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 6.13

Medical evaluation report I

Dissemination Level	**Public**
Due Date of Deliverable	Project Month M21, December 2016
Actual Submission Date	12 April 2017
Work Package	WP6, *Piloting and evaluation*
Task	Task 6.5, *Medical evaluation*
Lead Beneficiary	FSL
Contributing beneficiaries	NCSR-D
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 *Intermediate Phase* of pilot study using first prototype of the RADIO system at FSL premises. Four ADL methods were used to recognize: bed transfer, chair transfer, 4 meter walk and pill intake. 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 have been improved for the next round of pilot studies.

History and Contributors

Ver	Date	Description	Contributors
00	25 Nov 2016	Document structure	NCSR-D
01	12 Jan 2017	First draft	FSL
02	12 Mar 2017	Review by NCSR-D and additions of more details about the outcomes from FSL	FSL, NCSR-D
03	30 Mar 2017	Addition of more details on methodology (Section 2) and of duration distributions (Section 3.2).	FSL
04	7 Apr 2017	Evaluation results and conclusion from the perspective of improving the technical solution. Version sent to ethics board for review.	NCSR-D
05	11 Apr 2017	Further technical details were added.	NCSR-D, FSL
06	12 April 2017	Internal peer review	FZ
07	12 April 2017	Addresses peer review comments	NCSR-D, FSL
Fin	12 April 2017	Final preparation and submission	NCSR-D

Executive Summary

This deliverable reports the Medical Evaluation of the *Intermediate Phase* of pilot study using first prototype of the RADIO system at FSL premises. Four ADL methods were used to recognize: bed transfer, chair transfer, 4 meter walk and pill intake. 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. The evaluation results showed that 4 meter walk and pill intake methods suffered from a high number of no detections. Moreover, bed transfer and pill intake produced many false positives. Based on the results, technical explanations are given in the report. This information has 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
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 document the medical evaluation methods and the analysis according to these methods of the data collected during the Intermediate Phase Pilot Study.

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 *Intermediate Phase* pilot study at FSL premises, on nursing home residents and elderly people living in the community participating on the RADIO project. 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 *D6.2 Piloting plan.* The execution of trials and details about piloting, its outcomes and technical details are reported in *D6.6. Controlled pilot trials report II*. User evaluation results and the technical lessons learned from piloting are described in *D6.10 User Evaluation II*.

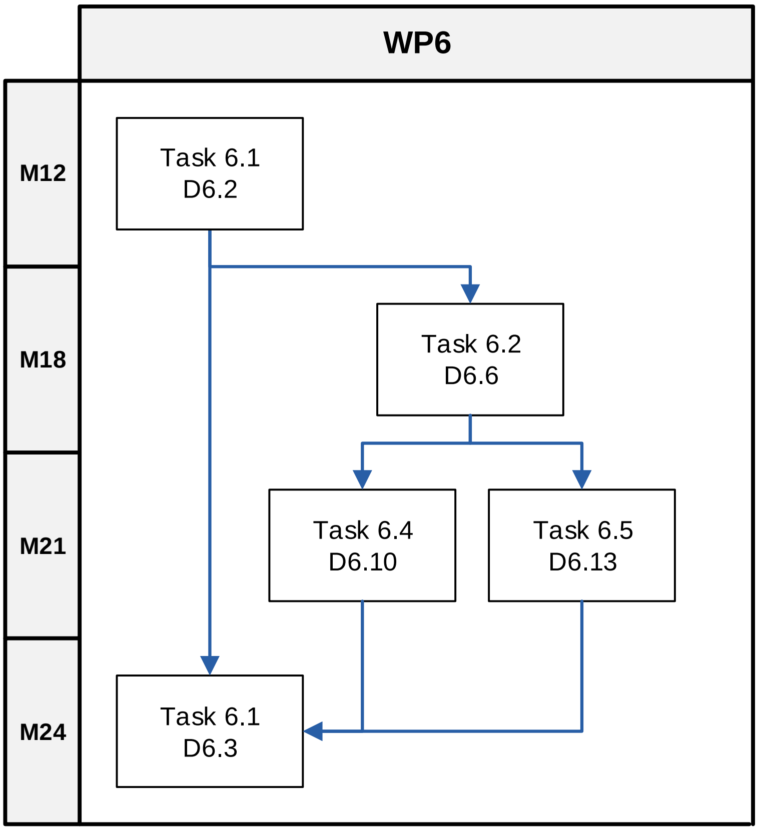


Figure 1. Dependencies between this deliverable and other deliverables.

