2008

DIFFERENTIATING EFFICACY AND EFFECTIVENESS IN STROKE REHABILITATION STUDIES: EVALUATION OF A SCALE

Laura L. Zettler
Western University

DIFFERENTIATING EFFICACY AND EFFECTIVENESS IN STROKE REHABILITATION STUDIES: EVALUATION OF A SCALE

(Spine Title: A Scale to Differentiate Efficacy and Effectiveness Studies)

(Thesis format: Integrated-Article)

by

Laura L. Zettler

2

Graduate Program
in
Epidemiology and Biostatistics

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science

School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada

ABSTRACT

Differentiating between efficacy and effectiveness research approaches has become increasingly recognized as a critical step in the evidence-based decision-making process. A critical review of these archetypal approaches, as well as an evaluation of a new tool that purports to distinguish between them, was undertaken. Three raters independently applied the tool to 151 randomized controlled trials that evaluated either a pharmacological or non-pharmacological intervention in stroke rehabilitation. Inter-rater reliability was assessed both for individual items and total scores. Validity was assessed by examining associations between the total scale score and key study characteristics consistent with the effectiveness design. Inter-rater reliability values were sub-optimal for most items; however, there was support for basic scale validity. Further item standardization is required before the scale can be incorporated into the critical appraisal process; however, the tool provides a solid foundation upon which to base further discussion of the differential criteria of efficacy-effectiveness trial design.

KEYWORDS: Randomized controlled trials; Research design; Rehabilitation; Validation study; Efficacy; Effectiveness; Explanatory studies; Pragmatic studies; Evidence-based medicine; Systematic review

CO-AUTHORSHIP

The manuscript derived from this thesis is also co-authored by:

1) Mark R. Speechley PhD, Associate Professor, Schulich School of Medicine and Dentistry, Department of Epidemiology and Biostatistics, The University of Western Ontario, London, Ontario; 2) Robert W. Teasell MD, FRCPC, Professor and Chair/Chief, Schulich School of Medicine and Dentistry, Department of Physical Medicine and Rehabilitation, Parkwood Hospital, St. Joseph's Health Care and The University of Western Ontario, London, Ontario; 3) Norine C. Foley MSc, and 4) Katherine L. Salter BSc, Research Associates, Department of Physical Medicine and Rehabilitation, Parkwood Hospital, St. Joseph's Health Care, London, Ontario.

Dr. Speechley was primarily responsible for generating the idea for the publication, as he found the recently published tool, and also contributed to writing and editing. Dr. Teasell was also involved in the editing process and both he and Dr. Speechley helped in obtaining funding for the student. Ms. Foley and Mrs. Salter served as data abstractors, helped with instrument standardization during the data collection stage, and also contributed to editing the manuscript.

Laura Zettler, the primary author, was responsible for the majority of the work presented in this thesis, including: literature review and searching of library databases; establishment of study objectives and direction of the project; study sample selection and study retrieval; data abstraction; database set-up and data entry; determination of statistical methods and conducting of statistical analyses; writing of thesis chapters and creating of tables and figures.

DEDICATION

I would like to dedicate this work to my boyfriend, Dean Leroux, whose great acts of patience and support, throughout my academic career, have made this thesis possible. I would also like to dedicate this thesis to my family, including my Mother (Marilyn), Father (Allan), and Sister (Lisa), whom have always expressed their sincere pride in my academic endeavors; and to my Grandparents (Marie and Jerome Zettler) and Godmother (Elaine Beaubien) for their motivational and financial support throughout my post-secondary education.

ACKNOWLEDGEMENTS

I would like to acknowledge, and express my sincere gratitude to, the many individuals who were instrumental in the successful completion of this thesis: to my supervisors, Drs. Mark Speechley and Robert Teasell, for their guidance, advice and support, throughout my career as a graduate student, including course work, the thesis topic-generating and writing process, as well funding applications; to fellow research associates, Norine Foley and Katherine Salter, for their motivational support and encouragement as well their contribution to the data collection and manuscript editing processes; and to Drs. Neil Klar and Yves Bureau for their statistical advice. I would also like to acknowledge CIHR, London Life, and UWO for their financial support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

ARHQ	Agency for Healthcare Research and Quality
CI	Confidence Interval
CONSORT	Consolidated Standards of Reporting Trials
COPD	Chronic Obstructive Pulmonary Disease
EBM	Evidence-Based Medicine
EBRSR	Evidence-Based Review of Stroke Rehabilitation
GRADE	Grades of Recommendation, Assessment, Development and Evaluation
ICF	International Classification of Functioning, Disability and Health
ICMJE	International Committee of Medical Journal Editors
IQR	Interquartile Range
ITT	Intention-to-Treat Analysis
NP	Non-pharmacological
P	Pharmacological
QOL	Quality of Life
RCT	Randomized Controlled Trial
SIGN	Scottish Intercollegiate Guidelines Network

CHAPTER 1. INTRODUCTION

1.1 Background and rationale

Randomized controlled trials (RCT) and systematic reviews of RCTs tend to be placed atop most hierarchies of evidence, and are considered to be the gold standards for determining the effects of a given intervention or treatment. Their ranking within the evidential hierarchy and their expected value to evidence-based decision-makers (practitioners, policy-makers, health economists, and patients) are largely based on issues of internal validity and the minimization of bias in the individual studies and aggregated bodies of evidence. While internal validity is certainly of utmost importance in all types of research, concerns have arisen regarding the relevance and generalizability or external validity of this high-quality scientific evidence and the relative neglect of these factors in levels of evidence systems [1]. With growing pressure for decisions in healthcare to be evidence-based, researchers are being urged to consider the perspective of the decision-maker when designing studies and disseminating research findings [2].

The primary objective of a study is an important determinant of its relevance to healthcare decision-makers and the applicability of its results to policy and practice. For some clinical trials, the objective is to estimate the effect of an intervention under highly controlled, optimal circumstances, while for others the goal is to examine how well an intervention works in usual clinical practice [3]. The distinction between studies with these different aims was first made by Schwartz and Lellouch, who used the terms explanatory and pragmatic, noting that the former approach aims to *understand* the biological mechanism by which a treatment has its effect, and that the latter approach aims to enable a *decision* to be made among treatment alternatives [4]. Explanatory trials

have also been labeled efficacy, fastidious [5] or regulatory trials [6], and pragmatic trials are often referred to as effectiveness, management [7], practical [8, 9], or public health trials [6, 8]. Efficacy and effectiveness have gained the most common usage in terms of describing these different study designs and are the preferred terms in this thesis.

A common criticism of efficacy trials is their limited external validity, in that the findings may neither be readily applicable to the intended target population, nor provide the necessary and relevant information required by decision-makers when faced with healthcare decisions [1, 9-12]. Efficacy trials have limited external validity because they typically enroll a highly selective, homogeneous patient sample; are conducted where practitioners and facilities are highly specialized; enforce strict treatment protocols to ensure patient and provider compliance; and have smaller samples that may lack the power to detect small, but worthwhile treatment effects on measures that are meaningful to patients. It should also be noted that while the two types of trials can be conceptualized, it is generally accepted that most studies exist on a continuum [8, 11, 13-17], or at least that hybrid efficacy-effectiveness studies are conceivable [4, 18, 19].

While efficacy trials are useful for understanding the mechanisms through which interventions produce outcomes, and seeing if they *can* work under ideal conditions, there is a growing recognition that healthcare decisions should be based on evidence from more realistic and generalizable effectiveness trials [9, 11, 14, 20-23]. Findings have shown that effectiveness trials infrequently replicate the results of efficacy studies of the same intervention, and patients involved in clinical trials often experience different (usually more beneficial outcomes) than those receiving similar care in usual clinical practice [12, 24-27]. This lack of agreement between the study designs and settings

highlights the importance of basing decisions on scientific evidence that is not only internally valid, but that can also be generalized to the intended target population and answer the questions that are important to health-care decision-makers. Nevertheless, since efficacy trials are regarded more favourably by those who grade research evidence, effectiveness studies may fail to meet the higher standards of evidence and are often excluded from systematic reviews and treatment guidelines [28]. Because decision-makers often consult systematic reviews of available evidence to justify decisions and evaluate treatment options, it would be useful if researchers could be more inclusive of effectiveness trials, and when included, make the distinction between trial types, in their reviews. Furthermore, it may be beneficial to place a more balanced emphasis on internal *and* external validity of findings in [1, 29], or incorporate the efficacy-effectiveness distinction into, evidence grading schemes. The efficacy-effectiveness distinction is also noteworthy in health economic analyses [30]. Specifically, there is growing acceptance that the demonstration of drug cost-effectiveness should most appropriately be conducted using data from effectiveness studies, as unrealistic and possibly biased results can occur when using cost and outcome data derived from efficacy studies [31-33].

Although the literature has outlined the typical characteristics of efficacy and effectiveness trials [6, 14, 15, 17, 19, 34-37], only recently was the first and only *tool* to differentiate between the study types published, by Gartlehner and colleagues [38]. The primary target audience for the instrument would likely be those conducting systematic reviews, and if the criteria could validly and reliably identify the different types of trials, or help to quantify where a study exists on the efficacy-effectiveness spectrum, it could prove to be quite useful for means of critical appraisal and evidence ranking within the

evidence-based decision-making process. Furthermore since the original application of the tool only focused on studies of pharmaceutical treatments that can be implemented in primary care settings, it would be interesting to evaluate whether the same criteria could be applied to studies of more complex, non-pharmacological interventions, and interventions (both pharmacological and non-pharmacological) that, by nature of the condition being treated, are not usually implemented in primary care. These interventions are quite often referred to as "black-boxes", most notably in the area of physical rehabilitation, in that it is difficult to address the specific components that produce the desired outcomes [39]. Therefore, it could be argued that the concept of efficacy may be more difficult to apply to complex interventions, especially if one considers that Schwartz and Lellouch's originally defined aim of efficacy (explanatory) research is to achieve *understanding* of a treatment's mechanism. Thus, the application of this new tool will allow for an evaluation of its psychometric properties and provide the context for a more thorough discussion of the implications of defining the efficacy-effectiveness continuum.

1.2 Objectives of the thesis

The primary objectives of this thesis are as follows:

(1) To provide an overview of the concepts of efficacy and effectiveness approaches in research, and introduce a recently published tool designed to differentiate between the trial types; and

(2) To apply the tool to a sample of RCTs in stroke rehabilitation, in order to:

a) assess its inter-rater reliability and discuss its applicability within a condition where both pharmacological and non-pharmacological interventions are common, and

b) attempt to validate the instrument by investigating associations between key study characteristics and total scale scores.

The secondary objectives of this thesis are as follows:

(3) To address any weaknesses of the scale and make suggestions for its improvement;

(4) To discuss how the differentiation between efficacy and effectiveness may depend on both the condition under investigation (treated in primary care versus specialized tertiary care setting), and the nature of the intervention (pharmacological and/or simple versus non-pharmacological and/or complex); and

(5) To identify how the differentiation between efficacy and effectiveness trials can be incorporated into the evidence-based decision-making process.

1.3 Overview of the thesis

This thesis is organized as follows: Chapter 2 reviews the literature on evidence-based decision-making and recognized short-comings of the current process, and more specifically, it reviews the differentiation between efficacy and effectiveness approaches in research, and describes a recently published tool designed to make this differentiation; Chapter 3 presents a version of a manuscript, that will be submitted to the Journal of Clinical Epidemiology, entitled "A scale to identify effectiveness studies appears to be valid, but reliance on individual ratings is problematic due to sub-optimal reliability"; and

Chapter 4 summarizes the key findings of the thesis and discusses the broader implications of these findings, the strengths and limitations of the thesis, and areas of future research. References, appendices, and a Curriculum Vitae follow the discussion.

1.4 References

1 Persaud N, Mamdani MM: External validity: the neglected dimension in evidence ranking. J Eval Clin Pract 2006;12:450-453.

2 Dobbins M, Jack S, Thomas H, Kothari A: Public health decision-makers' informational needs and preferences for receiving research evidence. Worldviews Evid Based Nurs 2007;4:156-163.

3 Haynes B: Can it work? Does it work? Is it worth it? The testing of healthcare interventions is evolving. BMJ 1999;319:652-653.

4 Schwartz D, Lellouch J: Explanatory and pragmatic attitudes in therapeutical trials. J Chronic Dis 1967;20:637-648.

5 Feinstein AR: An additional basic science for clinical medicine: II. The limitations of randomized trials. Ann Intern Med 1983;99:544-550.

6 Buyse M: Regulatory versus public health requirements in clinical trials. Drug Inf J 1993;27:977-984.

7 Sackett DL, Gent M: Controversy in counting and attributing events in clinical trials. N Engl J Med 1979;301:1410-1412.

