Evaluation Studies of Robotic Rollators by the User Perspective: A Systematic Review

Christian Werner\textsuperscript{a} Phoebe Ullrich\textsuperscript{a} Milad Geravand\textsuperscript{b} Angelika Peer\textsuperscript{c} Klaus Hauer\textsuperscript{a}

\textsuperscript{a}Department of Geriatric Research, AGAPLESION Bethanien Hospital/Geriatric Center at the University of Heidelberg, Heidelberg, and \textsuperscript{b}Department of Robot and Assistant Systems, Fraunhofer Institute for Manufacturing Engineering and Automation IPA, Stuttgart, Germany; \textsuperscript{c}Bristol Robotics Laboratory, University of the West of England, Bristol, UK

Key Words
Systematic review · Evaluation studies · Ambient assisted living · Robotics · Rollator · Walker · Self-help devices · Human-robot interaction · Mobility · User experience

Abstract
Background: Robotic rollators enhance the basic functions of established devices by technically advanced physical, cognitive, or sensory support to increase autonomy in persons with severe impairment. In the evaluation of such ambient assisted living solutions, both the technical and user perspectives are important to prove usability, effectiveness and safety, and to ensure adequate device application. Objective: The aim of this systematic review is to summarize the methodology of studies evaluating robotic rollators with focus on the user perspective and to give recommendations for future evaluation studies. Methods: A systematic literature search up to December 31, 2014, was conducted based on the Cochrane Review methodology using the electronic databases PubMed and IEEE Xplore. Articles were selected according to the following inclusion criteria: evaluation studies of robotic rollators documenting human-robot interaction, no case reports, published in English language. Results: Twenty-eight studies were identified that met the predefined inclusion criteria. Large heterogeneity in the definitions of the target user group, study populations, study designs and assessment methods was found across the included studies. No generic methodology to evaluate robotic rollators could be identified. We found major methodological shortcomings related to insufficient sample descriptions and sample sizes, and lack of appropriate, standardized and validated assessment methods. Long-term use in habitual environment was also not evaluated. Conclusions: Apart from the heterogeneity, methodological deficits in most of the identified studies became apparent. Recommendations for future evaluation studies include: clear definition of target user group, adequate selection of subjects, inclusion of other assistive mobility devices for comparison, evaluation of the habitual use of advanced prototypes, adequate assessment strategy with established, standardized and validated methods, and statistical analysis of study results. Assessment strategies may additionally focus on specific functionalities of the robotic rollators allowing an individually tailored assessment of innovative features to document their added value.

C.W. and P.U. contributed equally to this work.
Introduction

In older persons, the ability to move independently represents a hallmark of autonomous living [1] and quality of life [2], while being physically active is associated with numerous positive health outcomes [3, 4]. However, sensory, motor or cognitive impairments restrict mobility in frail, older persons [5]. Motor key functions such as standing, walking, or transfers are substantial challenges for their daily activities leading to high risk exposure of falls as documented in residents of senior homes [6]. Effects of motor impairment are augmented by sensory deficits such as visual impairment, leading to restricted functional independence [7], or by cognitive impairment, leading to spatiotemporal disorientation or executive dysfunction [8]. To overcome or compensate for such impairments and to improve the quality of life of affected persons, assistive devices as in walking aids (e.g. canes, walkers, rollators) have been developed with an early focus on mobility support. They provide support of postural stability and mobility [9], reduce risk of falling [10] and improve activity and participation [11]. However, such conventional mobility devices may not cover the needs of persons suffering from major functional or cognitive impairments.