1.3 Relation to other Work Packages and Deliverables

This document reports the medical evaluation results of the Intermediate Phase Controlled Pilot Trials. These trials were executed at FSL premises during July – September 2016.

The data collected during the trials reported were reported in *D6.6. Controlled pilot trials 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 *D6.10 User Evaluation II* and for medical evaluation reported in the current document *D6.13 Medical evaluation report I*. The evaluation results also include points to be considered in the design of the next piloting plan (D6.3).

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

As described in D6.06, after the patient completed the baseline evaluation, he was asked to perform 10 sessions (5 days of experimental sessions * 2 sessions per day), each of which included the execution of the four ADLs:

- Bed transfer: Lying to Standing
- Chair transfer: Sitting to Standing
- 4-meter walk
- Medication intake

At the end of each session, an email was sent informing clinical staff about whether the ADLs were detected and if so what was the duration of each activity (except pill intake where duration is not relevant). Due to technical reasons, 42 sessions were lost (38 emails were not sent and 4 times the system failed). Other pilot failures, resulted in loosing individual ADL measurements. 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 FSL researchers (ground truth). Details about how the duration measurements were performed and what instructions were given to the people collecting the ground truth are as it follows:

- Bed: the measurement at the ground truth started when the robot produced a ‘beep’ sound signaling that the subject could begin the movement of standing up (e.g., rolling over in the bed) and was stopped when the subject was standing in front of the bed.
- Chair: the measurement at the ground truth started when the robot produced a ‘beep’ sound signaling that the subject could begging the movement of standing up (e.g., moving the back from the seatback) and was stopped when the subject was standing in front of the chair.
- 4-meter walk: the measurement at the ground truth started when the robot produced a ‘beep’ sound signaling that subject could begin walking and was stopped at the 4th meter was covered.

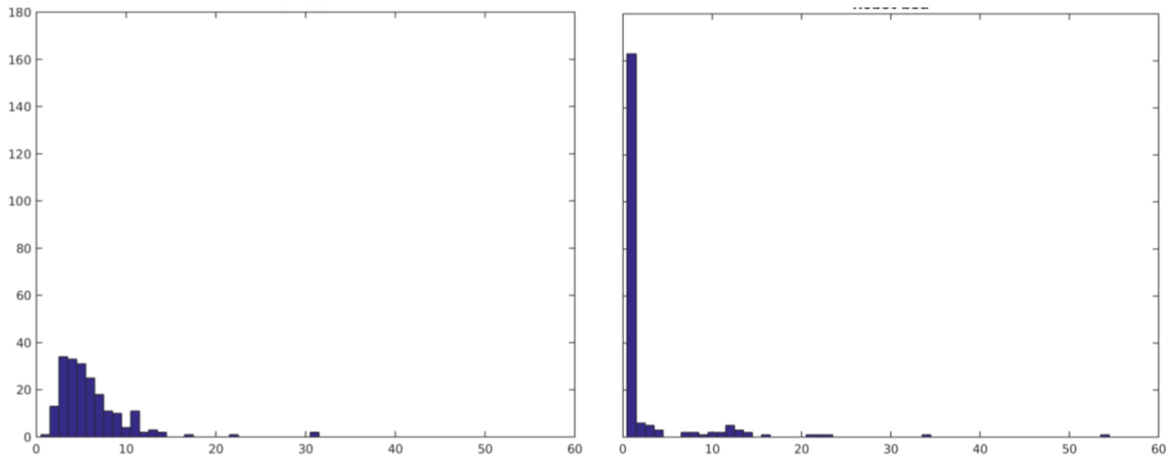
No duration measurement was performed for the medication intake ADL and only the detection (occurred/non occurred) was collected.