8 Depp C, Lebowitz BD: Clinical trials: bridging the gap between efficacy and effectiveness. Int Rev Psychiatry 2007;19:531-539.

9 Tunis SR, Stryer DB, Clancy CM: Practical Clinical Trials: Increasing the Value of Clinical Research for Decision Making in Clinical and Health Policy. JAMA 2003;290:1624-1632.

10 Rothwell PM: External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet 2005;365:82-93.

11 Zwarenstein M, Oxman A: Why are so few randomized trials useful, and what can we do about it? J Clin Epidemiol 2006;59:1125-1126.

12 Collins J: Which Randomized Controlled Trials Are Relevant to Clinical Practice? Obstet Gynecol 2005;106:216-218.

13 Kraemer HC: "Rules" of evidence in assessing the efficacy and effectiveness of treatments. Dev Neuropsychol 2003;24:705-718.

14 Macpherson H: Pragmatic clinical trials. Complement Ther Med 2004;12:136-140.

15 Nash J, McCrory D, Nicholson R, Andrasik F: Efficacy and Effectiveness Approaches in Behavioral Treatment Trials. Headache 2005;45:507-512.

16 Streiner DL: The 2 "Es" of research: efficacy and effectiveness trials. Can J Psychiatry 2002;47:552-556.

17 Tansella M, Thornicroft G, Barbui C, Cipriani A, Saraceno B: Seven criteria for improving effectiveness trials in psychiatry. Psychol Med 2006;36:711-720.

18 Armitage P: Attitudes in clinical trials. Stat Med 1998;17:2675-2683.

19 Fuhrer MJ: Overview of clinical trials in medical rehabilitation: impetuses, challenges, and needed future directions. Am J Phys Med Rehabil 2003;82(10 Suppl):S8-S15.

20 Helms PJ: 'Real world' pragmatic clinical trials: what are they and what do they tell us? Pediatr Allergy Immunol 2002;13:4-9.

21 Roland M, Torgerson DJ: Understanding controlled trials: What are pragmatic trials? BMJ 1998;316:285.

22 Rothwell PM: Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet 2005;365:176-186.

23 Treweek S, McCormack K, Abalos E, Campbell M, Ramsay C, Zwarenstein M: The Trial Protocol Tool: The PRACTIHC software tool that supported the writing of protocols for pragmatic randomized controlled trials. J Clin Epidemiol 2006;59:1127-1133.

24 Bauer MS, Williford WO, Dawson EE, Akiskal HS, Altshuler L, Fye C, Gelenberg A, Glick H, Kinosian B, Sajatovic M: Principles of effectiveness trials and their implementation in VA Cooperative Study #430: 'Reducing the efficacy-effectiveness gap in bipolar disorder'. J Affect Disord 2001;67:61-78.

25 Davidson MH: Differences between clinical trial efficacy and real-world effectiveness. Am J Manag Care 2006;12:405-411.

26 Lagomasino IT, Dwight-Johnson M, Simpson GM: Psychopharmacology: the need for effectiveness trials to inform evidence-based psychiatric practice. Psychiatr Serv 2005;56:649-651.

27 Weisz JR, Weiss B, Donenberg GR: The lab versus the clinic. Effects of child and adolescent psychotherapy. Am Psychol 1992;47:1578-1585.

28 Fritz JM, Cleland J: Effectiveness versus efficacy: more than a debate over language. J Orthop Sports Phys Ther 2003;33:163-165.

29 Atkins D, Fink K, Slutsky J: Better information for better health care: the Evidence-based Practice Center program and the Agency for Healthcare Research and Quality. Ann Intern Med 2005;142:1035-1041.

30 Farahani P, Levine M, Goeree R: A comparison between integrating clinical practice setting and randomized controlled trial setting into economic evaluation models of therapeutics. J Eval Clin Pract 2006;12:463-470.

31 Revicki DA, Frank L: Pharmacoeconomic evaluation in the real world. Effectiveness versus efficacy studies. Pharmacoeconomics 1999;15:423-434.

32 Simon G, Wagner E, Vonkorff M: Cost-effectiveness comparisons using "real world" randomized trials: the case of new antidepressant drugs. J Clin Epidemiol 1995;48:363-373.

33 Bombardier C, Maetzel A: Pharmacoeconomic evaluation of new treatments: efficacy versus effectiveness studies? Ann Rheum Dis 1999;58(Suppl I):I82-I85.

34 Alford L: On differences between explanatory and pragmatic clinical trials. New Zealand Journal of Physiotherapy 2007;35:12-16.

35 Bausewein C, Higginson IJ: Appropriate methods to assess the effectiveness and efficacy of treatments or interventions to control cancer pain. J Palliat Med 2004;7:423-430.

36 Godwin M, Ruhland L, Casson I, MacDonald S, Delva D, Birtwhistle R, Lam M, Seguin R: Pragmatic controlled clinical trials in primary care: the struggle between external and internal validity. BMC Med Res Methodol 2003;3.

37 Sekine I, Takada M, Nokihara H, Yamamoto S, Tamura T: Knowledge of Efficacy of Treatments in Lung Cancer Is Not Enough, Their Clinical Effectiveness Should Also Be Known. J Thorac Oncol 2006;1:398-402.

38 Gartlehner G, Hansen RA, Nissman D, Lohr KN, Carey TS: A simple and valid tool distinguished efficacy from effectiveness studies. J Clin Epidemiol 2006;59:1040-1048.

39 Whyte J, Hart T: It's more than a black box; it's a Russian doll: defining rehabilitation treatments. Am J Phys Med Rehabil 2003;82:639-652.

CHAPTER 2. LITERATURE REVIEW

2.1 Introduction

Due to the pressure that exists for decisions in healthcare to be evidence-based, the use of and demand for high-quality research evidence, by the many decision-makers in healthcare (practitioners, policy-makers, patients, etc.) are growing. Systems and tools are in place to aid decision-makers in the critical appraisal of research evidence and the establishment of levels of evidence for interventions; however these processes and the current high quality evidence-base have been criticized for neglecting concepts of external validity and generalizability, which are important factors to those wishing to incorporate research evidence into their decision-making. One important distinction recognized in the literature is that between efficacy (whether an intervention works in ideal, highly controlled circumstances) and effectiveness (whether an intervention works under the conditions of usual clinical practice) study designs. The latter type of study is becoming increasingly sought after by decision-makers in healthcare, who wish to base decisions on generalizable scientific evidence. The purpose of this chapter is to provide a general review of the literature regarding the uses of research evidence in healthcare decision-making, and the systems that establish levels of evidence in evidence-based decision-making. More specifically, this chapter will review the factors that differentiate between efficacy and effectiveness study designs; describe a new tool that purports to distinguish between the designs; and provide the rationale for making this distinction, in terms of the implications it has for the evidence-based decision-making process.

2.2 Evidence-based decision-making in healthcare: uses and users of research evidence and defining the "best evidence"

According to Sackett, evidence-based medicine (EBM) is defined as "the conscientious, explicit, and judicious use of the current best evidence in making decisions about the care of individual patients. The practice of evidence-based medicine means integrating individual clinical expertise with the best available external clinical evidence from systematic research" [1]. The application of the theory of EBM has extended beyond the realm of clinical practice to most aspects of decision-making related to healthcare. Peer-reviewed studies focusing on EBM and evidence-based decision-making have grown in abundance over recent years with a staggering number of citations retrieved in Pubmed (as of June 13, 2008) for the keywords evidence-based medicine (47,708 citations; introduced as a Mesh term in 1997), evidence-based practice (17,436), and systematic review (1,366,846).

In order to define the best evidence for healthcare decision-making, various systems have been developed to assist in grading the quality and strength of research evidence [2]. A large majority of these systems are based on a hierarchy of study design, assigning the highest level of evidence to randomized controlled trials (RCTs) and systematic reviews of RCTs and lower levels to observational studies and expert opinion [3, 4]. Several tools exist to evaluate the quality of research evidence, and they are largely based on the assessment of methodological and reporting quality of individual studies [5]. In 1995, when the assessment of trial quality was a relatively new practice, Moher and colleagues identified 23 published scales and 9 checklists that had been developed for RCT quality assessment [6]. Likewise, a recent systematic review of quality rating scales for RCTs

identified 21 scales and their modifications, this not including the abundance of quality *checklists* for RCTs, and tools used to assess the quality of other study designs [5]. Once individual studies have been critically appraised, the reviewer can make a determination regarding the strength of the aggregated evidence for a particular intervention [7], which is often categorized according to the following terms: strong, moderate, limited or conflicting, and no evidence [8]. These levels of evidence may then be incorporated into clinical guidelines and treatment recommendations.

Systematic reviews of RCTs, placed atop most hierarchies of evidence, and often considered the "best" research evidence [9], have been described as being at the heart or cornerstone of the EBM philosophy [10, 11]. They are an important tool and source of information for healthcare decision-makers [9, 10, 12-15], if not the necessary first step in the decision-making process [16]. These reviews are useful to a variety of decision-makers including clinicians, researchers, and economic and decision-analysts [15]. Within health policy, systematic reviews form the basis of health technology assessments (evidence-based policy) [17]; are commissioned when policy makers must tackle complex issues [16]; and can be used to investigate viable alternatives or justify implemented policy choices [14]. If not used to support health-care decisions, systematic reviews can at least "isolate[e] the state of science," by revealing problematic research and gaps in the evidence, and highlighting where evidence and practice are at odds [3]. Many organizations have recognized the important role of systematic reviews in healthcare decision-making. For instance, the Agency for Healthcare Research and Quality (AHRQ), has established Evidence-based Practice centers across North America

that perform systematic reviews in hopes of helping decision-makers translate evidence into policy or practice [18].

Tunis and colleagues maintain that the demand and use of high-quality scientific evidence is growing, noting the various users (patients, physicians, payers, health care administrators, and public health policy-makers) and uses (physician/patient decision-making, choosing plans or physicians, practice guidelines, quality measurement and improvement, product purchasing and formulary selection, benefit and coverage decisions, organizational and management decisions, program financing and priority setting, and product approval) of research evidence in decision-making [19]. Furthermore, Taylor et al. highlight the greater desire for policy-makers in particular to use research evidence to support decisions regarding hospital formularies, health insurance, guidelines of professional bodies, and cost-effectiveness of interventions [17].Thus, systematic reviews, and research evidence in general, have the potential to greatly impact the decisions made within the healthcare forum.

2.3 Putting research into practice: recognized shortcomings of high-quality research evidence and the current evidence-based decision-making process

There have been several criticisms of the current state of the evidence-base as well as of the systems that are frequently used to critically evaluate research evidence for use in healthcare decision-making. These shortcomings are largely related to the perceived relative neglect of external validity or generalizability of scientific evidence by researchers and reviewers, as well as the differing perspectives of researchers and users of research. It is noted that the systematic review is deemed by many to be the best source of information for clinical and policy decision-making and, over time, there have been

many improvements in systematic review methodology and reporting; however, clinicians, policy-makers, and patients may use systematic reviews less frequently than one would think [20]. Laupacis and Straus suggest that as much energy that is devoted to promoting methodological rigour in systematic reviews should now be focused on making them useful to decision-makers [20]. The AHRQ also support this notion in stating that if systematic reviews are to have an impact on policy and practice they must not only be comprehensive and rigorous, but produce evidence that is relevant to decision-makers in healthcare [18].

The generalizability and relevance of evidence deemed to be of high-quality, namely RCTs and systematic reviews of RCTs, are often questioned. Ideally, systematic reviews should ask clearly defined clinical questions, review evaluations of clinically relevant interventions conducted in representative patient samples and examine outcomes that are meaningful to patients and practitioners [21]; however, it could be argued that this is not common practice. There may be a reluctance on the part of decision-makers to incorporate research evidence into policy and practice due to the focus on highly selected and unrepresentative groups of patients [11, 17, 20, 22, 23] and practitioners [20]; surrogate or composite endpoints [22]; and outcomes that are irrelevant to patients [24] and policy-makers [12]. Furthermore these studies may inadequately report adverse events or measures of harm [20, 22-24]; fail to emulate routine practice conditions in terms of treatment protocols [17, 22, 23] and care settings [9, 22, 23]; use faulty comparators (such as placebos) [17, 22] or lack direct treatment comparisons [12]; be too short in duration [20]; and lack information on treatment costs [12]. Green further delineates these points when he asserts that "best practices" must be more than the

measurement of immediate or intermediate outcomes that are not linked to patient health outcomes and "[i]nvestigator-centered studies in unrepresentative patient samples" [25]. Taylor and colleagues note that current systematic reviews, focused on placebo-controlled RCTs, are not very useful to policy-makers, who want to understand the comparative effectiveness of therapies and combination therapies [17]. Additional concerns regarding the generalizability of systematic review evidence are the potential for funding bias (the inclusion of studies funded by commercial interests) and publication bias (the failure to include *all* RCTs on a given topic) [24].