In the context of ambient assisted living (AAL), robotically augmented rollators with various high-tech functionalities have been developed to provide physical, sensory and cognitive assistance, and/or health monitoring for further support [12]. The development and evaluation of such a robotic rollator (RR) is still a new, emerging research field mainly driven by technical engineering goals. However, as technical functionalities translate into assistive devices for the target population, the human-robot interaction and user perspective shift in the development focus. Apart from the sheer technical evaluation of concepts and functionalities, needs, requirements and preferences of potential users will have to guide the development and evaluation of assistive technology devices [13, 14]. In addition to technical testing, which verifies the functional capability of devices, an evaluation with focus on user performance, physical demands and subjective experiences of the RR is essential to prove the usability, ensure safety and demonstrate the added value for the intended user group. The change from technical to user perspective may, however, lead to specific methodological challenges including the study design and assessment strategy. To our knowledge, no systematic review on the evaluation of RRs with focus on the user perspective has been published. Therefore, the aim of this systematic review was to summarize the methodology of studies evaluating the human-robot interaction from a user perspective and to give recommendations for future evaluation studies.

Methods

Initial search terms were compiled and iteratively refined by team members with expertise in the clinical and in the technical research field. The literature search was conducted using the electronic databases PubMed and IEEE Xplore. Search terms included both controlled vocabulary (i.e. MeSH Terms, IEEE Terms) and keywords of relevance identified during searches. The detailed search strategy used in PubMed, which was modified for IEEE Xplore, is presented in the online supplementary table 1 (see www.karger.com/doi/10.1159/000444878 for all online suppl. material).

Manual searches were performed to identify additional studies by scanning reference lists of relevant articles and by reviewing key authors’ own databases. Studies were searched with focus on the evaluation of an RR (or robotic wheeled walker) by experiments, trials, or interventions in human beings independent of the type of outcome measurement. No restrictions regarding age or health status of the subjects were made. Single case reports were excluded. For the purpose of this review, the term ‘robotic’ includes the normal function of a rollator enhanced by additional physical, sensory, or cognitive robotic support while walking, also including sit-to-stand transfers. Studies evaluating solely monitoring functionalities without taking into account any user supporting functionalities or the subjective user experience were excluded. The search was limited to articles in the English language published up to December 31, 2014.

The selection process was conducted following the methodology as described in the method guidelines of the Cochrane Collaboration [15]. Titles and abstracts were identified by the standardized search strategy. For abstracts which met the inclusion criteria or for those with unclear status, full-text articles were analyzed for inclusion. Each step of study selection, based on predefined eligibility criteria, was performed independently by 2 reviewers (P.U. and C.W.). Any disagreements were resolved by consensus or third-party adjudication (K.H.). After inclusion, data on the user group, sample characteristics and the methodological approach were extracted by 1 researcher (C.W.) and confirmed by 2 other researchers (P.U. and D.S.). If an article described more than one study, the results for each study were extracted separately.

Results

A total of 8,989 articles were identified through database searching, and another 79 were added through manual searches. After removing duplicates, the initial search resulted in 8,876 articles. Of these, 235 were found to be related to the search topic based on title and abstract. After reviewing full texts, 148 articles were excluded as they did not meet the predefined inclusion criteria (fig. 1). An-
other 63 were discarded, as these articles described either identical experiments with the same RR, or various stages of development of a certain RR. In both cases, the article providing the most comprehensive information with focus on the user perspective was included. If different articles contained similar information, the one with the most recent development stage was included. Twenty-four articles published between 2001 and 2015 were identified for inclusion in the review. As 2 articles reported on 2 \[16, 17\] and 1 article on 3 independent studies \[18\], the final data extraction was based on 28 studies \(^1\). The detailed review results extracted for each study are presented in the online supplementary table 2, containing information on the names of devices, the definition of user groups, study sample, study object, study design and selected assessment methods.

