Towards Establishing Clinical Credibility for Rehabilitation and Assistive Robots Through Experimental Design

Katherine M. Tsui and Holly A. Yanco
University of Massachusetts Lowell
1 University Avenue
Lowell, MA 01854
Email: {ktsui, holly}@cs.uml.edu

***Abstract*— The number of clinical trials using robots has increased over the last ten years. It is not practical for all experiments to be clinical trials through the development cycle of a rehabilitation or assistive robot. How can system developers incorporate aspects of the clinical trials to gain credibility among clinicians during development period evaluations?**

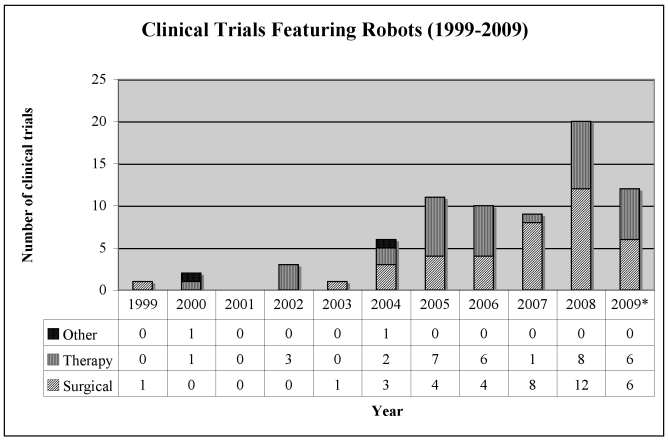
In this paper, we begin a discussion about how to bridge the gap between pre-clinical experiments and Phase 1 clinical trials using rehabilitation and assistive robots. We examine the importance of clearly defined inclusion and exclusion criterion. We also discuss establishing a baseline either with a control group or pre-experiment evaluation and the necessity for first conducting experiments with able-bodied participants. Also, we discuss the need for a common language between the system developers and the clinicians.

I. INTRODUCTION

Traditionally in robotics, the system developers conduct experiments on the physical robots and their control algorithms. Roboticists are primarily interested in performance measures such as time to task completion, accuracy, and power consumption. Thus, the experiments performed are focused around those measures. As robots become more commonplace in the real-world, experiments which have relevance to people who are not the system developers must also be conducted.

Over the last twenty years, robotics research in the domains of medicine and health care have dramatically increased. Many projects are in the development phase; however a number of the rehabilitation robots have made the transition from the laboratory setting to the clinic. Examples include the MIT-Manus [24] and the University of California Irvine's T-WREX [21], which are both upper limb rehabilitation devices. There have been several commercialized systems as well. For example, Intuitive Surgical has shipped over 1,171 units of the da Vinci Surgical System worldwide as of March 2009 [22]. DEKA's iBOT Mobility System power wheelchair sold 400 units in 2007 [9]. Hocoma's Lokomat, used for gait rehabilitation, had over 160 units installed in clinics worldwide as of June 2008 [20].

The number of clinical trials using robots has increased over the last ten years according to the US National Institute of Health's clinical trial listing ClinicalTrial.gov [35]. A total of 76 clinical trials involving robots as interventions were listed as open or completed as shown in Figure 1.[1] The majority of the clinical trials were surgical in nature (40 of 76). Robots have also been used as therapy in clinical trials (34 of 76), especially in the last five years. The remaining two applications of robots in clinical trials were used for telepresecence and as a non-theraputic assistive device. In this paper, we focus on non-surgical health care robots, specifically rehabilitation and assistive robots.



	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009*
Other	0	1	0	0	0	1	0	0	0	0	0
Therapy	0	1	0	3	0	2	7	6	1	8	6
Surgical	1	0	0	0	1	3	4	4	8	12	6

Fig. 1. The number of clinical trials found in ClinicalTrial.gov which used robots as the experimental intervention. Year indicates the study's start date. Data as of June 9, 2009.

Used in the US, European Union, and Japan, the Good Clinical Practice Protocol requires clearly stated objectives, checkpoints, and types and frequency of measurement [34]. It requires a detailed description of the proposed study and preventative biasing measures. The expected duration of the trial, treatment regiment and record keeping strategies must also be detailed. Further, discontinuation criteria for participants or the partial/whole trial must be clearly defined.

As rehabilitation and assistive robots are considered medical devices and can be prescribed to the end-user, it is logical to

[1]Search term used was "robot" and "robotics." Three trials, listed as terminated, are not included in this data.

leverage the existing protocol from clinical trials for medical devices. However, it is not practical for all experiments to be clinical trials through the development cycle of a rehabilitation or assistive robot. The question then is how can system developers incorporate aspects of the clinical trials to gain credibility among clinicians?