Several studies have attempted to assess the external validity of RCTs and systematic reviews in specific treatment areas, most of which have focused on the representativeness of the patient samples studied in clinical trials. For instance, it was found that only 1 in 20 people with COPD identified from a large general population survey would have met the entry criteria for the major RCTs that have informed COPD guidelines, suggesting the potential low external validity of this "gold standard" clinical resource [26]. Similarly, another study suggests that the clinical trials on which hypertension treatment recommendations are based, enrolled patients who were substantially different in many respects from those treated in general practice [27]. Also, Kandzari and colleagues found key differences in clinical characteristics and care patterns among acute coronary syndrome patients enrolled and not enrolled in clinical trials [28]. A systematic sample of exclusion criteria in RCTs published in high-impact medical journals revealed that a large segment of patient populations are excluded from investigations, often on the basis of common medical comorbidities and medication use, age, and female sex; it further revealed that drug treatment trials and multi-center trials were most likely to have

extensive exclusions [29]. Rothwell maintains that the limited external validity of RCTs is a main reason for the under-use of effective treatments in clinical practice [23]. Others have agreed, noting that a greater consideration of the external validity of research findings could facilitate the use of evidence by practitioners and policy-makers [30] and that studies that consider both internal and external validity are important for real-world decision-making [31].

The reluctance of decision-makers to take full advantage of the available scientific evidence highlights the differential perspectives of researchers and research-users. Researchers ask different types of questions than clinicians [20] and policy-makers [12], and all have different preferences for receiving research evidence [14, 20]. It has been suggested that reviews could prove more useful to decision-makers if researchers addressed clinically relevant questions, reduced the complexity of the published format, and provided their findings in a clinical context [20]. Dobbins and colleagues further emphasize the need for researchers to take account of the informational needs and preferences of their target audience [14]; in order to encourage the application of research findings to policy and practice decisions and facilitate evidence-based practice and policy, researchers must ensure that evidence is relevant to clinicians and policy-makers and that research results are generalizable [9, 14]. Gruen and colleagues believe that systematic reviews would be more useful to policy-makers if reviewers provided an evaluation of the generalizability of review findings, which would involve addressing the relative importance of the health problem; relevance of the outcome measures; and practicality, appropriateness, and cost-effectiveness of the intervention [32]. It has also been suggested that in order to maintain the validity of systematic reviews, appraisal of

individual studies should take place from the perspective of the clinician as well as the reviewer, so that important clinical details are not overlooked [33]. Fritz and Cleland state that defining the "best available evidence" in EBM will depend on the perspective of the reviewer, and whether they place more importance on the minimization of bias or clinical relevance [34].

Despite the wide recognition of the importance of the generalizability of research results, current evidence ranking systems often fail to capture this element that is so pertinent to decision-makers in healthcare. For example, in an evaluation of three systematic reviews of sciatica therapy, Hopayian found that they were all rated to be of high quality according to a validated scale, even though reviewers disregarded important clinical details in the individual studies, such as the relevance of the patient population, appropriateness of the intervention, and adequacy of the outcome measures [33]. Similarly, in a systematic review of RCT quality rating scales it was discovered that the majority of items addressed by the scales were related to internal validity, with very few dedicated to the assessment of external validity ("relevant outcomes were used" and "follow-up period adequate" being two exceptions); they further concluded that many of the scales were not adequately developed or validated and that their results should be interpreted with caution [5]. The heavy emphasis on internal validity and relative neglect of external validity and generalizability in quality of evidence ratings, best practice recommendations, and study reporting requirements have also been recognized by others [23, 25, 31, 35-38]. After critically appraising six current systems used for grading levels of evidence and strength of recommendations, The GRADE (Grades of Recommendation Assessment, Development and Evaluation) Working Group found important

shortcomings in all systems, including the inability to be used by all target decision-making groups (patients, practitioners, policy-makers) and address different questions, as well as low reproducibility of assessments [39]. Furthermore, with the numerous "levels-of-evidence" criteria that are available, Ferreira and colleagues have recognized the need for a single, reliable and valid system to be developed; this was determined after the application of different criteria to the same sets of studies led to different conclusions regarding the strength of the evidence [8]. Juni et al. reached similar conclusions when they applied 25 different quality rating scales to 17 RCTs within a meta-analysis, and found that the type of scale used had a dramatic impact on the interpretation of the meta-analysis [40]. Lohr also reviewed tools for rating the quality and strength of evidence and concluded that reviewers should match the topic and types of study under review to an appropriate grading tool because no one tool can be used for all cases [7].

In order to address the concerns of research users regarding the external validity of scientific evidence and to facilitate evidence-based decision-making, adaptations of current evidence-ranking systems [2, 41] as well as criteria to assess the external validity [23, 36-38] of individual studies have been proposed. Critical appraisal techniques should include an assessment of internal validity and external validity or generalizability [11, 24], and it has been suggested that measures of external validity could be used alongside the scales that are used to rate internal validity or study quality [36]. Furthermore, Rothwell recommends that there should be stricter external validity requirements for pharmaceutical licensing; an increased focus on external validity within CONSORT and Cochrane standards for reporting; and finally that the International Committee of Medical Journal Editors (ICMJE) require a "to whom do these results apply" section in submitted

manuscripts [23]. In an attempt to account for different study designs in research and the applicability of the available evidence, and to address the weaknesses of existing systems, the Scottish Intercollegiate Guidelines Network (SIGN) Grading Review Group developed a new system for grading recommendations in evidence-based guidelines [41]. Within this system a "Grade A" recommendation requires the highest level of evidence (based on study design and methodological quality) *and* the body of evidence must also be directly applicable to the target population [41]. Similarly, to address limitations that they had previously encountered, the GRADE Working Group developed a new system for grading levels of evidence, which emphasizes that in order assess the quality of evidence, one must take into account study design, study quality, consistency of results across studies, as well as the *directness* of the evidence [2]. The latter criterion refers to how alike the subjects, interventions, and outcomes within the body of evidence are to those of interest [2]. The AHRQ has also recognized the need for a balanced approach to assessing the strength of research evidence, which should include evaluations of study design, internal validity, consistency of findings, strength of association, directness of the evidence, and generalizability of the findings to the target populations [18]. Since the application of research evidence requires the evaluation of external validity, Persaud and Mamdani suggest that both internal and external validity should have equal consideration in evidence rankings, as opposed to viewing the latter of secondary importance, left for determination by the clinician [37]. Lastly, many recognize the complexity of healthcare decision-making, and realize that it may not always be appropriate to adhere to a rigid hierarchy of evidence [11, 42]. In a circular and integrative view of EBM for the study of complex interventions, different research methods can represent pathways for different

research questions and multiple methods are required in order to achieve a balance between internal and external validity of research evidence [42]. There is also support for a more flexible hierarchy, in which RCTs and observational studies can play complementary roles in establishing levels of evidence [4]. In a similar regard, Green and Glasgow have called for systematic reviews to weigh the wider range of research evidence, not just the strongest controlled studies [36].

2.4 Part of the debate: efficacy versus effectiveness approaches in research

In 1967, Schwartz and Lellouch were the first to make a clear distinction between two different aims in clinical research, which they termed explanatory and pragmatic [43]. In this seminal paper, explanatory and pragmatic approaches in clinical trials are differentiated in terms of the different kinds of research questions they address, and the resulting implications for treatment definitions, outcome assessment, subject selection, and statistical analyses. With regard to treatment definitions, the differentiation between explanatory and pragmatic approaches is framed in two ways: 'equalized' or 'optimal' conditions and 'laboratory' or 'normal' conditions. 'Equalized' or 'optimal' refers to either fixing or optimizing contextual factors of treatment between study groups – the former allowing a key treatment component to be studied, which provides information on its true effect, and the latter allowing the comparison of two modes of treatment, each of which absorbs the contextual factors inherent in their administration. 'Laboratory' or 'normal' refers to the treatment protocol, that can either require adherence to exacting conditions or to ordinary current practice. Regarding outcome assessment, the explanatory approach involves the measurement of one or more criteria of biological importance, whereas the pragmatic approach requires the assessment of a single criterion

of practical importance. The selection of subjects in an explanatory trial is strict, so as to achieve a homogeneous sample, with few expected withdrawals that will be excluded from analyses. On the other hand, the pragmatic approach adopts a wide subject selection that results in a heterogeneous sample, with many expected withdrawals that are incorporated into the analyses. Finally, the explanatory and pragmatic approaches offer two different methods of comparing treatment groups. In the former, statistical comparisons and power analyses aim to reduce Type I and II error probabilities and tests of significance are conducted. In the latter, the aim is to reduce Type III error probability (concluding that the inferior treatment is better), and significance tests are not required, as there is no consequence to choosing either of the treatments if they are in fact equal. Thus, the overall differentiation of explanatory and pragmatic comes down to the different problems they address: one seeks to understand the effect of treatment and verify a biological hypothesis and the other seeks to make a decision regarding the best mode of treatment.

In a similar regard, Archie Cochrane in 1971 made the distinction between 'effectiveness' and 'efficiency' [44]. He defined effectiveness in terms of research results, specifically, an intervention's effect in beneficially altering the course of a disease. Efficiency, on the other hand, he defined as an intervention's effectiveness when implemented in usual clinical practice, taking into consideration the general patient population, management strategies, and the optimum use of resources. Today, a similar distinction is often made; however, the terminology has changed somewhat, as noted in a work commemorating the 25th anniversary of Archie Cochrane's book [45]. Herein, it is stated that effectiveness is now termed efficacy, referring to results derived from RCTs in

idealized settings; effectiveness refers to the effect of an intervention on patient health in everyday settings; and efficiency implies cost-benefit, which incorporates effectiveness with the optimum use of personnel and resources [45]. Haynes further summarized the distinctions between efficacy ('can it work?'), effectiveness ('does it work?'), and efficiency ('is it worth it?'), making note that efficacy and effectiveness are both referred to in terms of the beneficial effect of an intervention, with the former under optimum, highly controlled conditions, and the latter under real-world conditions that more closely emulate usual clinical practice [46].

The writings of Schwartz and Cochrane have been the basis for further discussion of the methodologies employed by the different types of studies. In the literature, the term explanatory is now often used synonymously with efficacy, as is pragmatic with effectiveness [47-59]. Efficacy trials have also been referred to as fastidious [60] or regulatory trials [61] and effectiveness trials have been labeled management [62], practical [19, 63] or public health trials [61, 63]. Regardless of the terminology used, there are various but related ways in which the differences between efficacy and effectiveness approaches in research have been framed, including, but not limited to:

- understanding the biological mechanism of treatment versus making a decision between treatment modalities [43];
- whether a treatment works under optimal versus usual practice conditions [64];
- laboratory versus field research [65]
- underlying difference in effectiveness between treatments versus practical difference between treatment policies [66];
- pure versus applied science [67];

- evidence required for regulatory versus public health purposes [61];
- science versus technology [68];
- study of pharmaceuticals versus complex interventions [69];
- maximizing of internal versus external validity [51];
- RCT versus observational study designs [70];
- question of central tendency versus question of mediation and moderation [63].

It is also understood that efficacy and effectiveness are not likely a strict dichotomy, rather they exist as two polar opposites of a spectrum [34, 53, 56-58, 63, 69, 71-73]. Hoagwood and colleagues suggest a dimensional model of efficacy and effectiveness, that allows bidirectional movement between the two genres of study, and consists of three variables (intervention, outcomes, and validity), each of which ranges along a continuous scale [74]. Furthermore, hybrid study designs (blended methodology that retains elements of efficacy *and* effectiveness studies) are possible [43, 49, 64, 69]. Lastly, it is common in theory and practice to establish efficacy of an intervention prior to establishing its effectiveness [46, 49, 70, 72]. Despite the view that efficacy and effectiveness exist on a continuum, the literature tends to describe the two poles in terms of archetypal characteristics of the two research designs (See Table 2.1).

Table 2.1 Archetypal characteristics of efficacy and effectiveness studies cited in the literature.

