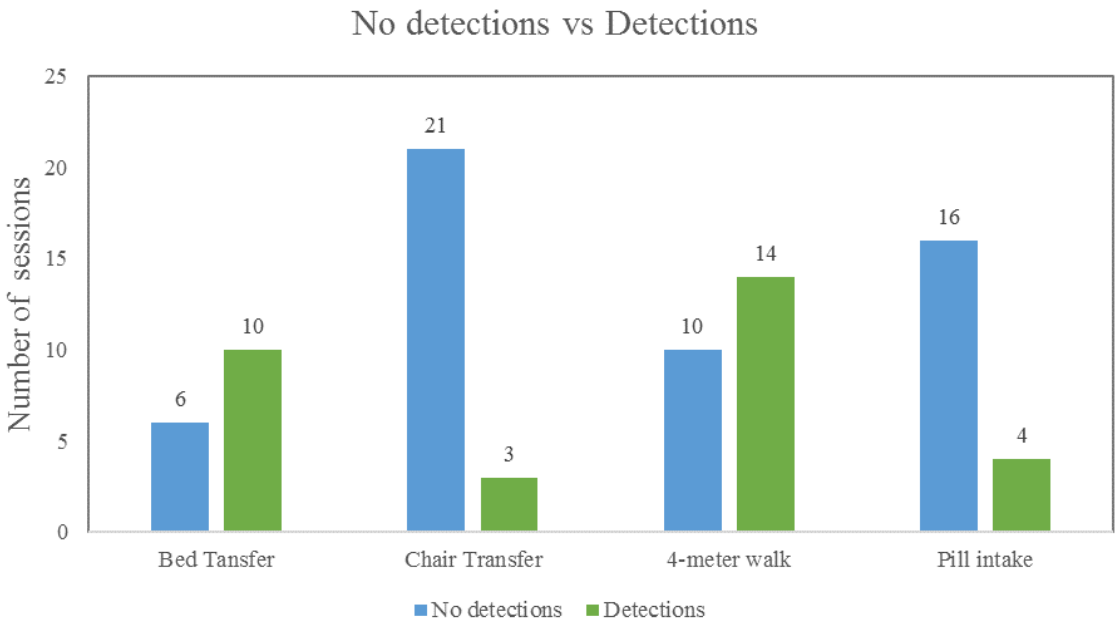


Figure 14. RADIO system's no detections and detections.