In this paper, we begin a discussion about how to bridge the gap between pre-clinical experiments and Phase 1 clinical trials using rehabilitation and assistive robots. First, we describe the different types of human-subjects experiments including clinical trials. We then discuss the importance of clearly defined inclusion and exclusion criterion, establishing a baseline either with a control group or pre-experiment evaluation, and the necessity for first conducting experiments with able-bodied participants. Also, we discuss the need for a common language between the system developers and the clinicians. Throughout this paper, we provide examples from our research as well as that of other research groups to illustrate the spectrum of methodologies, number and type of participants, types of data collected, and performance measures.

II. Progression of Experiments

The goal of all rehabilitation and assistive robots is to provide a service to an end-user. To improve the likelihood of a usable system, developers may create the system with the end-user in mind (known as user centered design [26]). Developers may also engage the end-user in the development of the system (known as participatory design [29]). Regardless of the style of development, continual evaluation is imperative for the success, robustness, and usability of a system.

The evaluation cycle, shown in Figure 2, begins with a user needs assessment which provides the domain grounding for the state-of-the-practice. From this, ideas for improving the state-of-the-practice can be implemented. The system then can be verified through case studies and insights from the case studies can be used to iterate on the system's design. Hypotheses about the quantity of improvement of the new system versus the state-of-the-practice can be formalized in controlled experiments.

A. Exploratory studies

When investigating a new research area, the first experiment that should be conducted is an *exploratory study*, which is a study in which the hypotheses are unknown. The findings of exploratory studies help keep the research area grounded in reality by understanding the state-of-the-practice. Additionally, exploratory studies may be used for data collection. At this stage, the goal is to examine how robots can be used; robots are not actually used during exploratory studies.

One method used for exploratory studies is ethnography. Ethnography strives to understand a situation holistically, employing first-hand observation of participants with unbiased notes about the environment, physical interactions, and conversation. Interviews may also be used to further explore details of a particular instance.

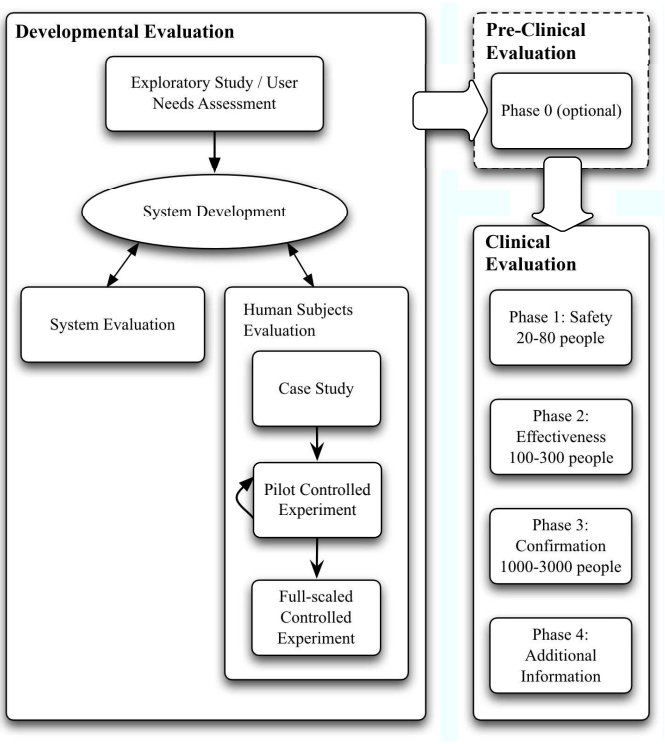


Fig. 2. Evaluation lifecycle of rehabilitation and assistive robots.

An example of an exploratory study that utilized ethnography is Choi et al.'s user needs assessment of eight people with amyotrophic lateral sclerosis (ALS, also known as Lou Gehrig's disease) [6]. The participants were given a digital camera to photograph objects that were dropped or were otherwise unreachable. They were also given a notebook to summarize the circumstances. The participants recorded these events for a period of one week. Afterward, the participants were interviewed with questions relating to the frequency of dropping objects, specifics about the object, and whether or not a care giver was already present in the participant's life. By conducting a user needs assessment of object used in activities of daily living, the researchers were better informed as to what their mobile manipulator El-E would need to be able to grasp.

B. Case studies

We define a *case study* or a *feasibility study* as a demonstration and/or validation of the system as used by a few people, usually one or two users. If the objective is to study the system itself (i.e., control algorithms, physical performance of a particular piece of hardware), any naive, able-bodied, cognitively capable participant can use the system. The system's performance can be compared to the performance of similar systems. However, if the objective is to test if an end-user can use the system, then people from the target population are needed. The quantitative performance measure is largely binary in that either the person was able to use the system or not. Typical qualitative measures include the participants' comments made during the experiment and post-experiment

interviews.