Characteristic	Efficacy/Explanatory	Effectiveness/Pragmatic
Research question or objective	Aimed at understanding the true biological effect or mechanism of a treatment (verifying a scientific hypothesis) [19, 43, 48, 49, 52, 54, 55, 61, 67, 68, 75, 76] How does a therapy produce its effect and can it work under ideal or restricted circumstances? [77] Evaluate the causal link between treatment and response [73] To determine the beneficial effect of an intervention under ideal or optimal	Aimed at making a decision regarding the best mode of treatment [19, 43, 48, 49, 54, 67, 68, 75, 76, 78] Confirming the benefit of treatment in a population at large [61] What are all the consequences of therapy and does it work under usual clinical circumstances? [77] Evaluate treatment response and feasibility in a real-world setting [73] Conditions of study closely match those of the target practice venues to which study results will be applied [84]

Characteristic	Efficacy/Explanatory	Effectiveness/Pragmatic
	conditions [34, 46, 48, 49, 51, 53, 56, 58, 63, 64, 70-72, 78-81] Estimate efficacy and safety of a specific clinical intervention [57]	To determine the beneficial effect of an intervention in ordinary (real life) clinical practice [46, 49, 51, 53, 55, 56, 64, 72, 79-81] To estimate relative benefits and risks of approved treatments, clinical interventions, programs or policies [57] To determine the impact of a new maneuver when introduced into practice [89] What changes in service delivery are required if efficacious treatment is to be delivered to the widest populations? [71]
Timing of study	Early stages of treatment development or for regulatory approval or licensing [46, 48, 49, 51, 52, 61, 63, 64, 70, 73, 74, 80] Before intervention is introduced [57] Associated with Phase *II* and *III* clinical trials [73, 81]	Later stages of treatment development; after efficacy has been established [48, 49, 52, 61, 70, 73, 80] Formulary approval [64] Post-implementation [57] Associated with Phase *IV* clinical trials [19, 73] Could be conducted prior to efficacy studies [74]
Type of condition best studied	Acute conditions [53, 69]	Chronic conditions [53, 69]
Typical characteristics of intervention under study	Simple interventions or effects [43, 53, 67] Pharmacological interventions [69, 82] Single activity or modality [63, 74] Newly developed intervention or a modification of a well established one [49] Intensive, specialized, standardized interventions [83]	Complex interventions or treatment policies [43, 53, 67, 82] Non-pharmacological interventions [69, 82] May be a single activity but involves the complex interactions between patient and provider [51] Multiple modality [74] "Black box" treatments [79] Brief, feasible interventions not requiring great expertise [83] Widely implementable and feasible interventions [79, 84] Interventions where placebo control is difficult or impossible to achieve [52] Easily adaptable/exportable to usual care setting [55, 72, 73] Clinically relevant and feasible interventions [38]
Implications for study design	RCT [19, 43, 48, 49, 51-54, 56, 57, 60, 61, 63-65, 68-70, 72, 75-77, 79-81, 84-86] ...small and rigorous [61, 68] ...blinded [49, 51-54, 57, 69, 73, 76, 79, 81, 86] ...randomization at the individual level [51]	RCT [19, 38, 43, 48, 49, 51-57, 60, 61, 63, 64, 68, 69, 72, 73, 75-77, 79-81, 84-86] ...large and simple [61, 68] ...open-label (perhaps blinded outcome assessment) [51, 53, 54, 57, 63, 64, 69, 73, 76, 79, 81, 86] ...stratified randomization according to sample heterogeneity [84] ...cluster randomization often necessary [51] Non-randomized designs (quasi-experimental or observational studies) [38, 49, 55, 63, 65, 70, 73, 79]
Study setting	Large tertiary care referral-based health centers [51] One setting with many resources and expert staff [83] Experimental setting or research clinic [53, 54, 69, 74] Highly controlled and specific clinical	Variety of practice settings to which the study will be applicable [19, 49, 63] Multiple settings [68, 83] Routine care setting [53, 54, 69, 80] Less controlled and representative clinical practice setting [55, 73] Real world or naturalistic settings [17, 74, 87]

Characteristic	Efficacy/Explanatory	Effectiveness/Pragmatic
	research setting [55, 73] Specialty academic or commercial research settings [63, 87] Academic hospital [55] Small number of experienced research sites [57]	Community hospital [55] Dozens of routine treatment sites [57] Clinical setting where the therapy is most likely to be used [52] Representative sample of settings [38]
Participating clinicians	Highly skilled, rigorously trained, closely monitored and supervised [55, 56, 73, 88] Skilled or specialized in the delivery of the intervention under study [63, 69, 79, 86] Motivated [81]	Variable levels of skill, training, monitoring, supervision, and experience [49, 55, 73] Usual providers [69, 79, 86, 87] Variety of physicians to which the study will be applicable [19] Representative sample of providers [38]
Who pays for cost of intervention	Provided at no cost to subjects [49, 63, 79] Patients typically compensated for their time and effort [73]	Implementation should not depend on utilization of research resources [84] Reliant on customary sources of sponsorship [49]
Subject sample	Strictly selected (numerous inclusion and exclusion criteria) [43, 46, 48, 49, 51-53, 55-57, 60, 61, 63, 64, 67, 68, 70, 72, 73, 76, 77, 80, 81, 85-88] Arbitrarily defined [43] Homogeneous [43, 49, 51, 53-56, 60, 63, 64, 68-70, 72, 73, 79, 80, 86-88] Low withdrawal rate (high compliance, motivation, responsiveness) [43, 46, 51, 56, 57, 60, 64, 73, 77, 86, 87]	Widely selected (few inclusion and exclusion criteria) [17, 19, 38, 43, 48, 49, 51, 53, 55-57, 61, 63, 64, 67, 68, 72, 73, 76, 77, 80, 81, 84, 86, 88] Exclusion criteria based on safety only [63] Exclusions on the basis of comorbidities, concurrent medications, and lower adherence should be avoided [38] Heterogeneous and representative of the target population [19, 38, 43, 48, 51, 53-57, 60, 64, 68-70, 72, 73, 78, 79, 84-86, 88] Many withdrawals (varying levels of compliance, motivation, responsiveness) [43, 51, 57, 64, 77]
Sample size	Small to moderate [61] Less than 1000 subjects [55] A few 100 subjects at most [57] Small samples usually sufficient [53, 69, 76] Large because statistical determination depends on Type I error [67, 75]	Large to huge [61] 1000 to 10,000 subjects [55] Larger ...to enable detection of small effects [19, 57] ...due to sample heterogeneity [76, 81] Large [49, 52, 53, 69] Smaller because statistical determination depends only on Type III error [67, 75]
Research protocol	Contextual factors equalized between groups (balanced contrast) [43, 53, 55, 60, 69] Strict, ideal (laboratory) conditions not usually met in clinical practice [43, 57, 61, 75, 77] Fixed [56, 57, 60, 64, 76] or flexible [57] dosage regulations Highly standardized or structured, clearly defined or specified and/or manual-based treatment protocol [48, 49, 53, 55-57, 63, 69, 73, 74, 79, 79, 81, 86, 88]	Contextual factors are optimized for each group [43, 53, 69, 75] Normal clinical practice conditions [43, 48, 57, 61, 75, 77, 85] Flexible dosage regulations [57, 60, 63, 64, 76] Highly flexible, loosely defined, individualized treatment protocol according to practitioner's discretion [48, 53, 54, 69, 86, 88] Treatment reflects therapy as it is actually given [56, 63, 81] Less structured treatment protocol [74]
Comparator	Placebo [46, 49, 51-53, 57, 60, 63, 64, 69, 70, 76, 81, 87, 88] Active control [57, 81] No intervention [49, 75, 88] Arbitrarily chosen comparator [64]	Active control [57, 60, 76, 87] Standard intervention or usual care [17, 49, 51, 57, 69, 75, 78, 88] Least expensive alternative [64] Best current treatment [46, 54, 64]

Characteristic	Efficacy/Explanatory	Effectiveness/Pragmatic
	Standard intervention [49, 55, 57, 70]	Clinically relevant alternative [19, 53, 63, 78, 85] No treatment [78]
Concurrent treatments	Concurrent or adjuvant therapies prohibited or very limited (therapy offered in isolation) [57, 80, 88]	Allows for the integration of concurrent or adjuvant therapies [56, 57, 80, 88]
Outcomes	Outcomes of biological importance that are linked to mechanism of action [43, 54, 61, 64] "Hard" endpoints [60] Outcomes are condition-specific [64] Outcomes with a short-term horizon [64] Short-term clinical outcomes [79] Intermediate outcomes [54] Single, objective, and often laboratory-based outcome [52, 69] Specific dichotomous or continuous outcome [51] Reliable measures that bear little meaning to patients [72] Specific characteristic [73] Symptom scales and other clinical parameters [57] Laboratory or biomedical endpoints [81] Surrogate outcomes [48] Outcomes that correspond to disease symptoms [63, 74]	Outcomes of practical importance that are meaningful to patients [43, 48, 53, 61, 69, 87] Patient-focused outcomes [17] Safety, risks, or adverse events [38, 43, 52, 55, 57, 63] "Soft" endpoints [60] Outcomes are comprehensive [64] Outcomes weakly linked to mechanism of action [64] Outcomes with short and long-term horizons [64] Long-term clinical and morbidity outcomes [79] Outcomes that represent the full range of health gains [54] Outcomes that are easy to measure and relevant [85] Broad range of health outcomes [19, 38, 63, 74] Economic or cost-effectiveness outcomes (costs, health-care resource use) [17, 19, 38, 52, 57, 63, 74, 81, 84-87, 89] Outcomes that are meaningful to practitioners [52, 81] Specific primary outcome; secondary outcomes used to explain mechanism of an intervention's effect [51] Clinically valid measures that are often unreliable [72] Clinically meaningful broad construct [73] Single well-defined clinically important primary outcome and multiple secondary outcomes [57] Outcomes that imply low respondent burden [63, 76, 84] Systems variables [74] User-friendliness [89] Patient satisfaction and burden [38] Behavior change at the patient and practitioner level [38]
Compliance with intervention or adherence to protocol	High or carefully measured [61] Ensured, maintained or encouraged [48, 51, 55, 56, 63, 79, 81]	As is or not fully measured or monitored [56, 61] Is an outcome to be measured [51, 63, 89] Fidelity to manual a key variable [57]
Follow-up/ study duration	Short follow-up [49, 53, 55, 69] Duration of 1-4 months [57] Short duration [63, 74, 81]	Longer follow-up [19, 49, 53, 55, 69, 87] Duration of 6 months or more [57] Long duration [74]
Analytical procedures	Aims to reduce Type I and II error [43] Analysis involves significance tests [43] Analysis is restricted to those who actually received intended treatment (per-protocol and/or completers only analysis) [43, 56, 64, 67, 69, 73, 75-77, 80, 81]	Aims to reduce Type III error [43] Significance tests are not necessary [43] Analysis includes all participating subjects according to their initial group allocation (intention-to-treat analysis) [43, 48, 53, 54, 56, 64, 67-69, 75, 76, 80-82, 84-86] Analysis includes only patients actually treated and subgroup analyses according to degrees of

Characteristic	Efficacy/Explanatory	Effectiveness/Pragmatic
	Analysis by intention-to-treat [48, 56, 60, 63, 70, 73, 80-82] Analysis of central tendency [63]	compliance and regulation [60] Analyses of mediation and moderation [63] Analyses should account for data not missing at random [84] Analyses should account for sample heterogeneity (variables that affect prognosis or those that may or may not be associated with the outcome(s)) [84, 85] Subgroup analyses [49, 55, 86]
Validity	Emphasis on maximizing internal validity [34, 49, 51, 53, 55, 61, 63, 69-71, 73, 74, 78, 79, 81]	Emphasis on maximizing generalizability/external validity [49, 51, 53, 55, 57, 61, 63, 69-71, 73, 78, 79, 81] Internal validity is ensured; external validity is maximized [74, 84] May not be generalizable at all [82] Results are frequently next to meaningless [78]

2.5 Evidence-based decision-making and the efficacy-effectiveness distinction

It has been suggested that pragmatic clinical trials (effectiveness studies) are better designed to meet the informational needs of healthcare decision-makers than explanatory trials (efficacy studies) [19, 52-54, 90] and that the usefulness of randomized trials to decision-making lies on a continuum, from explanatory to pragmatic trials [58]. Macpherson states that the aim of pragmatic trials is to inform policy-makers, practitioners or patients when choosing between interventions [53]. Efficacy studies, on the other hand, are noted to have important limitations for informing clinical and policy decisions [79]. Policy debates increasingly require information that is directly applicable to the target population [79] and evidence-based practice requires the evaluation of therapies under normal practice conditions [3], namely effectiveness studies. The notion that systematic reviews and evidence syntheses are largely focused on efficacy and not effectiveness, makes them necessary but insufficient tools for clinical decision-making [3]. Collins maintains that efficacy trials can highlight promising interventions, but effectiveness trials reveal interventions that are robust enough to overcome the heterogeneity of clinical circumstances, and it is the latter that practitioners should seek

out in order to find the best available, relevant evidence that is most likely to benefit their patients [48]. Pragmatic trials are said to be of high relevance to policy-makers because they tend to focus on the same criteria of effectiveness that are used by policy-makers, such as user-perceptions, important and visible outcomes, usual health service planning entities, and typical service limitations [90]. Due to their high emphasis on internal validity and the minimization of bias, and despite criticisms of their generalizability, efficacy studies tend to be judged more favorably by researchers establishing levels of evidence; this often results in the exclusion of effectiveness studies from many systematic reviews and practice guidelines, leaving their clinical applicability questionable [34].