**User Group Definitions**

Apart from 2 articles \[19, 20\], all mentioned a target user group for the RR; however, their definition differed substantially in accuracy and explicitness. Five articles provided a generic description in broad terms such as ‘elderly (disabled) people’ \[21–25\], 2 defined users by setting-specific characteristics such as ‘persons in nursing and assisted living homes’, partly amended by disease-related criteria (e.g. Alzheimer’s disease, stroke) \[26, 27\], and 10 provided brief information on users’ motor-functional (e.g. ‘with mobility problems’), cognitive (e.g. ‘with cognitive impairment’) and/or visual status (e.g. ‘visually impaired’) \[17, 18, 28–35\], but without staging impairment levels based on any screening or assessment instrument. Three articles described users by disease categories (e.g. Parkinson’s disease, hemiplegia) \[16, 36, 37\] without detailed information on the patients’ functional impairment level. Specific impairment-related definitions based on established, validated assessment methods (i.e. Walking Index for Spinal Cord Injury \[WISCI II\], Functional Ambulation Classification) were documented in only 2 articles \[12, 38\].

**Study Samples**

The mean sample size of the studies was 7.2 (standard deviation (SD) ± 4.3). The exact number of subjects was not reported in 5 studies \[18^1–3, 35, 37\]. No study presented a sample size calculation.

Samples differed considerably regarding age, impairments, or diseases. The age of subjects ranged from 14 \[22\] to 97 years \[31\] with age information lacking in half of the studies (14 of 28) \[16^1, 17^1, 18^1–3, 20, 23, 25, 27–29, 35, 37\].

\(^1\) When necessary, the individual studies of these articles are distinguished with numeric coding (i.e. \[16^1–2, 17^1–2, 18^1–3\]).
Thirteen studies included subjects with motor, functional, cognitive, visual and/or neurological impairments [12, 161, 171, 2, 26, 27, 30–32, 34, 36, 38], whereas a convenient (e.g. ‘ordinary adult males’) [19, 20, 23, 24, 33], mixed (e.g. ‘healthy subjects and subjects with motor and cognitive impairment’) [181–3, 21, 22, 29, 35, 37] or setting-specific sample (e.g. ‘residents of retirement facility’) [28] was used in 14 studies. In studies including impaired subjects, definitions and staging of the severity level of impairment were mostly absent (15 of 20) [171, 2, 181–3, 22, 26, 29–32, 34, 35, 37, 38]. In only 6 studies were motor-functional or cognitive impairment levels defined by established and validated screening or assessment instruments [e.g. Timed Up and Go (TUG), Mini-Mental State Examination] [12, 161, 21, 27, 36].

In 10 studies, subjects did not match with the predefined user group [181–3, 22–24, 27, 28, 33, 37]. However, due to the unspecific and wide-ranging user group definitions given in a number of articles, most studies (15 of 28) were carried out with subjects who were covered by these broad definitions [12, 161, 2, 171, 21, 26, 29–32, 34–36, 38]. In three studies, a user group definition and/or a description of the study sample was completely missing [19, 20, 25].

**Design of Studies**

Depending on study objectives, three different types of studies were performed: (1) observational; (2) comparative, or (3) interventional.

**Observational Studies**

Fourteen articles reported on observational studies [12, 18, 20, 22, 24, 29, 35, 37] or single observational experiments as part of their studies [16, 17, 23, 26, 28, 33], focusing predominantly on the verification of technical capability and/or the subjective user evaluation of the RRs. User performance was used as the study object in only one of these studies [26]. In observational studies/experiments, outcomes were only descriptively presented, without providing any reference values.

**Comparative Studies**

Fourteen articles included comparative studies [19, 21, 25, 27, 28, 30–32, 34, 38] or single comparative experiments in addition to observations [16, 17, 26, 33]. Comparisons were further distinguished into four categories: (1) ‘inter-device comparisons’ in which RRs and conventional devices (e.g. cane, folding/wheeled walker) or fully unassisted walking/sit-to-stand transfers were compared [19, 21, 26, 27, 30, 32, 34, 38]; (2) ‘intra-device comparisons’ in which different assistance levels (e.g. activated vs. nonactivated obstacle avoidance), interface designs, or development stages of the same RR were compared [17, 19, 25–28, 30, 31, 33, 34]; (3) comparisons in a pre/post-test study design with focus on the user experience [34] or the technical functionality [23], assessed before and after/over a series of trials, and (4) comparisons between outcomes of a newly developed robotic monitoring functionality and those of an external criterion measure as a reference measurement [16].