Many rehabilitation and assistive robotics experiments are case studies since it may be difficult to recruit end-users. Because of the small sample size, statistical analysis on performance measures are unlikely significant and conclusions do not scale to the general population. Therefore, the result of case studies is typically anecdotal.

Tijsma et al. conducted a case study of their human-robot interface with the Manus Assistive Robotic Manipulator (ARM) [10] with four end-user participants [30]. The interface was successfully integrated with only two of the four participants' wheelchair joysticks. The participants executed three tasks: picking up an upside-down cup and placing it right-side-up in another and picking up a pen and placing it in the same cup; putting two square blocks in a box of blocks; and retrieving two pens out of sight. Due to fatigue, the participants were only able to perform one trial per experimental condition. Data collected included the number of mode switches, task time, Rating Scale of Mental Effort [41] (at 5, 10, 20, and 40 minutes), and survey responses. The results were anecdotal due to the small sample size and insufficient data.

Case studies are most appropriate towards the beginning of a project; for example, when the first prototype has been completed; in this case, a case study may inform the researchers about the changes for the next revision. If a case study was particularly successful, a pilot study may be planned.

C. Controlled experiments

We define an experiment as "a test or procedure carried out under controlled conditions to determine the validity of a hypothesis or make a discovery" [7]. Controlled experiments are used to compare number of conditions with quantitative performance measures. For example, two conditions are compared in an *AB*-style experiment; that is, the one independent variable may have one of two values. This would be considered a simple hypothesis and is straight forward to test. For a simple hypothesis, users either participate in all conditions (within subjects study [19]) or in just one condition (between subjects study [17]). For a between subject study, a new group of users is needed for each variable tested. In an *AB*-style experiment, there would be two groups.

Some hypothesis can be more complex and may have more than one independent variable. For more complex hypotheses, the learning effect in a within subjects study may be too great or the number of participants needed to obtain statistical significance may be intractably large. Instead, a mixed-model design [18] of both a within and between subjects study may be needed. Bethel and Murphy [2] explain how to choose the appropriate type of study.

A *pilot study* is a scaled-down version of an experiment. In the progression of experiments for rehabilitation and assistive robots, pilot experiments should first be run with able-bodied, cognitively capable participants. As noted by Yanco [40], able-bodied, cognitively capable participants are more easily able to vocalize any discomforts and stop a trial quickly. A pilot is meant to test the experimental protocol itself and to verify that the data collected results in the intended performance measures. However, like case studies, it is difficult to find large populations of end-users. Therefore, many controlled experiments with end-users are pilot-sized with less than twelve participants.

An example of a controlled experiment with end-users exists in our own work [31]. We have created a visual interface for autonomously controlling a Manus ARM; a camera on the robot arm's shoulder provides a live-video feed for the interface. Based on the input device (touch screen or mouse-emulating joystick) and camera view (static or moving), we created four versions of a "flexible" interface. We conducted a controlled experiment for eight weeks with eight participants from the Crotched Mountain Rehabilitation Center. The participants used as many of the versions of the flexible interface as their manual dexterity and cognitive ability would allow. They participated as frequently as possible with a range of one session to eight. We recorded both quantitative data (e.g. trial run time, attentiveness rating, prompting level) and qualitative data (e.g. most/least liked interface, suggestions for improvements).

D. Clinical trials

There are five types of clinical trials: treatment trials, prevention trials, diagnostic trials, screening trials, and Quality of Life trials [36]. Used by pharmaceutical companies, the most well-known type of trial is the treatment trial which evaluate "experimental treatments, new combinations of drugs, or new approaches to surgery or radiation therapy" [36]. Robot-assisted surgery fits the category of treatment trials. Many robot-assistive therapies also may fit into the treatment trials, although some may fit into the category of Quality of Life trials which "explore ways to improve comfort and the quality of life for individuals with a chronic illness" [36].

According to ClinicalTrials.gov, clinical trials have four phases, although a preliminary phase may precede the first official phase [36]. In a Phase 1 trial, the intervention is tested with a small-sized group of participants (20-80 people) to evaluate its safety; the participants may include healthy people [36]. In Phase 2, the intervention is applied to a medium-sized group of participants (100-300 people) to evaluate the intervention's effectiveness and further test its safety [36]. In Phase 3, the intervention is give to large groups of participants (1000-3000 people) to confirm its effectiveness, compare it to other common interventions, and collect information so that the intervention can be safely used [36]. In Phase 4, additional information about the intervention's risks, benefits, and optimal usage is collected [36].