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

2.2 ADL detection

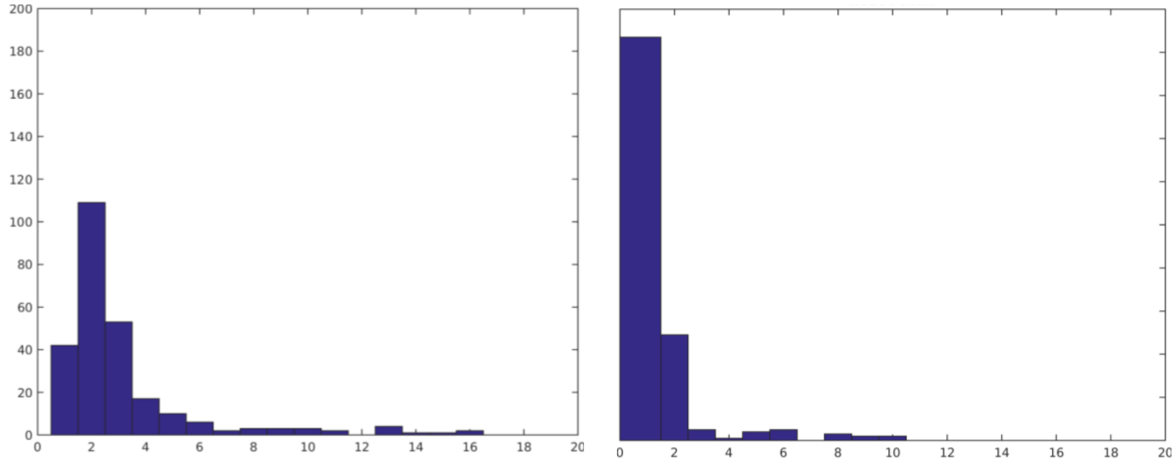
As mentioned before, due to lost emails (38) and system failures (4 times), the evaluation of the system for each ADL is based on 318 sessions (360 = 36 participants x 10 sessions, minus 42). Of those sessions there was a number of occurrences that the system did not detect the event (empty entry in the email). We will refer to these sessions as ***no detections.***

From the detected instances we had to further discriminate between ***correct detections*** and ***wrong (erroneous)***. To do so, we plotted the distribution of the measurements of both the RADIO system and the ground truth. Figures 2 to 4 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.



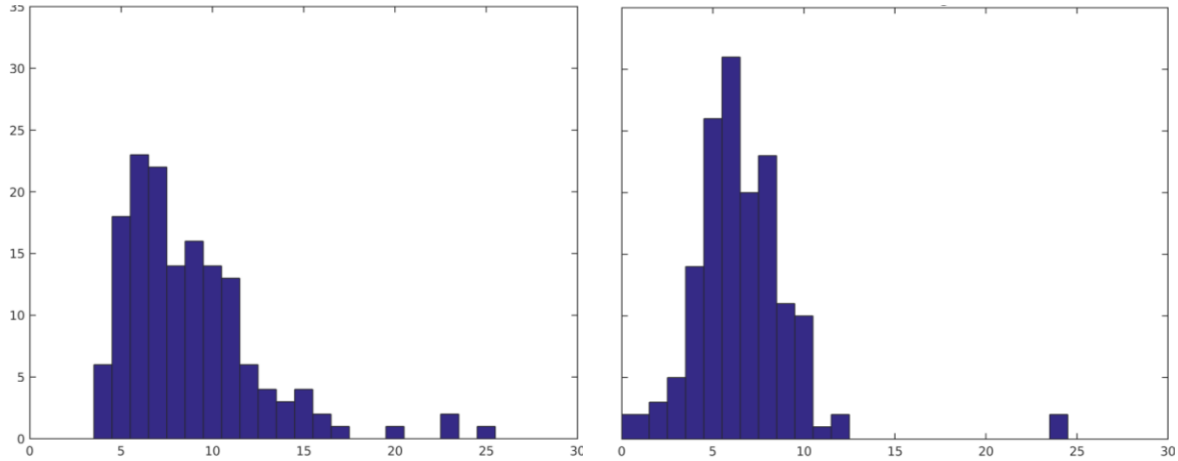
Distribution of ground truth measurements (sec) *Distribution of RADIO measurements (sec)*

Figure 2: Comparison of bed transfer ground truth and RADIO measurements



Distribution of ground truth measurements (sec) *Distribution of RADIO measurements (sec)*