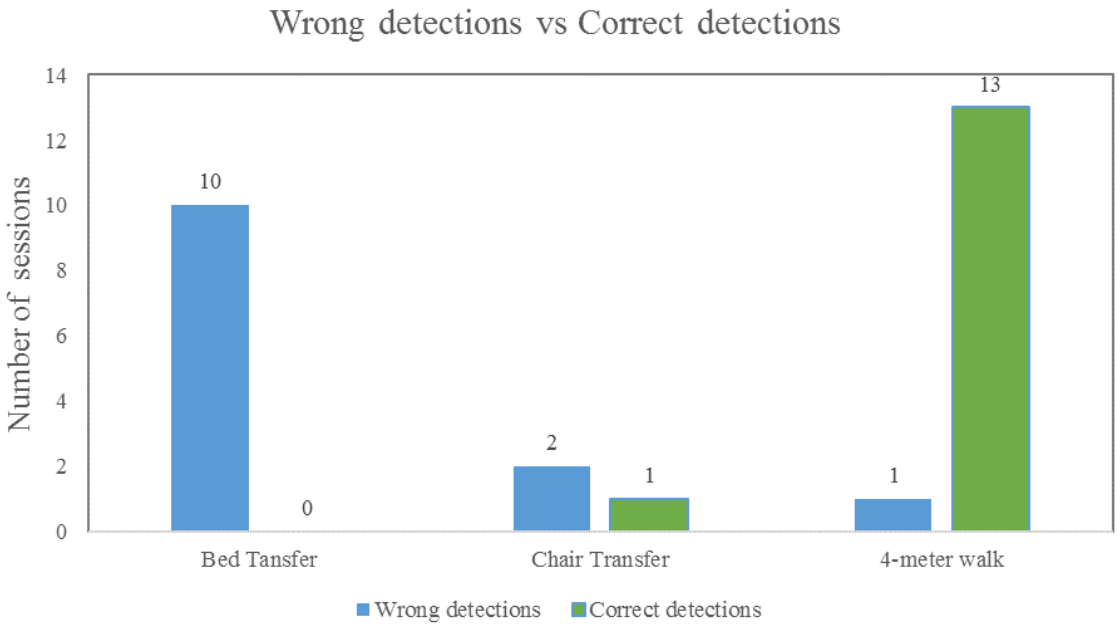


Figure 15. RADIO system's wrong and correct detections.

Table 1Table 4 presents the overall results of the detection sessions performed by the RADIO system divided, into correct detection, wrong detection, and no detection. The variables in Table 4 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 5.

Table 4. Overall detection results of the RADIO system

Detection	Bed Transfer	Chair Transfer	4-meter walk	Pill intake
Correct – True Positives	0	1	13	4
Wrong – False positives	10	2	1	N/A
No detection – False Negatives	6	21	10	16
Total	16	24	24	20

Table 5. Measures of fitness for purpose of the ADL recognition methods

Measure	Bed Transfer	Chair Transfer	4-meter walk	Pill intake
Precision	0	0.33	0.93	1
Sensitivity	0	0.045	0.56	0.20
F-measure	0	0.08	0.70	0.33

4.2 ADL Duration Measurement

4.2.1 4-meter walk

Figure 16 presents ground truth measurements against RADIO ones for the 4m walk ADL. The points presented in Figure 16 refer to the ***13 sessions*** where RADIO measurements were classified as ***correct detections***. In this case there is not a correlation between the two groups of measurements (r=-0.081, p=0.794).