Of the thirty-four clinical trials using therapy robots, sixteen were specifically listed the trial phase with two Phase 0, four Phase 1, five Phase 1/Phase 2, five Phase 2, and one Phase 3. We found that the number of expected participants for clinical trials using robots was smaller than the aforementioned groupings of participants in their respective phases. Of the thirty-four trials, thirty expected to have 10 to 90 participants ($\bar{X} = 36.6$, $SD = 22.8$), which included examples of all

phases. Two trials expected to have 127 and 160 participants, which were appropriately sized Phase 2 trials. Two trials did not disclose the number of expected participants.

Many robot systems are custom-built with custom software. Only one robot system may exist and therefore it is difficult to run a multiple site experiment, which is common in Phase 3 trials in order to recruit the large numbers of participants needed. Conversely, it is easier to conduct a Phase 1-sized experiment at a single site and potentially a Phase 2-sized experiment at a single site given a long enough data collection period. Also, it should also be noted that robots used for clinical trials are expected to work reliably for several hours per day for weeks or months at a time, which is the exception rather than the rule in the current state of robotics.

III. Defining Inclusion and Exclusion Criterion

A rehabilitation or assistive robot must provide a service to a end-user by definition. Experimental design is more complex due to the unique abilities of the end-users, thus generalizations cannot be easily made. For example, two end-users with the same medical diagnosis may present in completely different manners, and conversely two end-users with different diagnoses may be very similar in terms of presentation. Experimental design must therefore consider a person's physical, cognitive, and behavioral ability.

In practice during human subjects evaluation in the development evaluation, participants are usually grouped by their disability or diagnosis. For example, Tijsma et al. [30] worked with four people who used power wheelchairs. Choi et al. worked with eight people with ALS [6]. However, clearly defined inclusion and exclusion criterion show to whom the robot intervention evaluation can be applied.

An exemplary description of inclusion and exclusion criterion can be found in Brewer et al. [3]. The researchers at the University of Pittsburgh conducted an exploratory experiment with thirty people with Parkinson's disease [3]. The inclusion criteria was listed as a physician established diagnosis of Parkinson's disease, a reported loss of motor dexterity, a minimal score of 27 on the Mini Mental State Exam [13], no central nervous system diseases, no experience of abnormal/involuntary movements, and a minimal age of 18. The exclusion criteria was listed as restricted upper extremity movement, loss of sensory information in the hand, loss of vibration in the hand, inability to remain without medication for twelve hours prior to the experiment, and "inability to go to testing facility with companionship or unwillingness to be tested at their home."

Rehabilitation and assistive robots are also used to improve the quality of life for people with cognitive impairments. It is then not sufficient to only provide the participants' disabilities or diagnoses. Additional information, such as behavioral ability, visual ability, and cognitive ability, must also be included in the inclusion and exclusion criterion. Figure 3 shows the profiles of eight participants from our own work as given by our occupational therapist [31].

IV. Establishing a Baseline

Our goal when conducting an experiment is to quantify how much better or how much worse a given intervention is. Establishing a baseline is one method by which we can explicitly derive this quantification using statistical analysis. Traditionally in clinical evaluations, the baseline evaluation of an intervention is a separate control group. The control group would be one of two conditions in an *AB*-style, between subjects experiment. The participants in the control group do not receive the intervention and instead may receive traditional therapy, state-of-the-practice assistive device, or placebo in the case of pharmaceuticals. The participants in the second group do receive the intervention.

A. Using participants as their own controls

Having two conditions allows for an explicit comparison of the average performance of the two groups. However, it may be infeasible to obtain enough participants for two separate groups of large enough size so as to establish statistical significance (see Bethel and Murphy [2] for information on computing sample size). Participants who fit the inclusion criteria during the recruitment period may no longer by the time a study begins.

In this case, a pre-experiment evaluation can serve as the control group and a post-experiment evaluation as the experimental group. We then compare the average performance prior to the intervention against the average performance after the intervention has been delivered. Pre- and post-experiment evaluations can also be used in the case of a separate control group. We are then able to compare the change of average performance of the control group against the experimental group. Using pre- and post-experiment evaluations in this sense is most appropriate for evaluations with multiple sessions over a longer period of time.