Figure 3: Comparison of chair transfer ground truth and RADIO measurements



Distribution of ground truth measurements (sec) *Distribution of RADIO measurements (sec)*

Figure 4: Comparison of 4m walk ground truth and RADIO measurements

It becomes clear that many of the detected by the system ADLs were erroneously marked as detection as their values fall outside the collected range of ground truth measurements. This is quite evident in case of bed and chair transfer by the highly populated first bins in the RADIO measurements histograms (Figure 2 and 3, right hand distributions). In order to discriminate between correct detections and erroneous ones we assessed ***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 robot **did not detect** an actually occurring event (no email entry).
Chair			
4-meter walk			
Medication intake	The robot 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 this 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 it 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 there exists a correlation we proceed in calculating the linear regression and metrics that inform us about the sources of deviation 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.
    - ○ $LC = (1-r^2) * \frac{\sum y^2}{N}$

3 Results

3.1 ADL Detection

3.1.1 Bed Transfer

For the bed transfer ADL, we analyzed in total 306 sessions. Of these sessions, the RADIO system *did not detect* the ADL *104* instances. The rest of the data, as recorded by both the RADIO system and ground truth are presented in Figure 5[1]. As can be seen in Figure 5, most of RADIO values fall outside the min to max range of ground truth values: min and max values are 1.200 and 31.150 accordingly. Out of 202 actually detected bed transfers, only 47 can be classified as correct detections (true positives), while 155 are classified as wrong detections (false positives).

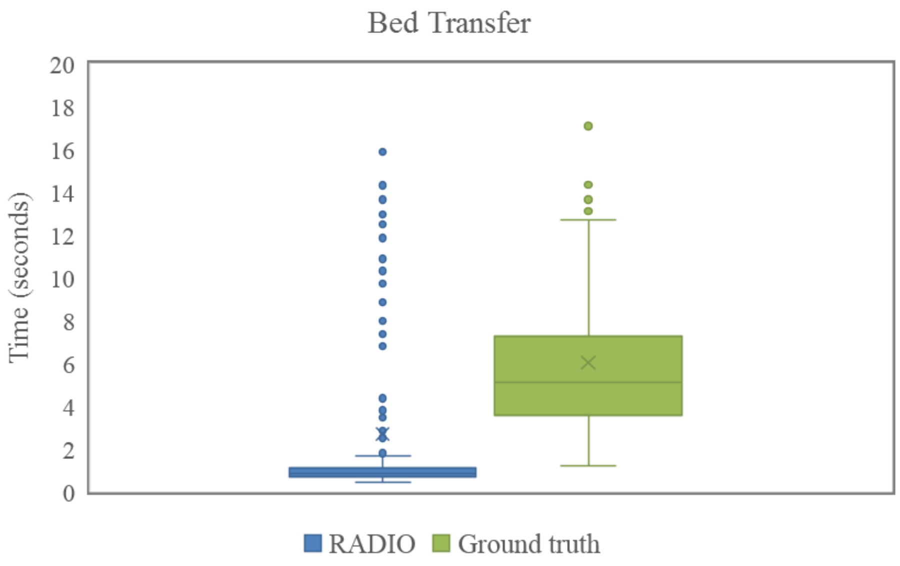


Figure 5. Box plots of detected bed transfers.

3.1.2 Chair transfer

For the chair transfer ADL, we analyzed in total 308 sessions. Of these sessions, the RADIO system *did not detect* the ADL *39* instances. The rest of the data, as recorded by both the RADIO system and ground truth are presented in Figure 6[2]. As can be seen in Figure 6, some of RADIO values fall outside the min to max range of ground truth values: min and max values are 0.830 and 16.180 accordingly. Out of 269 actually detected chair transfers, 139 can be classified as correct detections (true positives), while 130 are classified as wrong detections (false positives).

[1] Six data points that were outliers were omitted from Figure 5 for presentations purpose. There were 4 RADIO recordings (54, 22.65, 21.53, 21.30) and 2 ground observations (22.37, 31.03). None of the outlier points referred to the same session.