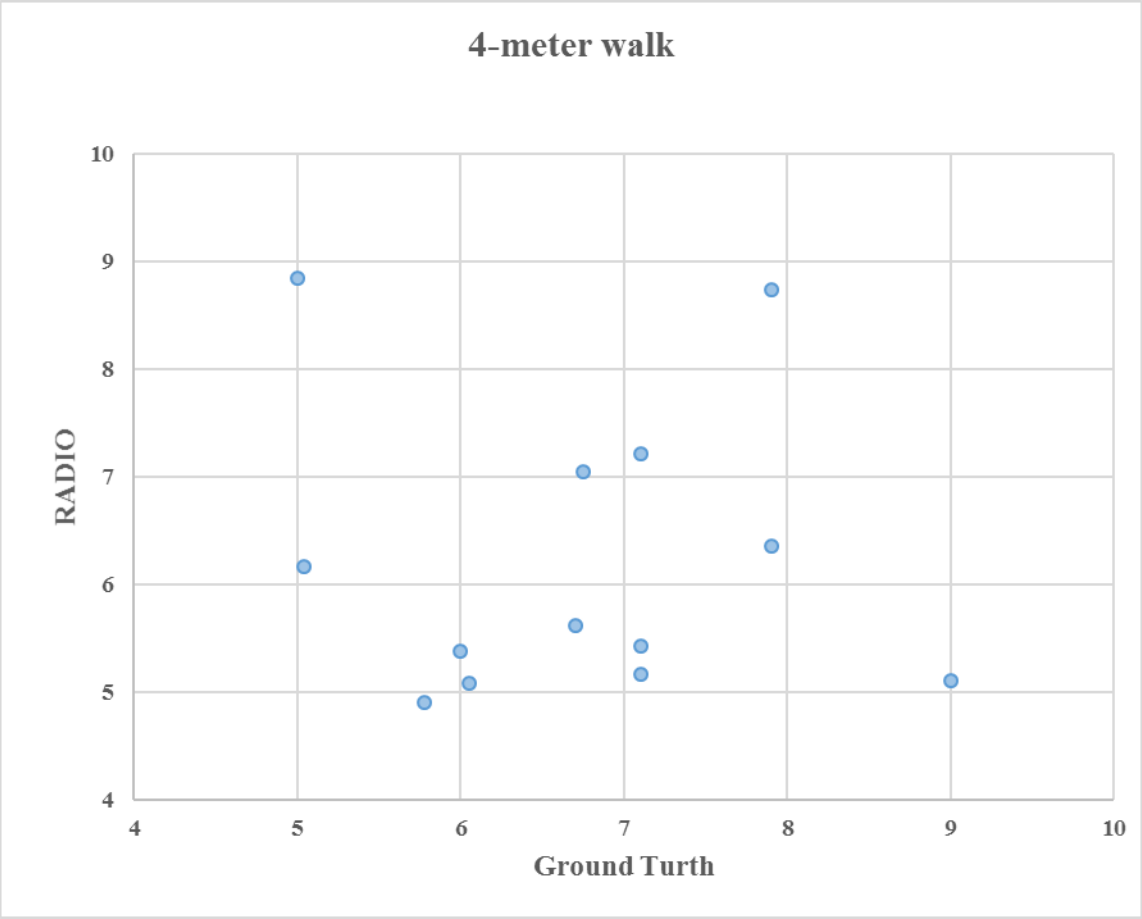


Figure 16. 4 meter walk ground truth data versus RADIO measurements.

5 DISCUSSION

This pilot study aimed to facilitate the medical evaluation of the integrated RADIO prototype as a support platform for ADL and IADL assessment. At that time the pilot study reveals the existence of a very high percentage of missing observations with a range between 6 and 56% of the records.

The **non-detected events** by the RADIO system in relation to the studied ADLs reveals the following percentages for FHAG and FZ respectively (in brackets percentages on the whole sample are indicated):

- Bed transfer: 55% and 37.5% (50%),
- Chair transfer: 55.9% and 87.5% (63%),
- 4-meters walk: 6.25% and 41.7% (12.5%),
- Pill intake: 44% and 80% (45.5%).

It seems that most of the no detections in transfer were because of network issues and not the method itself.

With respect to the clinometric characteristics of detected events, we observed measurement errors or **false positive** detections in:

- Bed transfer: 38.8% and 100% (60.7%)
- Chair transfer: 83.7% and 66.7% (82.5%)
- 4-meters walk: 18.1% and 7.1% (16.8%)
- Pill intake: not applicable.

These are very high percentages, particularly in transfers and to a lesser extent in walking. Observing the F values, as a global measure of performance, when considering both the pilots the range is 0-0.87.

The association between measurements by the assistant (ground truth) and the robot has been analysed by means of correlation tests between two numerical variables for bed and chair transfer and 4 meters walking. All three correlation coefficients show weak-moderate relationship. Being the 4 meters walking the ADL, which best perform with a moderate association however. The only statistically significant correlation was with the 4-m walking activity at FHAG data showing a moderate positive association (Pearson r = 0.32) but probably insufficient for a measurement test. FZ data did not show correlation between RADIO and ground truth assessment.

Therefore, at this time the only ADL acceptably detected by the RADIO system would be the 4 meters walking. The percentages of non-detection in transfers and taking medication are very high and require a detailed analysis of possible causes. There were network issues with transfers but also very high measurement errors in both FHAG and FZ data (FHAG chair transfers 93% and FZ 100% bed transfer).

This procedure required a lot of manual participation on the part of the researcher in the activation of the system prior to the measurement. This circumstance evidences a prototype situation at an experimental level that can only be manipulated in controlled laboratory circumstances yet, unlike the smart home sensors that have acted in a real application situation.

Therefore, the pilot experiment allows us to only assess the detections of the activities studied but not yet the concept of “monitoring the activities” since at this moment it forces the participation of the researcher in the activation of the system and the standardized schedule of the activities.