**Interventional Studies**

Two articles described studies that used an interventional approach, providing training opportunities with the RR [16, 36]. In one study, the subjects’ gait performance with the robotic gait assistance system was assessed on 6 consecutive days [16]. However, subjects seemed to use the RR only during test procedures and not in their daily routine. Although the ultimate research hypothesis for this ‘interventional’ approach was lacking, we assumed that the repeated use represented a type of training intervention in order for the subjects to get used to using the RR. In the other study, a 4-week randomized controlled trial was conducted to evaluate the effects of ambulation training with an RR compared to a traditional rehabilitation therapy method using parallel bars [36]. In this study, assessment methods were used to evaluate the subjects’ motor-functional performance after the robot-assisted training intervention.

**Statistical Analysis**

An inferential statistical analysis of outcomes was included in only 3 studies [19, 34, 36]. In 25 studies, outcomes were presented using solely descriptive or qualitative data (e.g. frequencies, means, SDs and user comments) [12, 161, 2, 171, 2, 181–3, 20–33, 35, 37, 38].

**Assessment Methods**

Assessment measures used in identified studies can be classified into five categories: (1) established clinical performance-based measures assessing subjects’ functional ability to perform a requested task by simple quantitative time-, range-, or rating-based outcomes (e.g. gait speed, walking distance, rating score) or by more detailed, qualitative outcomes captured by external technical measures (e.g. step time, double support time); (2) tailored assessment methods in terms of self-designed performance-based measures specifically tailored to specific functionalities of the RR (e.g. guidance system, obstacle avoidance). In addition to simple quantifiable time- or
count-based outcomes (e.g. walking time, number of collisions), these assessment methods predominantly used more technique-based and qualitative outcomes (e.g. path deviation, distance to obstacle); (3) assessment methods used to evaluate the subject’s physical and physiological demands during the use of the RR; (4) subjective evaluation measures to assess a user’s experience with the RR, and (5) technical evaluation measures to assess the technical capability of the RR.

As technical evaluation measures used in 9 studies [12, 16², 18¹, 20, 22–24, 33] exclusively focused on the technical verification of the RR with limited relevance for the user perspective, we do not further address and discuss these measures in this review.

Clinical Performance-Based Measures

Established clinical performance-based measures were used in 3 studies [21, 32, 36]. In one of these, the subjects’ gait and functional performance with the RR were assessed by the 4-meter walk test (4MWT), a modified version of the TUG, and spatiotemporal gait parameters (i.e. step time, double support time) captured by video camera during both tests [21]. Other studies documented the subjects’ motor performance by the 6-min walk test (6MWT), 10-meter walk test (10MWT) and Performance Oriented Mobility Assessment (POMA) [36], or only by the 10MWT [32]. The most frequently used outcomes were gait speed [21, 32, 36], completion time [21], or walking distance and rating scores for functional performance (POMA) [36].

In one study, an established screening test for assessing the functional ability of subjects to perform activities of daily living (ADL) was used (Barthel ADL Index) [36].