An example of using the state-of-the-practice as an active control can be found in Au [1]. Researchers at the Massachusetts Institute of Technology conducted a pilot experiment of a novel lower-limb robotic prosthesis with three unilateral, transtibial amputees. The participants completed three sessions in the experiment. In the first session, the participant was fit with the robotic prosthesis and walked a thirty foot length at a self selected pase. In the second session, the participant walked on an indoor track for five minutes while his metabolic cost of transport was measured via oxygen consumption and carbon dioxide generation. There were three conditions for the prosthesis. In the first condition (baseline), the participants used the robot prosthesis without power. In the second and third conditions, the participants used the robotics prothesis with two different control algorithms. In the last session, the participants walked along a walkway in a motion capture setting. The joint torque, joint angle, and center of mass of the robotic prosthesis and unaffected leg was collected in each of the three prosthetic conditions. As in the second session, the unpowered robot prosthesis served as the baseline.

Another example of using participants are their own controls comes from Housman et al. [21]. The researchers conducted a

	Age	Diagnosis	Cognition	Behavior	Vision	Wheelchair Type	Computer Access
P1	26	Spinal Cord Injury, Traumatic Brain Injury	Not significantly impaired	None	Typical	Manual	Standard mouse, keyboard
P2	60	Traumatic Brain Injury	Distractible	Very sociable; follows prompts well	Left inattention; right visual processing disorder	Manual	Standard mouse, keyboard
P3	17	Spinal Bifida	Good memory and receptive/expressive language	Low frustration tolerance; needs encouragement	Reduced acuity	Manual	Standard mouse, keyboard (limited dexterity)
P4	20	Cerebral Palsy	Good receptive language and expression with AAC; able to learn new skills; mild limitation with problem solving	Aggressive when frustrated; can express need for break	Functional	Manual, Power	Dynavox with keyguard
P5	20	Cerebral Palsy	Below age level; moderate decision making ability	Needs encouragement	Functional	Power	Standard mouse, keyboard (limited dexterity)
P6	37	Traumatic Brain Injury	Challenged by multi-step process; short term memory impairment	None	Functional	Manual	Standard mouse, keyboard
P7	20	Osteogenesis Imperfecta	Mild deficits; slight prompting needed due to vision	Cooperative	Mild perceptual impairment	Power	Standard mouse, keyboard
P8	18	Cerebral Palsy	Mild deficits; slightly below age level; slight prompting needed	None	Functional	Power	Standard mouse, keyboard (limited dexterity)

Fig. 3. Participant profiles for Tsui et al.'s 2008 end-user evaluation of a "flexible" visual interface of a wheelchair-mounted robot arm [31].

small-scale clinical trial of the Therapy Wilmington Robotics Exoskeleton (T-WREX) at the Rehabilitation Institute of Chicago (RIC) and Northwestern University. Housman et al. evaluated twenty-three people post-stroke over sixteen weeks comparing robot-assisted therapy to a traditional rehabilitation therapy regiment. Eleven stroke survivors exercised with T-WREX for one hour, three times per week for eight weeks. An active-control group of twelve patients exercised with a physical therapist for the same duration. Blood pressure readings and pain ratings were taken before and after each session. Time working directly with a therapist was recorded. After eight weeks, the groups switched to allow for subjective comparison thereby allowing each person to act as their own baseline measurement.

B. Using able-bodied, cognitively capable participants

A baseline may be established using able-bodied, cognitively capable participants prior to conducting the end-user participants. Testing with able-bodied, cognitively capable participants may provide a means of normalizing an end-user's performance or may provide an upper performance limit. For example, Römer et al. propose an absolute measure for time to task completion, in which the time is compared to that of an able-bodied person's performance [27].

An example of an experiment conducted with able-bodied participants as an evaluation baseline exists in our own work [33]. We designed a controlled experiment with twelve able-bodied, cognitively capable participants to compare our visual interface with the hierarchical menu provided by the manufacturer for the Manus ARM. The input method for both systems was single switch scanning. We hypothesized that able-bodied, cognitively capable people would have the best possible performance because 1) there would not be any difficulties with learning the operation procedure for each system given adequate training and 2) there would not be any difficulties with physically operating the switch. The participants moved the robot arm towards a target while we recorded the number of clicks and the time to task completion.

V. Clinical performance measures

Existing performance measures for most of assistive robotic technologies do not provide sufficient detail for experimental and clinical evaluations. Tsui et al. [32] surveyed the performance measures of six domains in rehabilitation and assistive robotics: intervention for Autism Spectrum Disorders, eldercare, post-stroke rehabilitation, intelligent wheelchairs, assistive robotic arms, and external limb prostheses. We also discussed the ubiquity of functional performance measures (i.e., relating to an activity of daily living and administered in a realistic setting). Tsui et al. provided guidelines for choosing appropriate and meaningful performance measures:

1) Consult a clinician who specializes in the particular domain.