[2] Seven data points that were outliers were omitted from Figure 6 for presentations purpose. All were ground truth observations (7.98, 16.18, 13.08, 6.25, 10.35, 14.24, 8.16, 8.95, 13.06)

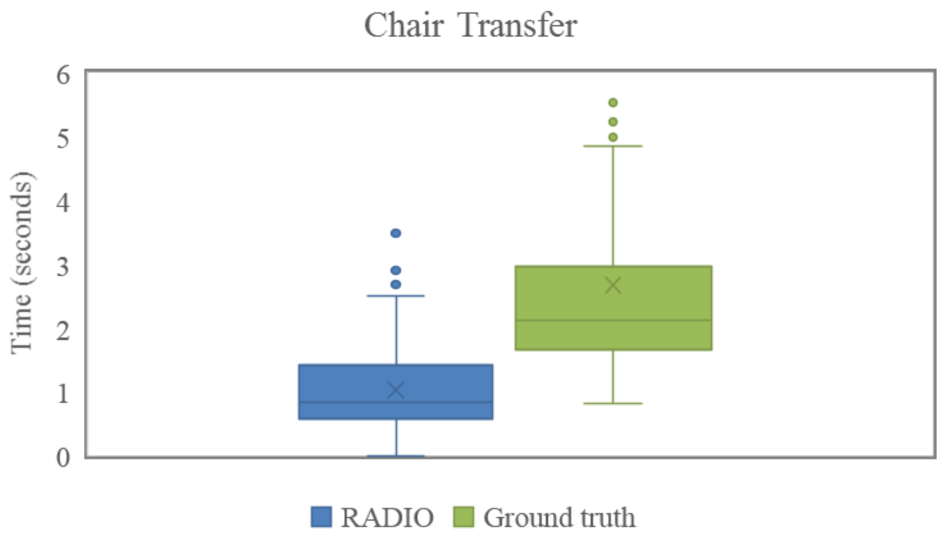


Figure 6. Box plots of detected chair transfers.

3.1.3 4-meter walk

For the 4-meter walk ADL, we analyzed in total 309 sessions. Of these sessions, the RADIO system *did not detect* the ADL *159* instances. The rest of the data, as recorded by both the RADIO system and ground truth are presented in Figure 7. As can be seen, some of RADIO values fall outside the min to max range of ground truth values: min and max values are 3.710 and 25.270 accordingly. Out of 150 actually detected chair transfers, 138 can be classified as correct detections (true positives), while 12 are classified as wrong detections (false positives).

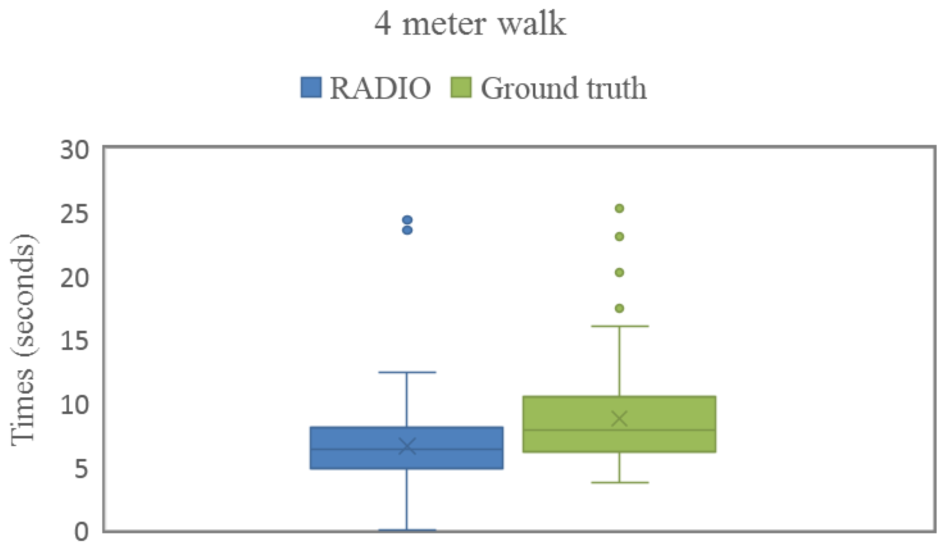


Figure 7. Box plots of detected 4-meter walks.

3.1.4 Pill intake

For the pill intake ADL, we analyzed in total 318 sessions. Of these sessions, the RADIO system *did not detect* the ADL *169* instance and *detected 149* pill intakes.