Existing gaps between research and practice have led researchers, practitioners, policy-makers, and government officials to call for more practice- and policy-relevant research [31]. It is common to find conflicting results between efficacy and effectiveness trials of the same intervention, or differences in outcomes in patients involved in highly controlled clinical trials and those receiving care in usual clinical practice [48, 84, 87, 91, 92] – a phenomenon often labeled the "efficacy-effectiveness gap" [84]. On the other hand, a recent empirical evaluation comparing the outcomes of highly active anti-retroviral therapy patients enrolled in a pragmatic randomized trial to those outside the trial setting, found only small differences in outcomes, suggesting that pragmatic trials *can* provide realistic estimates of a treatment's effect in usual practice [93]. It has become widely recognized that policy-makers and clinicians should base treatment decisions on evidence from effectiveness trials [19, 52-54, 58, 90] and that there is a pressing need for *more* effectiveness or pragmatic trials to inform decision-makers [38], provide reliable estimates of treatment effects in subgroups [94], and address the broader needs of policy-

makers [17]. Current trends in research are focused on acquiring generalizable scientific evidence, leading to applied studies with a greater emphasis on external validity and less protection of internal validity [79]. Furthermore, a shift towards favouring effectiveness studies in evidence syntheses (guideline development and systematic reviews) will make them more relevant to actual practice and ultimately improve the clinical outcomes they address [3, 34]. Lastly, experts in pharmacoeconomics are recognizing that biased results are possible when evaluations are conducted using cost and outcome data from efficacy studies, and that in order to obtain realistic estimates of the cost-effectiveness of drugs, data from effectiveness studies should be incorporated into economic evaluations [64, 81, 86].

An understanding of the distinction between efficacy and effectiveness approaches has important implications for those who evaluate and apply research evidence [34], and is necessary for judging the relevance of research results [79]. Kraemer stresses that where and how to pitch a given study along the efficacy-effectiveness spectrum should be appreciated, with trials on the effectiveness side warranting clinical application [72]. It has been suggested that efficacy and effectiveness research should require different levels of evidence [65] and different tools to rate their quality and strength [7]. Undoubtedly, the range of available evidence will consist of efficacy, effectiveness and hybrid efficacy-effectiveness studies; while some authors are uncertain of the clinical or social utility of these blended designs [79], others believe they may be the key to bridging the gap between efficacy and effectiveness research [71, 88, 95]. Furthermore, although hybrid designs may be more common than the two archetypal designs themselves, it is still thought that one can label a study as one type or the other, depending on the prevailing

design elements that are present [34, 43, 64]. Armitage makes a similar point when he states that "the two [research] attitudes are likely to co-exist, and compete for ascendancy, in any one trial,..." [96]. Much of this commentary suggests that it may be beneficial and possible to differentiate between efficacy and effectiveness trials or quantify where a study exists on the efficacy-effectiveness spectrum, in hopes of aiding in the evaluation of the generalizability of research findings, the establishment of levels of evidence in evidence-based decision-making, and the determination of intervention cost-effectiveness.

2.6 A new rating instrument

Gartlehner and colleagues have developed the first and only tool designed to distinguish between efficacy and effectiveness studies in the literature [50]. The intended applications of the tool, as suggested by the authors, are in the establishment of inclusion criteria for and interpretation of systematic reviews, and in the evaluation of the generalizability of individual studies. The rating instrument comprises seven criteria of study design, each chosen on the basis that they influence a study's external validity. In light of the lack of validated definitions for efficacy and effectiveness trials, the authors invited the directors of 12 evidence-based practice centers in the United States to identify examples of each study type in the literature. The authors then applied the proposed criteria to the 24 nominated pharmaceutical trials, and determined the sensitivity of the instrument in correctly identifying the example effectiveness trials that had been provided.

The first of Gartlehner et al.'s seven criteria is 'Populations in primary care'. This item was included on the basis that study settings in efficacy trials tend to be

unrepresentative of those where patients with the condition of interest typically would receive care. Furthermore, the setting has implications for the available study sample, in that patients with access to such specialized facilities and providers are likely to differ in many ways from the general patient population.

The second item, 'Less stringent eligibility criteria', highlights the notion that effectiveness trials should study a treatment's effect when applied to a heterogeneous patient sample, as opposed to a group that is highly selected and unrepresentative of those to which the treatment would be applied in usual clinical practice.

Third is 'Health outcomes' – Gartlehner et al. differentiate between subjective, objective, and health-related outcomes, stating that the latter should be of principal interest in effectiveness studies, whereas the former two are more commonplace in efficacy trials.

The fourth item, 'Long study duration, clinically relevant treatment modalities', emphasizes the need for effectiveness studies to implement protocols that mimic clinical practice, including diagnostic standards, treatment modalities, and treatment duration.

'Assessment of adverse events' is the fifth proposed criterion. It is stated that using "objective scales with predefined symptoms" for adverse events determination would be ideal in the effectiveness trial, but that the assessment may be limited to known critical issues.

Item 6 of the tool is 'Adequate sample size to assess a minimally important difference from a patient perspective'. The authors maintain that efficacy studies do not have adequate power to detect small, but clinically meaningful differences between treatments. It is suggested that an effectiveness study should be adequately powered to detect a

minimally important difference in quality of life, and larger samples are required when rare outcomes are of primary interest.

The final proposed criterion is 'Intention-to-treat (ITT) analysis'. The authors make note that efficacy trials tend to exclude protocol deviators from analyses whereas data from effectiveness trials must be assessed according to ITT, in order to account for the various reasons why patients discontinue or alter their course of treatment.

Gartlehner and colleagues concluded that these criteria can distinguish between efficacy and effectiveness studies, with a cut-off of six out of seven criteria providing the optimal balance between sensitivity and specificity in terms of correctly identifying effectiveness trials [50]. The authors promote the use of their tool and are confident of their results; however, they do note that their analyses were based on a rather small sample of studies [50]. They further claim that the same criteria are likely to be applicable to studies of other types of interventions beyond pharmaceuticals.

2.7 Conclusion

In conclusion, it is apparent that decision-makers in healthcare including, but not limited to, practitioners, policy-makers, and health economists, have concerns regarding the generalizability of high-quality scientific evidence and how they might best incorporate research findings in the decision-making process. Perhaps the distinction between efficacy and effectiveness studies, or the notion of pitching studies along a continuum from efficacy to effectiveness, would be of value to reviewers and their target audience, if incorporated into critical appraisal methodology and levels of evidence systems. The focus of the next chapter of this thesis is to present a manuscript titled "A scale to identify effectiveness studies appears to be valid, but reliance on individual

ratings is problematic due to sub-optimal reliability". This manuscript describes a study that applied Gartlehner and colleagues' tool to a sample of RCTs of pharmacological and non-pharmacological interventions in stroke rehabilitation.

2.8 References

1 Sackett DL: Evidence-based medicine. Spine 1998;23:1085-1086.

2 Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, Guyatt GH, Harbour RT, Haugh MC, Henry D, Hill S, Jaeschke R, Leng G, Liberati A, Magrini N, Mason J, Middleton P, Mrukowicz J, O'Connell D, Oxman AD, Phillips B, Schunemann HJ, Edejer TT, Varonen H, Vist GE, Williams JW, Jr., Zaza S: Grading quality of evidence and strength of recommendations. BMJ 2004;328:1490.

3 Coulter ID: Evidence summaries and synthesis: necessary but insufficient approach for determining clinical practice of integrated medicine? Integr Cancer Ther 2006;5:282-286.

4 Barton S: Which clinical studies provide the best evidence? The best RCT still trumps the best observational study. BMJ 2000;321:255-256.

5 Olivo SA, Macedo LG, Gadotti IC, Fuentes J, Stanton T, Magee DJ: Scales to assess the quality of randomized controlled trials: a systematic review. Phys Ther 2008;88:156-175.

6 Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S: Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. Control Clin Trials 1995;16:62-73.

7 Lohr KN: Rating the strength of scientific evidence: relevance for quality improvement programs. Int J Qual Health Care 2004;16:9-18.

8 Ferreira PH, Ferreira ML, Maher CG, Refshauge K, Herbert RD, Latimer J: Effect of applying different "levels of evidence" criteria on conclusions of Cochrane reviews of interventions for low back pain. J Clin Epidemiol 2002;55:1126-1129.

9 Clarkson JE: Getting research into clinical practice - barriers and solutions. Caries Res 2004;38:321-324.

10 Wille-Jorgensen P, Renehan AG: Systematic reviews and meta-analyses in coloproctology: interpretation and potential pitfalls. Colorectal Dis 2008;10:21-32.

11 Pirmohamed M: Best evidence and the clinical decision making process. Postgrad Med J 2002;78:316.

12 Bero LA, Jadad AR: How Consumers and Policymakers Can Use Systematic Reviews for Decision Making. Ann Intern Med 1997;127:37-42.

13 Cook DJ, Mulrow CD, Haynes RB: Systematic Reviews: Synthesis of Best Evidence for Clinical Decisions. Ann Intern Med 1997;126:376-380.

14 Dobbins M, Jack S, Thomas H, Kothari A: Public health decision-makers' informational needs and preferences for receiving research evidence. Worldviews Evid Based Nurs 2007;4:156-163.

15 Mulrow CD: Systematic Reviews: Rationale for systematic reviews. BMJ 1994;309:597-599.

16 Logan S: Systematic reviews and making decisions. Child Care Health Dev 1998;24:255-257.

17 Taylor RS, Niv D, Raj PP: Exploration of the evidence. Pain Pract 2006;6:10-21.

18 Atkins D, Fink K, Slutsky J: Better information for better health care: the Evidence-based Practice Center program and the Agency for Healthcare Research and Quality. Ann Intern Med 2005;142:1035-1041.

19 Tunis SR, Stryer DB, Clancy CM: Practical Clinical Trials: Increasing the Value of Clinical Research for Decision Making in Clinical and Health Policy. JAMA 2003;290:1624-1632.

20 Laupacis A, Straus S: Systematic reviews: time to address clinical and policy relevance as well as methodological rigor. Ann Intern Med 2007;147:273-274.

21 Bigby M, Williams H: Appraising systematic reviews and meta-analyses. Arch Dermatol 2003;139:795-798.

22 Scott IA, Greenberg PB: Cautionary tales in the clinical interpretation of therapeutic trial reports. Intern Med J 2005;35:611-621.

23 Rothwell PM: External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet 2005;365:82-93.

24 Farquhar C, Vail A: Pitfalls in systematic reviews. Curr Opin Obstet Gynecol 2006;18:433-439.

25 Green LW: From research to "best practices" in other settings and populations. Am J Health Behav 2001;25:165-178.

26 Travers J, Marsh S, Caldwell B, Williams M, Aldington S, Weatherall M, Shirtcliffe P, Beasley R: External validity of randomized controlled trials in COPD. Respir Med 2007;101:1313-1320.

27 Uijen AA, Bakx JC, Mokkink HG, van Weel C: Hypertension patients participating in trials differ in many aspects from patients treated in general practices. J Clin Epidemiol 2007;60:330-335.

28 Kandzari DE, Roe MT, Chen AY, Lytle BL, Pollack CV, Jr., Harrington RA, Ohman EM, Gibler WB, Peterson ED: Influence of clinical trial enrollment on the

quality of care and outcomes for patients with non-ST-segment elevation acute coronary syndromes. Am Heart J 2005;149:474-481.

29 Van Spall HG, Toren A, Kiss A, Fowler RA: Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review. JAMA 2007;297:1233-1240.

30 Ferguson L: External validity, generalizability, and knowledge utilization. J Nurs Scholarsh 2004;36:16-22.

31 Mercer SL, DeVinney BJ, Fine LJ, Green LW, Dougherty D: Study designs for effectiveness and translation research :identifying trade-offs. Am J Prev Med 2007;33:139-154.

32 Gruen RL, Morris PS, McDonald EL, Bailie RS: Making systematic reviews more useful for policy-makers. Bull World Health Organ 2005;83:480.

33 Hopayian K: The need for caution in interpreting high quality systematic reviews. BMJ 2001;323:681-684.

34 Fritz JM, Cleland J: Effectiveness versus efficacy: more than a debate over language. J Orthop Sports Phys Ther 2003;33:163-165.