2) Choose an appropriate clinical measure for the domain. A domain's "gold standard" will provide the best validity to clinicians, if one exists.
3) Include a functional performance measure appropriate for the domain.
4) Choose an appropriate method to capture a participant's emotional and mental state.
5) Consider an appropriate quality of life measurement.
6) Administer the clinical performance measures at least once before and after the experiment or study.
7) Consider coding open-ended responses, comments, and/or video.
8) Concretely define each enumeration on Likert and differential semantic scales.

The guidelines help build a bridge between the current state of rehabilitation and assistive robot experiments and clinical experiments, which is especially true of the second and sixth guidelines. However, the guidelines do not address how to gain clinical credibility so that the system can be tested with the target population. To create clinical credibility prior to working with the target population, we must create a common language.

One means for building a common language between clinicians and system developers is to base the developmental-phase experiments on experiments that would be done in the clinic. For example, Dollar and Howe evaluated their Shape Deposition Manufacturing robot hand by picking up common household objects (e.g., telephone, broom, glass), which were identified as "practice objects" by Klopsteg et al. in 1968 [8, 23]. In Wada et al.'s case study ($n = 1$) of their Robotic Gait Trainer [39] clinical performance measures included the six-minute walk test (6MWT) [16] and the timed get-up-and-go test (TGUG) [37]. In Miller et al.'s pre-clinical evaluation ($n = 6$) [25], the functional testing of an upper limb prosthesis was comprised of a series of standard tests: box and blocks, clothespin relocation, Assessment of Motor and Process Skills (AMPS) [12], and the University of New Brunswick prosthetic function [28].

Another means for building a common language is to correlate the system developers' robotic performance measures to clinical measures. The motivation for building common languages between clinicians and system developers is two-fold. First, it is the primary means by which rehabilitation and assistive robots will gain credibility. Second, it creates the need for a meaningful mapping between robotic performance measures and clinical performance measures. Robotic performance measures have a high-level of detail and can provide continuous evaluation instead of periodic evaluations which are the state-of-the-practice. However, the robotic performance measures may create intractably huge data sets which would likely overwhelm the clinician and would unlikely have much meaning to him/her.

The goal is to correlate the robotic performance measures with existing clinical performance measures. For example, Celik et al. examined trajectory error and smoothness of motion with respect to Fugl-Meyer [14] in the context of post-stroke rehabilitation [5]. At the University of Pittsburgh, Brewer et al. have developed the Advanced Sensing for Assessment of Parkinson's disease (ASAP) protocol, which uses machine learning techniques on sensor data to predict the score of a person with Parkinson's disease on the Unified Parkinson Disease Rating Scale (UPDRS) [4, 3, 11]. At the University of Missouri, Wang et al. have developed a fuzzy logic-based augmentation of an existing evaluation tool, the Short Physical Performance Battery (SPPB) [15] which measures balance, gait, strength, and endurance tasks, to provide finer-grained performance measure for day-to-day monitoring [38].

To better understand how to validate the creation of a common language between clinicians and system developers, we summarized the experimental design and validation procedure Brewer et al. used for predicting UPDRS scores [3]. As previously discussed in Section III, Brewer et al. laid out specific inclusion and exclusion criteria.

The participant exerted force on two 6-axis force/torque sensors using his/her index finger and thumb to track the target wave form displayed on a screen [3]. The experiment varied the wave form and cognitive load. The wave form was either a sine wave or a pseudorandom wave. The cognitive load was one of three conditions. In the first minute, the participant had no cognitive task and only performed the primary task of modulating his/her force on the sensors to follow the waveform. In the second minute, the participant performed the primary task and performed the secondary task of counting backwards from 100 in steps of 1. In the third minute, the secondary cognitive task changed to counting backwards from 100 in steps of 3. The use of a secondary task to gauge cognitive load originated in psychology and has been adopted into human-computer interaction (HCI) and human-robot interaction (HRI) experiments.

Brewer et al. computed three "summary variables": the tremor integral, the root-mean-square error between the wave and the exerted force, and the time delta between the wave modulation and the participant's response of exerted force [3]. Then they computed thirty-six predictor variables for each participant (2 hands $\times$ 2 waveforms $\times$ 3 secondary cognitive task conditions $\times$ 3 summary variables) [3]. To the predictor variables, they applied principle-component analysis and two types of least-squares regression to predict the UPRDS score. They found significant correlation between the predicted and actual UPRDS scores ($p = 0.004$ with $R = 0.54$, $p < 0.001$ with $R = 0.87$, and $p < 0.001$ and $R = 0.78$ respectively).

VI. Conclusion

Over the last twenty years, robotics research in the domains of medicine and health care have dramatically increased and there has been an increase in clinical trials using robots over the last ten years. Full-scale clinical trials are intractable during the development phase. In this paper, we began a discussion about how to bridge the gap between pre-clinical experiments and Phase 1 clinical trials using rehabilitation and assistive robots. We believe that good experimental methodology in conjunction with replication of experiments will help

to establish clinical credibility for rehabilitation and assistive robots.