3.1.5 Overall detection evaluation of the ADL methods

Figure 8 presents the bar charts of detection vs no detection sessions across all methods. Figure 9 presents the correct vs wrong detections again across all methods (besides pill intake as this classification is not applicable in this case).

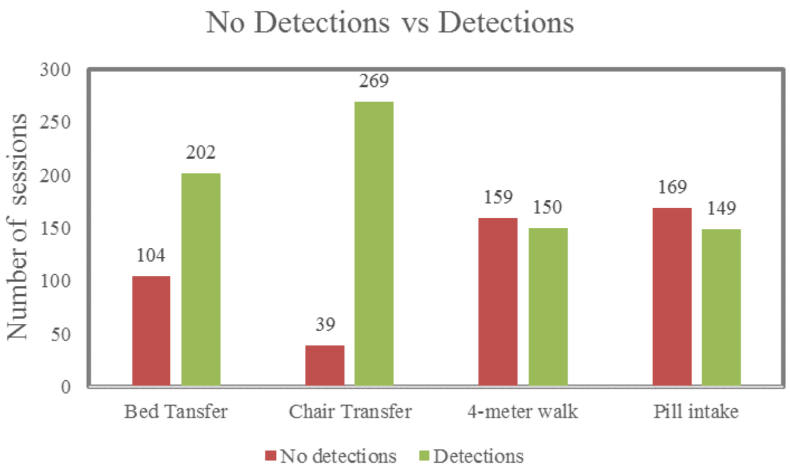


Figure 8. RADIO system's no detections and detections.

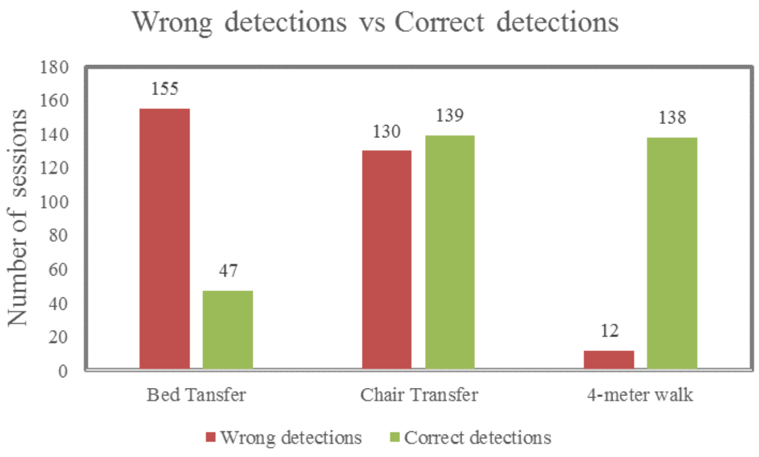


Figure 9. RADIO system's wrong and correct detections.

Table 2 presents the overall results of the detection sessions performed by the RADIO system divided, into correct detection, wrong detection, and no detection. 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**
Correct – True Positives	47	139	138	149
Wrong – False positives	155	130	12	0
No detection – False Negatives	104	39	159	169
Total	306	308	309	318

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

Measure	**Bed Transfer**	**Chair Transfer**	**4-meter walk**	**Pill intake**
Precision	0.23	0.51	0.92	1.00
Sensitivity	0.31	0.78	0.46	0.47
F-measure	0.27	0.62	0.62	0.64

3.2 ADL Duration Measurement

3.2.1 Bed Transfer

Figure 10 presents ground truth measurements against RADIO ones. The points presented in Figure 10 refer to the ***47 sessions*** where RADIO measurements were ***classified as correct detections***. As can be observed there is no correlation between the two groups of measurements (r=-0.048, p=0.496). Based on the plotted points, we could further infer that even though some points were classified as correct detections (based on the rules of Table 1), in reality they constitute erroneously detecting either early movements during the transfer (orange marker) or belatedly marking the completion of the event due to dropped frames and network congestion (green marker).