Tailored Assessment Measures

In 10 studies, assessment strategies included self-designed performance-based measures specifically tailored to specific robotic functionalities [16¹, 17², 19, 25–28, 31, 34]. Obstacle avoidance and guidance systems were evaluated while subjects completed walking paths [25, 28] or obstacle courses [17², 31, 34], navigation and localization systems while performing navigational tasks [26, 27], and gait assistance systems by analyzing the subject’s gait during robot-assisted walking [16¹, 19]. Simple quantifiable outcomes of these tests included number of collisions [26, 31, 34], reorientations [34], navigational mistakes [27] or abnormal gait patterns [16¹, 2], walking time [34], or achievement of task [26]. More specifically tailored, technique-based outcomes, as used in 8 studies, comprised of deviations from an optimal path [17², 25, 28, 31], distance to obstacles [17, 26], maximum speed and walking distance [26], mean and SD of robot’s velocity [19] and gait variability (i.e. SD of gait speed/step length) [16¹, 2]. To obtain such technically advanced outcomes, 5 studies used the data flow created by the technical systems installed on the RR, including laser rangefinders (LRF) [16¹, 2, 28], a video camera and sonar sensors [17²], or a web camera [31]. In the other 3 studies, information on the technical measure to capture these outcomes was nonexistent [19, 25, 26]. Out of the studies that determined outcomes with the robot-integrated technical systems, only one seemed to process raw data (LRF data) into outcome variables (i.e. path deviation) by using an already established method for robust position estimation of mobile robots in indoor environments (‘Monte Carlo localization’) [28]. In the other 4 studies, it remained unclear whether raw data were analyzed by self-designed or potentially established methods [16¹, 2, 17², 31].

In 2 inter-device comparative studies, a bicycle speedometer attached to the conventional device [16] or an LRF placed in the test environment [26] was used to assess technically advanced outcomes such as walking distance or gait variability also when not using the RR. However, a reference, or any information on the psychometric quality of these methods, was missing in both studies.

In 4 studies including tailored assessment measures, test procedures appear to be nonstandardized [16², 26, 34] or have been insufficiently described [28].

Evaluation of Physical and Physiological Demands

Four studies assessed subjects’ physical and physiological demands with motorized RRs during time-based performance-based measures (i.e. navigational trail, 10MWT) [26, 32] or during walking with standardized gait speed [19, 33]. In 2 studies, the exertion of force applied to steer the RR was measured using the force/torque sensors integrated on the robot’s handles [19, 26]. One also reported on forces required to operate a conventional walker, but did not mention the method to capture these forces [26]. The other study additionally evaluated the oxygen consumption (VO₂) and metabolic cost of transport (metabolic cost per unit of mass and distance travelled) during robot-assisted gait using open-circuit respirometry [19]. In the remaining 2 studies, the muscle activity in the lower extremities was recorded by electromyography (EMG) [32, 33], and one also measured torso kinematics by a tri-axial accelerometer attached to the subject’s back [32].
Subjective Evaluation Measures

Nineteen studies included measures to evaluate the subjects’ experience with the RR [12, 161, 171, 2, 181, 3, 19, 22–24, 26–30, 34, 35, 37, 38]. However, assessment instruments to perform such subjective evaluations varied widely in methodological quality. Nine studies documented solely nonspecific comments of nonstandardized surveys [161, 172, 181, 22, 24, 28, 29, 35, 37], 3 used standardized (dichotomous) questions [27, 30, 38], 4 used self-designed structured questionnaires, each with different multistage rating scales (e.g. 1–5, 0–100) [12, 171, 19, 34], 2 mentioned the use of questionnaires but did not provide detailed information on contents or a reference [183, 26], and 1 presented results of the subjective evaluation by response categories referring to different items but without mentioning the assessment instrument used for this purpose [23]. Most frequently used outcomes of standardized surveys included maneuverability [12, 171, 38], safety [12, 30, 38] and comfort [12, 19, 34].

Discussion

The aim of this systematic review was to summarize the methodology of evaluation studies of RRs with focus on the user perspective. Identified studies showed large heterogeneity in definitions of potential users, study population, study design and assessment methods. We found major methodological shortcomings related to insufficient sample descriptions and sample sizes, lack of appropriate, standardized and validated assessment instruments, and lack of statistical analysis of study results. No generic methodology to evaluate RRs could be identified.

User Group Definitions

The majority of user group definitions seemed inadequate to guide a technical development of an AAL system. Generic, setting-specific, nonspecific impairment-based or disease-oriented definitions do not relate to specific functional impairments of potential users, but cover users with a wide range of different functional abilities and requirements. The effective design of AAL systems in such heterogeneous user groups may not be feasible. The main goal of an AAL system should rather be to overcome or compensate for specific impaired functions. Clear impairment-related definitions are therefore mandatory to specifically tailor AAL developments for specific impairments of users and to ensure that innovative functionalities effectively address a user’s needs. When such specific impairment-related definitions are additionally based on standardized and validated assessment methods with established cutoff values, a general comparability of developments and evaluations will be feasible.