35 Mullen PD, Ramirez G: The promise and pitfalls of systematic reviews. Annu Rev Public Health 2006;27:81-102.

36 Green LW, Glasgow RE: Evaluating the relevance, generalization, and applicability of research: issues in external validation and translation methodology. Eval Health Prof 2006;29:126-153.

37 Persaud N, Mamdani MM: External validity: the neglected dimension in evidence ranking. J Eval Clin Pract 2006;12:450-453.

38 Glasgow RE, Magid DJ, Beck A, Ritzwoller D, Estabrooks PA: Practical clinical trials for translating research to practice: design and measurement recommendations. Med Care 2005;43:551-557.

39 Atkins D, Eccles M, Flottorp S, Guyatt GH, Henry D, Hill S, Liberati A, O'Connell D, Oxman AD, Phillips B, Schunemann H, Edejer TT, Vist GE, Williams JW, Jr.: Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches The GRADE Working Group. BMC Health Serv Res 2004;4:38.

40 Juni P, Witschi A, Bloch R, Egger M: The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 1999;282:1054-1060.

41 Harbour R, Miller J: A new system for grading recommendations in evidence based guidelines. BMJ 2001;323:334-336.

42 Walach H, Falkenberg T, Fonnebo V, Lewith G, Jonas WB: Circular instead of hierarchical: methodological principles for the evaluation of complex interventions. BMC Med Res Methodol 2006;6:29.

43 Schwartz D, Lellouch J: Explanatory and pragmatic attitudes in therapeutical trials. J Chronic Dis 1967;20:637-648.

44 Cochrane AL: Effectiveness and efficiency: random reflections on health services. London, Nuffield Provincial Hospital Trust, 1972.

45 Cochrane AL, Maynard A, Chalmers I: Non-random reflections on health services researchon the 25th anniversary of Archie Cochrane's Effectiveness and efficiency. London, BMJ Pub. Group, 1997.

46 Haynes B: Can it work? Does it work? Is it worth it? The testing of healthcare interventions is evolving. BMJ 1999;319:652-653.

47 Armitage P: Attitudes in clinical trials. Stat Med 1998;17:2675-2683.

48 Collins J: Which Randomized Controlled Trials Are Relevant to Clinical Practice? Obstet Gynecol 2005;106:216-218.

49 Fuhrer MJ: Overview of clinical trials in medical rehabilitation: impetuses, challenges, and needed future directions. Am J Phys Med Rehabil 2003;82(10 Suppl):S8-S15.

50 Gartlehner G, Hansen RA, Nissman D, Lohr KN, Carey TS: A simple and valid tool distinguished efficacy from effectiveness studies. J Clin Epidemiol 2006;59:1040-1048.

51 Godwin M, Ruhland L, Casson I, MacDonald S, Delva D, Birtwhistle R, Lam M, Seguin R: Pragmatic controlled clinical trials in primary care: the struggle between external and internal validity. BMC Med Res Methodol 2003;3.

52 Helms PJ: 'Real world' pragmatic clinical trials: what are they and what do they tell us? Pediatr Allergy Immunol 2002;13:4-9.

53 Macpherson H: Pragmatic clinical trials. Complement Ther Med 2004;12:136-140.

54 Roland M, Torgerson DJ: Understanding controlled trials: What are pragmatic trials? BMJ 1998;316:285.

55 Sekine I, Takada M, Nokihara H, Yamamoto S, Tamura T: Knowledge of Efficacy of Treatments in Lung Cancer Is Not Enough, Their Clinical Effectiveness Should Also Be Known. J Thorac Oncol 2006;1:398-402.

56 Streiner DL: The 2 "Es" of research: efficacy and effectiveness trials. Can J Psychiatry 2002;47:552-556.

57 Tansella M, Thornicroft G, Barbui C, Cipriani A, Saraceno B: Seven criteria for improving effectiveness trials in psychiatry. Psychol Med 2006;36:711-720.

58 Zwarenstein M, Oxman A: Why are so few randomized trials useful, and what can we do about it? J Clin Epidemiol 2006;59:1125-1126.

59 Sheikh A, Smeeth L, Ashcroft R: Randomised controlled trials in primary care: scope and application. Br J Gen Pract 2002;52:746-751.

60 Feinstein AR: An additional basic science for clinical medicine: II. The limitations of randomized trials. Ann Intern Med 1983;99:544-550.

61 Buyse M: Regulatory versus public health requirements in clinical trials. Drug Inf J 1993;27:977-984.

62 Sackett DL: Why randomized controlled trials fail but needn't: 2. Failure to employ physiological statistics, or the only formula a clinician-trialist is ever likely to need (or understand!). CMAJ 2001;165:1226-1237.

63 Depp C, Lebowitz BD: Clinical trials: bridging the gap between efficacy and effectiveness. Int Rev Psychiatry 2007;19:531-539.

64 Bombardier C, Maetzel A: Pharmacoeconomic evaluation of new treatments: efficacy versus effectiveness studies? Ann Rheum Dis 1999;58(Suppl I):I82-I85.

65 Leichsenring F: Randomized controlled versus naturalistic studies: a new research agenda. Bull Menninger Clin 2004;68:137-151.

66 Newcombe RG: Explanatory and pragmatic estimates of the treatment effect when deviations from allocated treatment occur. Stat Med 1988;7:1179-1186.

67 MacRae KD: Pragmatic versus explanatory trials. Int J Technol Assess Health Care 1989;5:333-339.

68 Charlton BG: Understanding randomized controlled trials: explanatory or pragmatic? Fam Pract 1994;11:243-244.

69 Alford L: On differences between explanatory and pragmatic clinical trials. New Zealand Journal of Physiotherapy 2007;35:12-16.

70 Bausewein C, Higginson IJ: Appropriate methods to assess the effectiveness and efficacy of treatments or interventions to control cancer pain. J Palliat Med 2004;7:423-430.

71 Roy-Byrne PP, Sherbourne CD, Craske MG, Stein MB, Katon W, Sullivan G, Means-Christensen A, Bystritsky A: Moving treatment research from clinical trials to the real world. Psychiatr Serv 2003;54:327-332.

72 Kraemer HC: "Rules" of evidence in assessing the efficacy and effectiveness of treatments. Dev Neuropsychol 2003;24:705-718.

73 Nash J, McCrory D, Nicholson R, Andrasik F: Efficacy and Effectiveness Approaches in Behavioral Treatment Trials. Headache 2005;45:507-512.

74 Hoagwood K, Hibbs E, Brent D, Jensen P: Introduction to the special section: efficacy and effectiveness in studies of child and adolescent psychotherapy. J Consult Clin Psychol 1995;63:683-687.

75 Bond J, Atkinson A, Gregson BA, Newell DJ: Pragmatic and explanatory trials in the evaluation of the experimental National Health Service nursing homes. Age Ageing 1989;18:89-95.

76 Gallin JI: Principles and practice of clinical research. San Diego, Calif, 2002.

77 Sackett DL, Gent M: Controversy in counting and attributing events in clinical trials. N Engl J Med 1979;301:1410-1412.

78 Ernst E, Canter PH: Limitations of "pragmatic" trials. Postgrad Med J 2005;81:203.

79 Wells KB: Treatment research at the crossroads: the scientific interface of clinical trials and effectiveness research. Am J Psychiatry 1999;156:5-10.

80 Arce JC, Nyboe Andersen A, Collins J: Resolving methodological and clinical issues in the design of efficacy trials in assisted reproductive technologies: a mini-review. Hum Reprod 2005;20:1757-1771.

81 Revicki DA, Frank L: Pharmacoeconomic evaluation in the real world. Effectiveness versus efficacy studies. Pharmacoeconomics 1999;15:423-434.

82 McMahon AD: Study control, violators, inclusion criteria and defining explanatory and pragmatic trials. Stat Med 2002;21:1365-1376.

83 Glasgow RE, Lichtenstein E, Marcus AC: Why don't we see more translation of health promotion research to practice? Rethinking the efficacy-to-effectiveness transition. Am J Public Health 2003;93:1261-1267.

84 Bauer MS, Williford WO, Dawson EE, Akiskal HS, Altshuler L, Fye C, Gelenberg A, Glick H, Kinosian B, Sajatovic M: Principles of effectiveness trials and their implementation in VA Cooperative Study #430: 'Reducing the efficacy-effectiveness gap in bipolar disorder'. J Affect Disord 2001;67:61-78.

85 Hotopf M, Churchill R, Lewis G: Pragmatic randomised controlled trials in psychiatry. Br J Psychiatry 1999;175:217-223.

86 Simon G, Wagner E, Vonkorff M: Cost-effectiveness comparisons using "real world" randomized trials: the case of new antidepressant drugs. J Clin Epidemiol 1995;48:363-373.

87 Lagomasino IT, Dwight-Johnson M, Simpson GM: Psychopharmacology: the need for effectiveness trials to inform evidence-based psychiatric practice. Psychiatr Serv 2005;56:649-651.

88 Clarke GN: Improving the Transition From Basic Efficacy Research to Effectiveness Studies: Methodological Issues and Procedures. J Consult Clin Psychol 1995;63:718-725.

89 Conine TA, Hershler C: Effectiveness: a neglected dimension in the assessment of rehabilitation devices and equipment. Int J Rehabil Res 1991;14:117-122.

90 Treweek S, McCormack K, Abalos E, Campbell M, Ramsay C, Zwarenstein M: The Trial Protocol Tool: The PRACTIHC software tool that supported the writing of protocols for pragmatic randomized controlled trials. J Clin Epidemiol 2006;59:1127-1133.

91 Davidson MH: Differences between clinical trial efficacy and real-world effectiveness. Am J Manag Care 2006;12:405-411.

92 Weisz JR, Weiss B, Donenberg GR: The lab versus the clinic. Effects of child and adolescent psychotherapy. Am Psychol 1992;47:1578-1585.

93 Eg Hansen AB, Gerstoft J, Kirk O, Mathiesen L, Pedersen C, Nielsen H, Jensen-Fangel S, Sorensen HT, Obel N: Unmeasured confounding caused slightly better response to HAART within than outside a randomized controlled trial. J Clin Epidemiol 2008;61:87-94.

94 Rothwell PM: Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet 2005;365:176-186.

95 Carroll KM, Rounsaville BJ: Bridging the gap: a hybrid model to link efficacy and effectiveness research in substance abuse treatment. Psychiatr Serv 2003;54:333-339.

96 Armitage P: Attitudes in clinical trials. Stat Med 1998;17:2675-2683.

CHAPTER 3. MANUSCRIPT

A scale to identify effectiveness studies appears to be valid, but reliance on individual ratings is problematic due to sub-optimal reliability.

Laura L. Zettler BHSc, MSc (cand)[1,2], *Mark R. Speechley PhD*[1], *Katherine L. Salter BSc*[2], *Norine C. Foley MSc*[2], *and Robert W. Teasell MD*[2]

Department of Epidemiology and Biostatistics, Schulich School of Medicine and Dentistry, The University of Western Ontario, London, Ontario[1]

Department of Physical Medicine and Rehabilitation, Parkwood Hospital, St. Joseph's Health Care, London, Ontario[2]

Note: A version of this chapter will be submitted to the Journal of Clinical Epidemiology for consideration for publication.

3.1 Introduction

Randomized controlled trials (RCT) and systematic reviews of RCTs, with their place atop most hierarchies of evidence, are considered gold standards for determining the effects of a given intervention or treatment. Randomized studies differ amongst themselves, however; for some the objective is to estimate the effect of an intervention under highly controlled, optimal circumstances, while for others the goal is to examine how well an intervention works in usual clinical practice [1]. The distinction between studies with these different objectives was first made by Schwartz and Lellouch, who used the terms explanatory and pragmatic, respectively [2]. Explanatory trials are also referred to as efficacy, fastidious [3] or regulatory trials [4], and pragmatic trials are called effectiveness, management [5], practical [6, 7], or public health trials [4, 6]. Because the terms efficacy and effectiveness have gained the widest usage, they will be used in this paper.

A common criticism of efficacy trials is their limited external validity, in that the findings may neither be readily applicable to the intended target population, nor provide the necessary and relevant information required by decision-makers when faced with treatment decisions [7-11]. Efficacy trials have limited external validity because they typically enroll a highly selective, homogeneous patient sample; are conducted where practitioners and facilities are highly specialized; enforce strict treatment protocols to ensure patient and provider compliance; and have smaller samples that may lack the power to detect small, but worthwhile treatment effects on measures that are meaningful to patients. Although the two types of trials can be conceptualized, it is generally

accepted that most studies exist on a continuum [6, 9, 12-16], and that hybrid studies are possible [2, 17, 18].