ACKNOWLEDGMENT

This work is funded in part by the National Science Foundation (IIS-0546309).

REFERENCES

[1] S. Au. *Powered Ankle-Foot Prosthesis for the Improvement of Amputee Walking Economy*. PhD thesis, Massachusetts Institute of Technology, 2007.

[2] C. L. Bethel and R. R. Murphy. Use of Large Sample Sizes and Multiple Evaluation Methods in Human-Robot Interaction Experimentation. In *AAAI Spring Symposium on Experimental Design for Real-World Systems*, 2009.

[3] B. R. Brewer, S. Pradhan, G. Carvell, and A. Delitto. Application of Modified Regression Techniques to Validate a Quantitate Assessment for the Motor Signs of Parkinson's Disease. To appear in *IEEE Transactions on Neural Systems and Rehabilitation Engineering*.

[4] B. R. Brewer, S. Pradhan, G. Carvell, P. Sparto, D. Josbeno, and A. Delitto. Application of Machine Learning to the Development of a Quantitative Clinical Biomarker for the Progression of Parkinson's Disease. In *Rehabilitation Engineering Society of North America Conference*, 2008.

[5] O. Celik, M. K. O'Malley, C. Boake, H. Levin, S. Fischer, and T. Reistetter. Comparison of Robotic and Clinical Motor Function Improvement Measures for Sub-Acute Stroke Patients. In *IEEE International Conference on Robotics and Automation*, 2008.

[6] Y. S. Choi, C. Anderson, T. Deyle, and C. C. Kemp. Human-Robot Interaction Studies for Autonomous Mobile Manipulation for the Motor Impaired. In *AAAI Spring Symposium on Experimental Design for Real-World Systems*, 2009.

[7] Dictionary.com. Experiment. http://dictionary.reference.com/browse/experiment. Accessed Dec. 2007.

[8] A. Dollar and R. Howe. The SDM Hand as a Prosthetic Terminal Device: A Feasibility Study. In *IEEE International Conference on Rehabilitation Robotics*, 2007.

[9] M. Drummond. The End of the iBOT: The Innovative Wheelchair Has Been Taken Off the Market. http://disabilities.suite101.com/article.cfm/the_end_of_the_ibot. Accessed Apr. 23, 2009.

[10] Exact Dynamics. Assistive Robotic Manipulator, 2004. http://www.exactdynamics.nl/. Accessed Mar. 1, 2009.

[11] S. Fahn, R. Elton, et al. Unified Parkinson's Disease Rating Scale. *Recent Developments in Parkinson's Disease*, 2:153–163, 1987.

[12] A. Fisher. AMPS: Assessment of Motor and Process Skills Volume 1: Development, Standardisation, and Administration Manual. *Ft Collins, CO: Three Star Press Inc*, 2003.

[13] M. Folstein, L. Robins, and J. Helzer. The Mini-Mental State Examination. *Archives of General Psychiatry*, 40(7):812–812, 1983.

[14] A. Fugl-Meyer, L. Jaasko, I. Leyman, S. Olsson, and S. Steglind. The Post-Stroke Hemiplegic Patient. 1. A Method for Evaluation of Physical Performance. *Scandinavian Journal of Rehabilitation Medicine*, 7(1):13–31, 1975.

[15] J. Guralnik, E. Simonsick, L. Ferrucci, R. Glynn, L. Berkman, D. Blazer, P. Scherr, and R. Wallace. A Short Physical Performance Battery Assessing Lower Extremity Function: Association with Self-Reported Disability and Prediction of Mortality and Nursing Home Admission. *Journal of Gerontology*, 49(2):M85–94, 1994.

[16] G. Guyatt. The 6-minute walk: a new measure of exercise capacity in patients with chronic heart failure. *Canadian Medical Association Journal*, 132(8):919–923, 1985.

[17] R. Hall. 2×2 Between Subjects Design. http://web.mst.edu/∼psyworld/between_subjects.htm. Accessed Jun. 9, 2009.

[18] R. Hall. 2×2 Mixed Factorial Design. http://web.mst.edu/∼psyworld/mixed_design.htm. Accessed Jun. 9, 2009.

[19] R. Hall. Within Subjects Design. http://web.mst.edu/∼psyworld/within_subjects.htm. Accessed Jun. 9, 2009.

[20] Hocoma. Lokomat® & Pediatric Lokomat® References. http://www.hocoma.ch/en/products/lokomat/referenzen-2008-060/. Accessed Apr. 23, 2009.