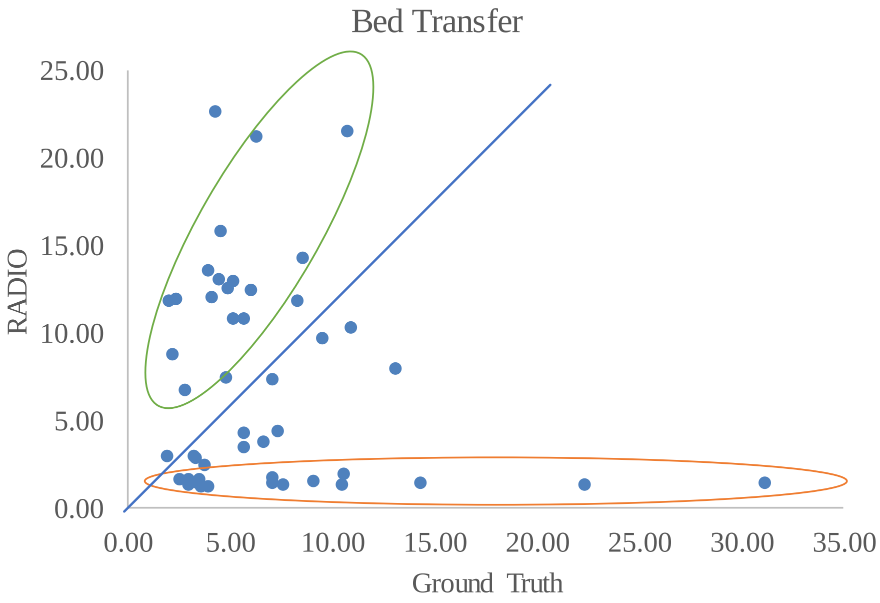


Figure 10. Bed transfer ground truth data versus RADIO measurements.

3.2.2 Chair Transfer

Figure 11 presents ground truth measurements against RADIO ones for the chair transfer. The points presented in Figure 11 refer to the ***139 sessions*** where RADIO measurements were classified as ***correct detections***. As can be observed there is no correlation between the two groups of measurements (r=-0.068, p=0.2.65). There seem to be two clusters (points in green and orange marker) around a mean of 1 and 1.75 of RADIO measurement. We attribute these values to false segmentation of the movement by the object tracking algorithm which makes RADIO calculations fall into different "value bins". This have been addressed since this study by smoothing (cf. page 13, Section 2.2.3, D3.5).

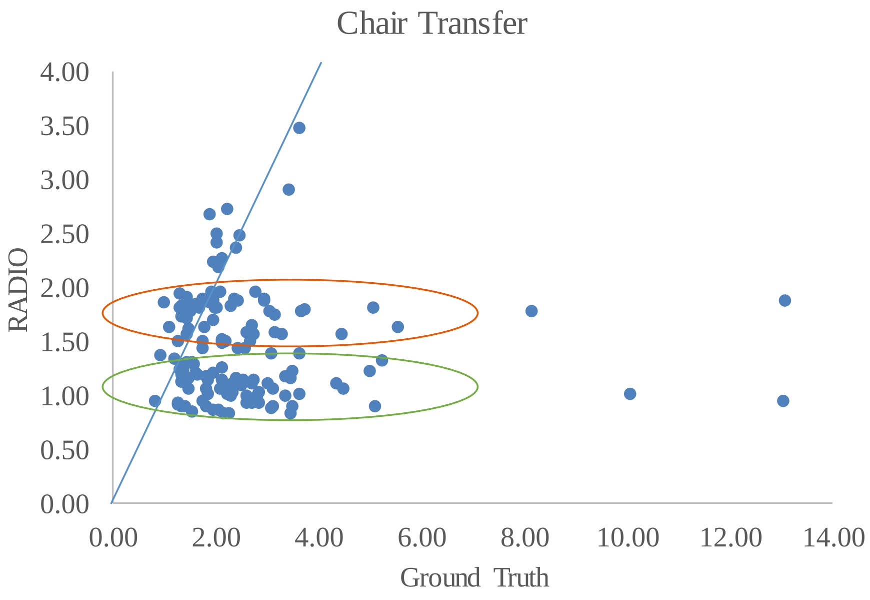


Figure 11. Chair transfer ground truth data versus RADIO measurements.