Definitions according to impairment levels will in turn allow specifications such as risk stratification of potential users. With this, the user group will be further classified, opening up the option to exclude persons with no or minor impairment, with no need for assistive devices, or with advanced impairment or unacceptable risk exposure when using the device (triage). Another specification may focus on the main function of the specific device. For example, when an AAL system such as a RR basically supports gait performance, a specific definition based on standardized and validated gait assessment (e.g. 10MWT) will be superior compared to less specific definitions such as general functional scores (e.g. Barthel ADL Index).

As the user group of RRs may be old and multimorbid persons, also highly prevalent age-associated impairments might be included in the definitions, depending on the specific functionalities or complexity of devices (e.g. inclusion of cognitive impairment with respect to navigation functions in disoriented persons).

Study Samples

Overall, sample sizes seemed rather limited to give a consistent picture of the user perspective. Surprisingly, the statistical analysis of documented data was not in the focus of studies as only a very limited number included such analyses (3 of 28) and none of these presented a sample size calculation as a prerequisite of statistical analysis.

A remarkable number of studies (10 of 28) evaluated RRs in persons who were not covered by the predefined user group, considerably limiting the user perspective of these studies. Study results with inadequate, convenient or insufficiently described samples may not suffice to allow conclusions for persons with specific impairments which may represent the potential users of the RR. To ensure that RRs meet a user’s needs and requirements and become successful on the market, it seems mandatory to involve the intended users at all stages of the design and evaluation process of such assistive robotic technologies [39–41].

Design of Studies

Observational Studies

The most heterogeneous group of studies covered observational studies that used solely descriptive data presentations without providing any reference or comparative values. Findings and conclusions of these studies were thus mainly based on the authors’ subjective percep-
tion and appraisal. However, when using standardized and validated outcome measures with well-established cutoff values or other assistive mobility devices for comparison, such observations lose their merely subjective and study-specific nature and enable the objective appraisal of outcomes related to other studies or the documentation of an added value of the RR compared to other devices. From a user as well as a technical perspective, observational studies that descriptively presented non-classifiable or noncomparable outcomes therefore seem to have limited value.

Comparative Studies

The documentation and perception of an added value of the RR is of utmost importance for potential users. Innovative high-tech developments may be fascinating and mandatory for engineering research; however, they may also lead to rather complicated devices for everyday use, not easy to maneuver, too complex to operate, or too expensive to afford. A comparison of RRs with established, low-tech devices ('inter-device comparative study design') may therefore be useful to demonstrate to users the benefit of RR usage.

Comparisons may also be used for the evaluation of single functionalities to document the effect of a specified functionality (e.g. activated guidance system) or the progress of a new development stage. Such an ‘intra-device comparative’ study design allows a tailored assessment of the subjects’ functional performances, physical and physiological demands, and user experience in specific assistance levels or development stages of the RR.

Frail, older persons may initially be intimidated by the robot’s appearance in early stages of development (e.g. without casing, exposed hardware) which may in turn result in a more negative user perception before actually having used the RR. Subjective user evaluations, in a pre-/posttest study design, provide the opportunity to assess the subjects’ initial impressions of the RR and whether there are potentially negative prejudices, which may, however, be overcome after actual use of the RR.

Independent of different types of comparative studies, such a study design should definitely include a statistical analysis to compare results, which was however seldom used in the identified studies.

Interventional Studies

An interventional study design represents a new aspect in evaluation studies with strong focus on the user perspective. Newly developed RRs may not necessarily meet a user’s acceptance or provide usability and efficiency when using them for the first time. Insufficient training opportunities or instruction prior to assessment measures may jeopardize study outcomes [42]. An adequate practice time therefore seems mandatory to prevent initial problems in operating the RR and may further increase the impact on outcomes. Particularly when comparing RRs with a subject’s own conventional assistive devices, brief instructions may not be sufficient, as subjects are already much more familiar and better trained with their own devices.