While efficacy trials are useful for understanding the mechanisms through which interventions influence outcomes, there is a growing recognition that policy-makers and clinicians should additionally base treatment decisions on evidence from effectiveness trials, [7, 9, 13, 19-22]. Because decision-makers often rely on systematic reviews of available evidence for particular interventions, it would be useful if researchers could make the distinction between trial types in their systematic reviews. This is important to decision-makers because studies of the same intervention have produced conflicting results between the types of trials, or differences in outcomes in patients involved in clinical trials and those receiving care in usual clinical practice [11, 23-26]. The efficacy-effectiveness distinction is also important in studies of cost-effectiveness [27]. Specifically, there is growing acceptance that the demonstration of drug cost-effectiveness should most appropriately be conducted using data from effectiveness studies, as unrealistic and possibly biased results can occur when using cost and outcome data from efficacy studies [28-30].

Although the literature has outlined the typical characteristics of efficacy and effectiveness trials [4, 13, 14, 16, 18, 31-34], only recently was the first and only *tool* to differentiate between the study types published, by Gartlehner and colleagues [35]. The primary target audience for the instrument would likely be those conducting systematic reviews, and if the criteria can validly and reliably identify the different types of trials, it could prove to be quite useful for means of critical appraisal and evidence ranking within the systematic review process. The objectives of this study were: (1) to apply a published

instrument, that purports to differentiate between efficacy and effectiveness trials, to a sample of trials in stroke rehabilitation; (2) to assess the inter-rater reliability of the instrument, overall and within studies of pharmacological and non-pharmacological interventions; and 3) to attempt to validate the instrument by investigating associations between key study characteristics relevant to the effectiveness trial design, and total scale scores.

3.2 Methods

3.2.1 Study selection

The studies were selected from a systematic review of interventions in stroke rehabilitation [36]. Since its original publication in 2003, several updates have taken place, and the entire Evidence-Based Review of Stroke Rehabilitation (EBRSR), currently in its 10th edition, is freely available online [37]. The detailed review methodology of the EBRSR has been previously published [38]. Briefly, all studies were entered into a database and described according to a number of variables including year and journal of publication, study design, type of intervention, quality rating, as well as a number of variables dedicated to the coding of outcome measures. As of 2007, this database contained 768 studies, 634 specifically dedicated to stroke rehabilitation and the remaining 134 focusing on the secondary prevention of stroke.

A study was eligible for inclusion in the present study if it met the following criteria: 1) true randomized controlled trial (RCT) by design; 2) evaluated either a pharmacological or therapy-based (non-pharmacological) intervention to prevent or rehabilitate deficits following stroke; and 3) published after the release of the CONSORT Statement [39] (published during or after 1997). These inclusion criteria served to

assemble a relatively large, yet homogenous sample of studies in terms of study design, reporting quality, and the characteristics of the patient populations. Studies focused on the secondary prevention of stroke were not eligible because they were thought to represent a different genre of research consisting primarily of pharmacological trials. Furthermore, their inclusion would have added much heterogeneity into the sample, with regard to reporting quality as well as the patient populations and areas of intervention. Lastly, although Gartlehner and colleagues included only studies of medications in the development the tool, they did suggest that their criteria should be applicable to other types of interventions [35] – this prompted the decision to include trials of both pharmacological (P) *and* non-pharmacological (NP) interventions in the present study.

3.2.2 "Effectiveness tool"

The tool used in this study was that developed by Gartlehner et al [35]. The authors proposed seven criteria of study design that they believed would influence a trial's external validity, and thus would reliably distinguish between efficacy and effectiveness studies. The authors applied the tool to a sample of 24 pharmacological trials that had been identified *a priori,* as either efficacy or effectiveness trials, by the directors of 12 Evidence-based Practice Centers in North America. Each item was rated as either present or absent to arrive at a total score out of seven. Although these authors acknowledged that efficacy and effectiveness exist on a continuum, they believed that it may be necessary to classify studies as more on one side than the other. The authors concluded that using a cut-off of six criteria yielded the optimal balance between sensitivity and specificity in terms of correctly identifying effectiveness trials.

3.2.3 Rater calibration and modifications of criteria

The raters for this study (KS, NF, LZ) varied in terms of their experience with the stroke rehabilitation literature, but were all well matched in terms of knowledge of research methodology. Since the criteria included in the tool (see Table 3.1) lacked operational definitions, and were only textually described, some were difficult to apply even by the more experienced raters. For example, regarding Item 2, raters had a hard time deciding whether the eligibility criteria in a given study were "less stringent" and whether this item should be judged according to the number of criteria listed or by giving more weight to certain criteria than others. In an attempt to resolve some of these issues, all three raters independently applied the tool to ten trials from the EBRSR database that were not included in the main study sample. What followed was a discussion of the problems encountered when using the instrument and an attempt to tailor the individual items to make the tool more applicable to the stroke rehabilitation literature, while attempting to resolve some of the subjectivity in scoring. Table 3.1 describes the criterion operational definitions employed in the present study.

Table 3.1 Criterion operational definitions implemented due to application difficulties. Based on the application of the original tool to a sample of studies (n=10) from the EBRSR, not included in the main study sample.

Scale criterion*	If and why standardization was necessary	Operational definition employed
Item 1 ***Populations in primary care***	The definition states that the primary care setting criterion may be inadequate in some cases. Stroke rehab is rarely implemented in primary care; however, trial settings can range from typical facilities available to the majority of stroke survivors to those that are highly specialized, only accessible to a unique segment of the population.	To fulfill this criterion, the study setting should exemplify that where the average stroke patient could receive care (outpatient rehab department, homecare, general medical ward or rehab unit of non-academic hospital). The criterion will not be met if the trial setting is a specialized stroke care facility (stroke unit, academic hospital, rehab hospital, research laboratory).
Item 2 ***Less stringent eligibility criteria***	Some eligibility criteria limit the available study population more than others, making it difficult to rate this item based solely on the number of eligibility criteria in a given study. Having a thorough knowledge of the stroke rehab literature, the raters discussed the criteria that place the greatest limits on the	Study ineligibility based on cognitive impairment and/or prior stroke excludes a large portion of the general stroke population, resulting in this criterion not being met in these cases. Raters should refer to the study's participant flowchart, if available, to compare the number of

	eligible study population.	subjects initially screened to that included in the final sample, to help in the rating of this item.
Item 3 ***Health outcomes***	It was not definitively clear what would constitute a "health outcome" for the purposes of using this tool, therefore, raters incorporated the International Classification of Functioning Disability and Health (ICF) [40]. Since research in stroke rehabilitation primarily assesses outcomes related to function and disability, nearly all studies would assess health outcomes to some extent according to the ICF. In order to allow variation on this item, raters thought it would make sense to differentiate studies based on the level at which function/disability was measured.	To fulfill this criterion, at least one of the main study outcomes must fall under the *Activities and Participation* domain of the ICF. Outcomes that belong within the domain of *Body Functions and Structures*, will not be considered health outcomes for the purposes of using this tool, as these measure function and disability at the lowest level. Since outcomes such as clinical markers (eg. bone mineral density) or event rates (eg. mortality incidence) fall outside the ICF framework, they are also not considered health outcomes. The health outcome does not have to be the primary outcome, so long as it is a main study outcome, clearly emphasized in the results.
Item 4 ***Long study duration, clinically relevant treatment modalities***	It appeared that studies were often long enough in duration, but treatments were not clinically relevant and vice versa. Furthermore, although the authors provided examples of drug treatment modalities that would not be clinically relevant, it was harder to judge clinical relevance in the case of non-pharmacological interventions.	To fulfill this criterion, the study will have both features: long duration (an appropriate length, given the intervention and outcomes under study) and clinically relevant treatment. Regarding non-pharmacological interventions: flexible, individualized therapy, at an intensity and with resources typically seen in practice will be deemed clinically relevant. Following this, any new devices/equipment will likely not be clinically relevant, as they have yet to be integrated into practice.
Item 5 ***Assessment of adverse events***	This item was relatively clear; however, the definition states that compliance or discontinuation rates may reflect adverse events, bringing some confusion into the assessment of this item.	To fulfill this criterion, the study will either report any adverse events/side effects (or lack thereof) that occurred over the course of treatment. The criterion will not be met if a study assesses only compliance/discontinuation rates, without mention of any adverse events.
Item 6 ***Adequate sample size to assess a minimally important difference from a patient perspective***	With this item, raters were unsure whether they were to make their own judgment about the adequacy of sample size, or rely on the published report for this justification. Authors mention that the sample should be large enough to detect a minimally important difference in QOL; however, many studies are not powered to detect such a difference, if QOL is not the primary outcome.	To fulfill this criterion, the study will provide some justification of sample size by referring to prior pilot work or other background literature that corroborates the clinical meaningfulness of the difference in outcome on which the sample size is based.
Item 7 ***ITT analysis***	This criterion was straightforward.	To fulfill this criterion, the study will state that data were analyzed according to the intention-to-treat principle.

*Criteria adapted from Gartlehner et al. [35]

3.2.4 Data abstraction

2.4.1 Study scoring

The same three raters (KS, NF, LZ) were responsible for applying the tool to the final sample of studies. All were familiar with the item definitions and were provided with a series of reference notes that combined the original authors' definitions and any operational definitions that were discussed (see Appendix A). Raters applied the tool independently and each progressed through the entire sample of studies in a different randomized order. Raters used the same two category (yes/no) scoring system as intended by the original authors; however, raters also made note of whether inadequate reporting was the reason for a study not fulfilling the given criteria.

2.4.2 Validation hypotheses

Because there was no gold standard by which to assess the validity of this instrument could be assessed, data on other study characteristics, not included in the tool, but potentially relevant to the differences between efficacy and effectiveness trials, were abstracted.

Type of intervention: It was hypothesized that studies of non-pharmacological interventions would be more likely to receive a high score (closer to the 'effectiveness' end of the spectrum) than those of pharmacological interventions. This is based on the premise that the effectiveness trial may be better suited to the study of non-pharmacological alternatives, whereas the efficacy trial design may be more appropriate for evaluating pharmaceuticals [13, 19, 31, 41].

Type of comparison group: It was hypothesized that studies in which the comparison group received a different active intervention (either usual care or another mode of

treatment) would score highly compared to studies in which the comparison group received a placebo or sham intervention, the same intervention with one factor varied (ie. intensity, dosage, equipment, practitioner), or no treatment. The rationale here is that efficacy trials seek to explain if and why a given intervention works, whereas effectiveness trials seek to confirm the incremental benefit of one treatment modality over another [2, 3, 7, 13, 18, 42].

Number of participating centers: Another hypothesis was that multi-center trials would tend to have higher scores than single-center trials, because multi-center trials may capture a large, broader sample of patients across a variety of treatment settings, and are the common method for effectiveness RCTs [18, 43].

Sequence of RCT: It is often emphasized that in a given area of research there should be a sequencing from the demonstration of efficacy to that of effectiveness [1, 12, 14, 18, 32, 44], therefore, each study was recorded as being a pilot study, a full-scale RCT, or RCT follow-up, in order to test the hypothesis that full-scale studies and follow-ups would be more likely to receive high scores, whereas pilot studies would more often fall within the lower range of scores.

Quality rating scale score: Because all studies in the sample had been previously incorporated into the EBRSR, each had an associated quality rating according to the PEDro scale [45]. Because the difference between efficacy and effectiveness research is often framed in terms of validity (a trade-off favouring internal validity in the former and external validity in the latter) [14, 18, 31, 33, 41], it was hypothesized that studies with a low quality rating (an index of internal validity) would score highly on the effectiveness scale compared to studies with a high quality rating. Appendix B provides a detailed

description of the PEDro scale items as used in the EBRSR, and Appendix C outlines the validation variable definitions.

3.2.5 Statistical analyses

2.5.1 Descriptive analyses

In order to perform an overall descriptive analysis a single dataset was created using the results from individual raters. Each item was coded as present or absent according to how the majority of raters (2 of 3) scored that item, with the total score then tabulated for each study accordingly. This allowed for the calculation of the median total score, total score frequencies, as well as the percentage of studies that fulfilled each criterion for both pharmacological and non-pharmacological trials, without having to arrive at consensus for all studies in the sample. Descriptive statistics are in the form of frequencies and/or percentages, or medians and interquartile ranges, as appropriate.

2.5.2 Inter-rater reliability

Inter-rater reliability for each item and the dichotomized total score (<3 vs. ≥ 3) was assessed separately for trials of pharmacological and non-pharmacological interventions, using multi-rater kappa statistics and 95% confidence intervals (CI) [46]. These analyses were carried out using an Excel template [47] based on the calculations presented by Fleiss [46, 48]. The inter-rater reliability of the total score (not dichotomized) was further evaluated by assessing differences in score distributions across raters, using the non-parametric Friedman two-way analysis of variance [49]. Refer to Appendix D for a more detailed description of the reliability statistics used.