[21] S. J. Housman, V. Le, T. Rahman, R. J. Sanchez, and D. J. Reinkensmeyer. Arm-Training with T-WREX After Chronic Stroke: Preliminary Results of a Randomized Controlled Trial. In *IEEE International Conference on Rehabilitation Robotics*, 2007.

[22] Intuitive Surgical, Inc. Intuitive Surgical - Investor FAQ. http://investor.intuitivesurgical.com/phoenix.zhtml?c=122359&p=irol-faq#22324. Accessed Apr. 23, 2009.

[23] P. Klopsteg and P. Wilson. *Human Limbs and Their Substitutes*. Hafner Pub. Co., 1968.

[24] H. Krebs, M. Ferraro, S. Buerger, M. Newbery, A. Makiyama, M. Sandmann, D. Lynch, B. Volpe, and N. Hogan. Rehabilitation Robotics: Pilot Trial of a Spatial Extension for MIT-Manus. *Journal of NeuroEngineering and Rehabilitation*, 1(5):1–15, 2004.

[25] L. Miller, K. Stubblefield, R. Lipschutz, B. Lock, and T. Kuiken. Improved Myoelectric Prosthesis Control Using Targeted Reinnervation Surgery: A Case Series. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 16(1):46–50, 2008.

[26] D. Norman and S. Draper. *User Centered System Design: New Perspectives on Human-Computer Interaction*. L. Erlbaum Associates Inc. Hillsdale, NJ, USA, 1986.

[27] G. Römer and H. Stuyt. Compiling a Medical Device File and a Proposal for an International Standard for Rehabilitation Robots. *IEEE International Conference on Rehabilitation Robotics*, pages 489–496, 2007.

[28] E. Sanderson and R. Scott. UNB Test of Prosthetic Function: A Test for Unilateral Amputees [test manual]. *Fredericton, New Brunswick, Canada, Univ. of New Brunswick*, 1985.

[29] D. Schuler and A. Namioka. *Participatory Design: Principles and Practices*. Lawrence Erlbaum Associates, 1993.

[30] H. Tijsma, F. Liefhebber, and J. Herder. Evaluation of New User Interface Features for the Manus Robot ARM. In *IEEE International Conference on Rehabilitation Robotics*, pages 258–263, 2005.

[31] K. Tsui, H. Yanco, D. Kontak, and L. Beliveau. Development and Evaluation of a Flexible Interface for a Wheelchair Mounted Robotic Arm. In *Internaional Conference on Human Robot Interaction*, 2008.

[32] K. M. Tsui, D. J. Feil-Seifer, M. J. Matarić, and H. A. Yanco. Performance Evaluation Methods for Assistive Robotic Technology. To appear in *Performance Evaluation and Benchmarking of Intelligent Systems*. Editors R. Madhavan, E. Tunstel, and E. Messina. Springer.

[33] K. M. Tsui and H. A. Yanco. Simplifying Wheelchair Mounted Robotic Arm Control with a Visual Interface. In *AAAI Spring Symposium on Multidisciplinary Collaboration for Socially Assistive Robotics*, pages 247–251, March 2007.

[34] US Food and Drug Administration. Guidance for Industry, E6 Good Clinical Practice: Consolidated Guidance. *Federal Register*, 10:691–709, 1997.

[35] U.S. National Institutes of Health. Home - ClinicalTrials.gov. http://clinicaltrials.gov/. Accessed Jun. 9, 2009.

[36] U.S. National Institutes of Health. Understanding Clinical Trials - ClinicalTrials.gov. http://clinicaltrials.gov/ct2/info/understand. Accessed Jun. 9, 2009.

[37] J. Wall, C. Bell, S. Campbell, and J. Davis. The Timed Get-up-and-Go test revisited: measurement of the component tasks. *Journal of rehabilitation research and development*, 37(1):109, 2000.

[38] S. Wang, J. Keller, K. Burks, M. Skubic, and H. Tyrer. Assessing Physical Performance of Elders Using Fuzzy Logic. *IEEE International Conference on Fuzzy Systems*, pages 2998–3003, 2006.

[39] J. A. Ward, S. Balasubramanian, T. Sugar, and J. He. Robotic Gait Trainer Reliability and Stroke Patient Case Study. In *IEEE International Conference on Rehabilitation Robotics*, 2007.

[40] H. Yanco. Evaluating the Performance of Assistive Robotic Systems. *Performance Metrics for Intelligent Systems (PerMIS) Workshop*, pages 21–25, 2002.

[41] F. Zijlstra. *Efficiency in Work Behaviour: A Design Approach for Modern Tools*. PhD thesis, Delft University, 1993.