Overall, we identified a lack of studies investigating usability of RRs in natural environments with adequate long-term evaluation of habitual use. The development and evaluation of RRs seemed to occur rather in engineering laboratories than in clinical settings, as already reported for other robotic assistance systems (e.g. service robots, robotic exoskeleton) [43]. This may be explained by the fact that most of the identified studies evaluated research prototypes in rather early development stages, not yet ready for market launch. In such stages, it is important to manipulate specific variables of a prototype in order to investigate their effects precisely and to optimize technical functionalities accordingly [41]. Since laboratory evaluations also require less time and provide highly standardized conditions, a restricted experimental study design may have been favored. However, for the ultimate goal of RRs to assist mobility of impaired persons in daily life, tests for habitual use seem to be mandatory documenting risk, experience-based perception of use, and quality of life with high relevance for users as well as caregivers.

Assessment Methods

Internationally well-established, clinical performance-based measures allow a worldwide comparability of results, but may be insufficient to cover the particular added value of specific robotic functionalities (e.g. obstacle avoidance, navigation assistance) as the outcome variables do not necessarily refer to the subjects’ abilities potentially affected by the RR [42]. In addition, clinical assessment methods may be limited by subjective rating (POMA) or limited with respect to less detailed, unidimensional outcomes such as gait speed (4MWT, 10MWT) or TUG. Augmenting such measures with technical assessment systems (e.g. video analysis system) allows a multidimensional analysis of the subjects’ gait, including outcomes related to insecure gait or postural (in-)stability (e.g. width of base of support, double vs. single limb support) and reduction of the risk of falls as a main target of RRs.
Even established and validated assessment methods may have their limitations when inadequately used. Outcomes such as gait speed (4MWT) and task completion time (TUG) may be inappropriate when comparing a nonmotorized, conventional device with a motorized RR with limited maximum speed. In such comparisons, a superior outcome for the low-tech device seems almost mandatory and may indicate an insufficient selection of a study outcome. The use of ADL scales (e.g. Barthel ADL Index) to evaluate the effects of a robot-assisted ambulation training appears also inappropriate, since they include, if any, only very few sub-items targeting the subject’s walking ability.

Another potential methodological pitfall may be related to performance-based outcome variables with ambiguous consequences: a motorized RR will improve gait speed in less impaired persons without substantial risk. However, improved performance may be traded off by a substantially higher risk of falling in more impaired persons.

Tailored Assessment Methods

The quality of an assessment strategy substantially depends on the appropriateness of methods with focus on the newly developed functionalities to document the added value of RRs. Clinical performance-based measures may be attractive because of their well-established psychometric properties; however, they have been developed for clinical purpose and may not cover new functionalities in innovative assistive technologies [42]. An assessment strategy specifically tailored to the specific functionality to be evaluated may help to achieve this goal. In RRs, depending on the functionalities installed, a huge data flow created by the robot-integrated sensing technique already exists to control motor or cognitive assistance systems. Using this data flow for assessment purposes may allow highly qualitative and quantitative tailored assessments exactly tuned to the newly developed functionality in order to document the added value of the RR. For example, when focusing on functionalities providing navigational assistance, the data flow from laser sensors, which is used to feed back the position of the RR, could be processed into a superior assessment of walking trajectories during a navigational task. When using such data for the purpose of assessment, it seems mandatory to examine or to provide sufficient information on the psychometric qualities of the robot-integrated sensor technique and the analysis method used to process raw data into the outcome variables. However, it appeared that only 1 study used an already established method for this approach [28]. Furthermore, to ensure reliable, reproducible and comparable outcomes, the test procedure of tailored assessment measures has to be also clearly standardized.