3.2.3 4-m walk

Figure 12 presents ground truth measurements against RADIO ones for the 4m walk ADL. The points presented in Figure 12 refer to the ***138 sessions*** where RADIO measurements were classified as ***correct detections***. In this case there is a correlation between the two groups of measurements (r=674, p<.001). The linear regression between the two groups of measurements is given by y = 0.5x+ 2.4. The mean standard deviation (MSD) between the RADIO and ground truth measurements is 11. This is partitioned in squared bias (SB) of 3.66 (translation of unity slope), non-unity slope (NU) of 3 (rotation of unity slope) and lack of correlation (LC) of 4.4 (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.

The bias present could be explained by a bias in the ground truth measurements (e.g. delay of stopping the timer) and/or a RADIO method bias coming out of delays in the computational pipeline. It becomes difficult to further distinguish between deviations from “ideal measurements”. Data points in the green marker in Figure 12 show an underestimation (as measured by RADIO) of the ground truth measurement. It is difficult to say what is the origin of these errors. One possible explanation is that it is impossible to set up an experiment where the exact same 4m are measured by the human observer and by the robot, leading to systematically underestimated measurements. This has led to a refinement of the experimental setup for the final pilot studies.

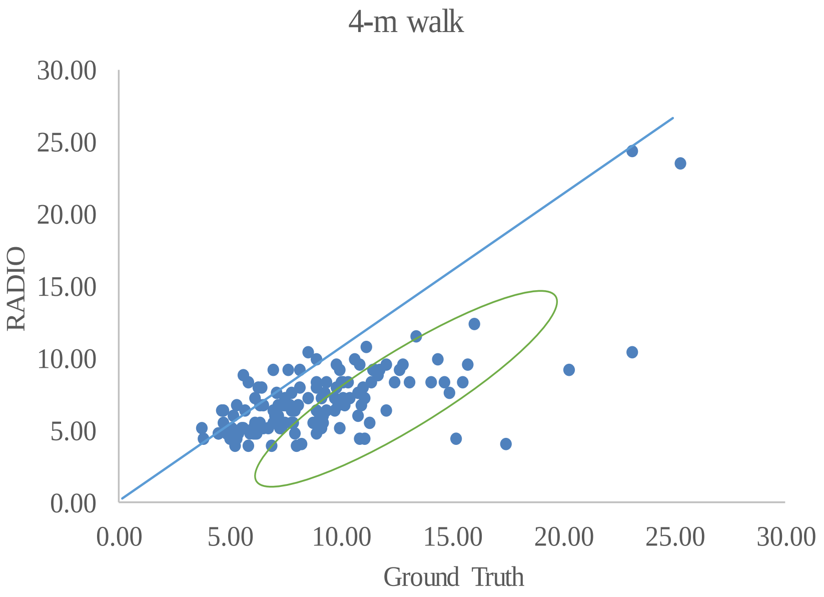


Figure 12. 4 meter walk ground truth data versus RADIO measurements.

4 DISCUSSION

About the *precision* of the system, only pill intake ADL shows a 100% precision, thus indicating the 100% of likelihood that an event recognized as occurred actually did. The other three ADLs present lower percentages, due to the presence of many false positives.

Many of the false positives are also a result of evaluating raw, individual recognition events rather than daily reports where the system would have had the opportunity to merge recognitions or to otherwise sanitize its reporting. What has been observed, in particular, is that multiple instances of the same activity were recognized by the system during the time needed for the subject to carry out the activity once. This is particularly the case for sitting-standing ADL. In getting out of bed, it was also common that reaching out to the bed handlebars was recognized as a separate event that the subsequent getting out of bed. A natural consequence of the above it that the system systematically underestimates the duration of the detected ADLs compared to the ground truth.

As for the *sensitivity* of the system, no ADL recognition achieved the cut-off percentage of 80% [2] as all the four ADL recognition tests present many false negatives. Particularly, many false negatives are present in the 4 meter walk and pill intake ADL. Indeed, during the experimental session of the pill intake ADL, it has been observed that the patient had to take the pills carefully covering the detection box with his body, otherwise most of the times the event was not detected. Interestingly, sitting-standing ADL, presenting the second lowest precision rate after lying-standing, shows the highest sensitivity rate among the four ADLs, being the ADL presenting the lowest number of false negatives.

REFERENCES

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

[2] Lalkhen AG, McCluskey A. Clinical tests: sensitivity and specificity. Contin Educ Anaesth Crit Care Pain. 2008;8(6): 221-223