Evaluation of Physical and Physiological Demands

Measures such as EMG, respirometry, accelerometry, or measurements of applied steering forces to the RR allow a detailed insight into relevant physical and physiological effects on objective parameters, which may be indicators for the subject’s individual physical exertion (e.g. VO₂, muscle activity). However, some of these rather laborious measures (e.g. EMG, respirometry) seem less amenable for old and multimorbid persons and may have therefore been used predominantly in studies including only young, healthy adults [19, 33]. To prevent overtaxing by test conditions, alternative methods to evaluate physical exertion are available which may increase amenability by standardized and validated subjective rating [e.g. 44].

Subjective Evaluation Measures

In studies including subjective evaluation measures, a wide range of methods (e.g. nonspecific comments, self-designed questionnaires) related to a variety of different aspects of the subject’s experience with the RR was used which may considerably limit the comparability of outcomes. The overall lack of already established, validated questionnaires for the subjective evaluation of assistive technology [e.g. 45–47] might be due to two reasons: (1) established questionnaires have been developed for a generic evaluation of a wide range of assistive technology devices but may be limited for evaluating specific functionalities of individual devices [45]; (2) some questionnaire items may also be inappropriate to evaluate prototypes after a short-term experiment in a restricted test scenario, covering aspects such as quality of life, usability in daily routine, durability, or services [45–47] whose assessment may only be feasible after habitual use of the devices over an extended period of time. However, the subjective evaluation measures used in the identified studies rather targeted the subject’s actual experience directly after using the RR. This may explain the use of self-developed questionnaires including items already assessable after short-term use in an artificial setting (e.g. maneuverability, safety, ease of use). However, only once these questionnaires have been validated before application and internationally established cutoff values are available, will such assessment instruments guarantee high psychometric quality and allow comparability of study results [48].
Limitations

Only information available in the articles was evaluated in this review, although the authors may have used additional or more detailed methodology, not stated in the articles. The evaluation of AAL prototypes may require elaborate and costly ethical application, and study procedures (‘Medical Product Act’) may have prevented RRs to be tested in comprehensive studies with adequate sample sizes and the target user group as well as in natural environments with adequate long-term evaluation of habitual use. The role of clinical partners in AAL research projects may offer opportunities to solve such problems. Clinical partners may be able to provide specific impairment-based user group definitions, to recruit a satisfactory number of potentially adequate subjects and to investigate the habitual use of AAL systems in natural environments.

Conclusions

Apart from the heterogeneity, methodological deficits in most of the identified studies became apparent. Recommendations for future evaluation studies include: (1) clear definition of the target user group by valid, specific impairment-based criteria; (2) adequate selection of subjects with predefined inclusion criteria representative of potential users; (3) inclusion of other assistive mobility devices for comparison; (4) inclusion of the habitual use of advanced prototypes in evaluation rather than mere short-term, restricted, experimental test scenarios for single functionalities of prototypes not finalized for use in the target user group; (5) selection of established, standardized and validated assessment methods; (6) implementation of a specifically tailored assessment strategy, focusing on specific functionalities of the RR, and (7) statistical analysis of study results. These recommendations, given for RRs, may also apply in general to the development and evaluation of AAL systems with focus on the user perspective.

Acknowledgements

This systematic review was carried out within the 7th Framework Program of the European Union, ICT Challenge 2, Cognitive Systems and Robotics, contract ‘EU-FP7-ICT-2011-9 2.1 – 600769 – MOBOT: Intelligent Active MOBility Assistance RoBOT Integrating Multimodal Sensory Processing, Proactive Autonomy and Adaptive Interaction’. The authors are solely responsible for the content of this review, which does not necessarily represent the views of the European Union. We thank Denis Schrank (D.S.) for assistance with data extraction and Alex Smith for checking the English of the manuscript.

Disclosure Statement

The authors have no conflict of interests to declare.

References

Evaluation of Robotic Rollators

Gerontology 2016;62:644–653
DOI: 10.1159/000444878