2.5.3 Validity assessment

Using the dataset created from the majority ratings, the total "effectiveness" score was dichotomized into "low" (<3) and "high" (≥3). The following variables were investigated for potential associations with the total score: type of intervention (pharmacological vs. non-pharmacological), type of comparison group (placebo/sham/no treatment or treatment varies by a factor vs. active treatment), number of participating centers (single center vs. multi-center), quality rating (high, PEDro score ≥ 7 vs. low, PEDro score < 7), and sequence of RCT (pilot vs. full-scale or follow-up RCT). The total score (high vs. low) was cross-tabulated with each independent variable and associations were analyzed using the Pearson chi-squared statistic with a continuity correction. For simplicity, in the case of the independent variable "type of comparison group", only trials with two treatment arms were included in the cross-tabulation; otherwise, all trials were included in the analysis. With the exception of multiple-rater kappa calculations, all data analyses were carried out using SPSS version 15.0 [50]. A two-sided alpha of less than 0.05 was considered statistically significant.

3.3 Results

3.3.1 Study selection

After applying the inclusion criteria, and removing any duplicates found by visually scanning the database, 414 eligible studies remained, of which 80 evaluated pharmacological and 334 non-pharmacological interventions. Based on the results of a power analysis (see Appendix E), it was decided to abstract data from all available pharmacological trials, as well as a simple random sample of the same size of non-pharmacological trials. Then the eligibility of all 160 trials was further verified, due to the

possibility of coding errors and further duplicates in the EBRSR database. Nine studies (two pharmacological and seven non-pharmacological) had to be excluded. Therefore, the final sample consisted of 151 RCTs, 78 of which evaluated a pharmacological and 73 a non-pharmacological intervention. See Figure 3.1 for a full description of the study selection process.

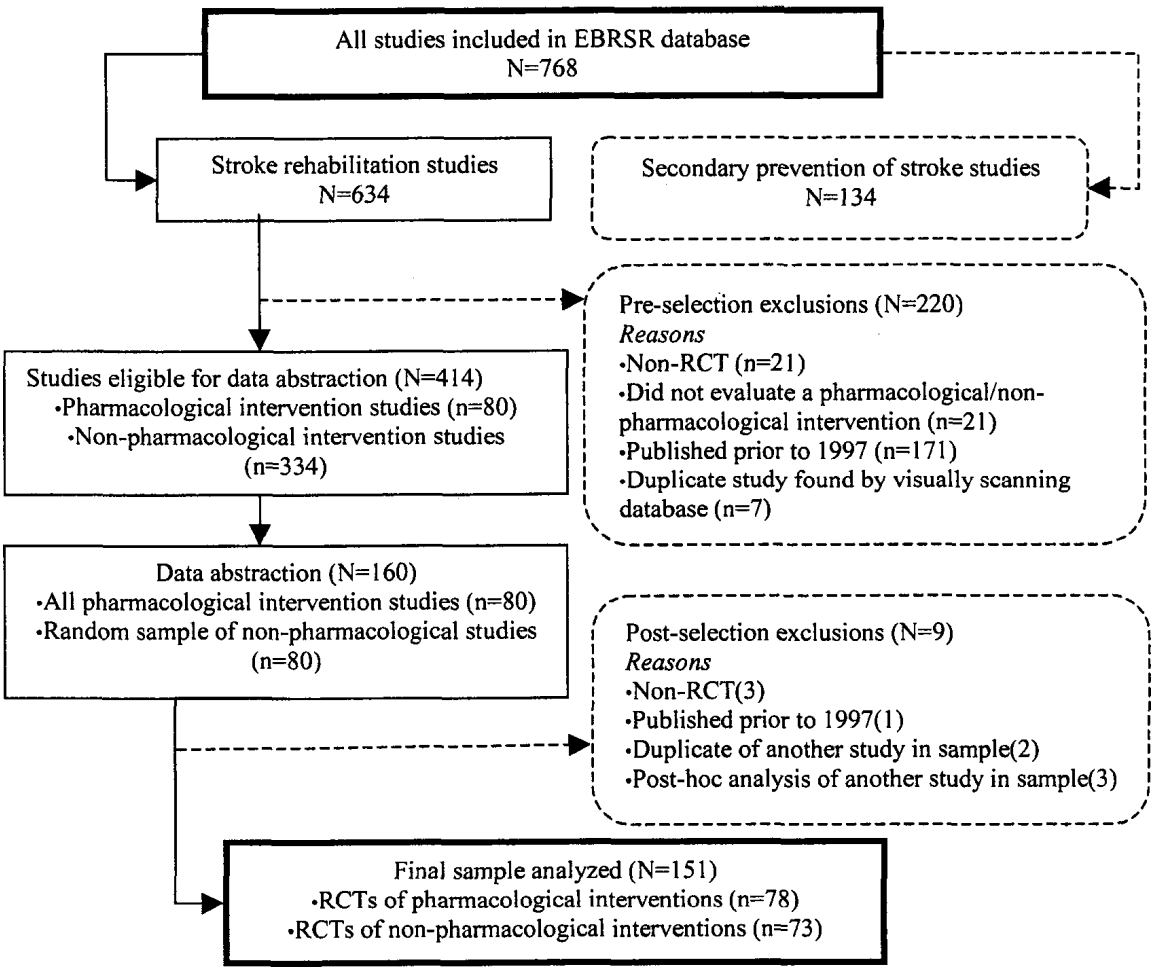

Fig. 3.1 Flow chart of study selection process.

3.3.2 Descriptive results

The RCTs included in the final sample investigated a range of general conditions within stroke rehabilitation (Table 3.2). The majority of pharmacological RCTs were focused on the amelioration of spasticity (n=17) or depression (n=13), whereas most non-pharmacological trials were aimed at upper and/or lower extremity function (n=46, inclusive). See Appendix F for the table of included studies.

Table 3.3 outlines the proportions of both pharmacological and non-pharmacological trials fulfilling each criterion according to the majority rating. For pharmacological trials, the item that was present most often (77% of trials) was Item 5 (assessment of adverse events). In non-pharmacological trials the most frequent criterion was Item 3 (health outcomes) which was seen in 67% of trials. For both types of trials, Item 6 (adequate sample size to assess a minimally important difference from a patient perspective) was present least often (12% and 14% of non-pharmacological and pharmacological trials, respectively). Most studies, regardless of intervention type, received low scores, with the most frequent score being two out of seven (27% of all trials); not one trial met all seven criteria (Figure 3.2). Median total scores were 3 and 2 for pharmacological and non-pharmacological trials, respectively (Figure 3.2). Inadequate information provided in the published report, as the reason for recording a "no" rating, was an issue for all raters on Item 1 (populations in primary care) (Table 3.4).

Table 3.2 General conditions under investigation in RCTs of pharmacological and non-pharmacological interventions included in the final sample (n=151).

RCTs of pharmacological interventions	No. of Studies	RCTs of non-pharmacological interventions	No. of studies
Spasticity	17	Upper extremity function	20
Depression	13	Lower extremity function	16
Aphasia	8	General physical function	10
Osteoporosis	7	Models of care delivery	10
Cognition/dementia	7	Education or reintegration	6
Acute thromboembolic complications	5	Spasticity	3
Paresis/hemiplegia	5	Urinary incontinence	2
Central pain	5	Dysphagia	2
Acute neurological/functional recovery	4	Depression	1
Hemiplegic/spastic shoulder pain	4	Perception	1
Memory	1	Memory	1
Nutritional status	1	Complex regional pain	1
Dysphagia	1		
Total	78	Total	73

Table 3.3 Proportion of trials of pharmacological (P) and non-pharmacological (NP) interventions fulfilling each criterion (in the case of rater disagreement, the majority rating was used).

Scale criterion	No. of studies (%)	
	P (n=78)	NP (n=73)
Item 1: Populations in primary care	12 (15.4)	23 (31.5)
Item 2: Less stringent eligibility criteria	32 (41.0)	33 (45.2)
Item 3: Health outcomes	27 (34.6)	49 (67.1)
Item 4: Long study duration; clinically relevant treatment modalities	37 (47.4)	36 (49.3)
Item 5: Assessment of adverse events	60 (76.9)	17 (23.3)
Item 6: Adequate sample size to assess a minimally important difference from a patient perspective	9 (11.5)	10 (13.7)
Item 7: Intention-to-treat analysis	30 (38.5)	16 (21.9)

Table 3.4 Proportion of times a "no" rating was due to inadequate information provided in the published report; by item and rater.

Scale criterion	Frequency (%)		
	Rater 1	Rater 2	Rater 3
Item 1: Populations in primary care	47 (44.3)	43 (35.5)	34 (29.6)
Item 2: Less stringent eligibility criteria	6 (6.7)	3 (4.1)	3 (3.7)
Item 3: Health outcomes	1 (1.5)	0 (0)	0 (0)
Item 4: Long study duration; clinically relevant treatment modalities	54 (48.6)	3 (6.3)	0 (0)
Item 5: Assessment of adverse events	2 (2.7)	8 (9.6)	0 (0)
Item 6: Adequate sample size to assess a minimally important difference from a patient perspective	25 (19.7)	8 (6.5)	0 (0)
Item 7: Intention-to-treat analysis	0 (0)	7 (6.4)	0 (0)

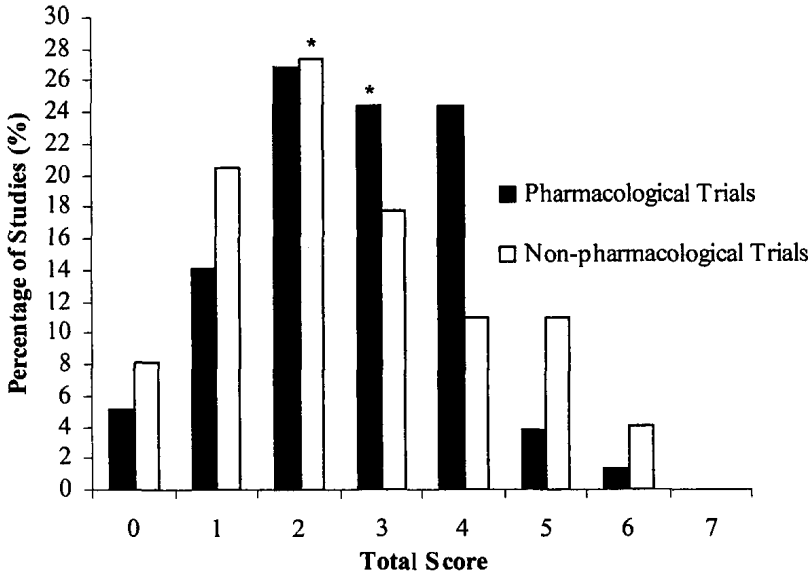

Fig. 3.2 Total score distributions for trials of pharmacological (P) and non-pharmacological (NP) interventions.
*Median total score [IQR] = 3 [2-4], P trials; 2 [1-4], NP trials

3.3.3 Instrument reliability

Table 3.5 presents the individual item reliability (multiple-rater kappa and 95% CI) for studies of each intervention type and for all trials combined. The lowest reliability, for trials of both pharmacological and non-pharmacological interventions, was for Item 4 (health outcomes), with an overall kappa value of 0.11 (95% CI: 0.02-0.20). Items 5 and 7 (assessment of adverse events and ITT analysis) had the highest reliability estimates, with overall kappa values of 0.81 (95% CI: 0.71-0.90) and 0.82 (95% CI: 0.73-0.91), respectively. Regarding inter-rater reliability of the total score (dichotomized), the multiple-rater kappa for pharmacological trials was 0.43 (95% CI: 0.31-0.56) and for non-pharmacological trials it was 0.51 (95% CI: 0.37-0.64). The results of the Friedman test showed evidence of distributional differences in scores among the three raters for pharmacological trials [2.5 (IQR:1-4) vs. 3 (IQR:2-4) vs. 2.5 (IQR:1-3); χ^2=8.567;

p=0.014]; for non-pharmacological trials, differences across raters were not statistically significant [2 (IQR:1-4) vs. 3 (IQR:1.5-4) vs. 3 (IQR:1.5-4); χ^2=2.966; p=0.227].

Table 3.5 Inter-rater reliability of individual scale items for trials of pharmacological (P) and non-pharmacological (NP) interventions and for all trials combined.

Scale criterion	Multiple-rater kappa (95%CI) for inter-rater reliability		
	P (n=78)	NP (n=73)	All trials (n=151)
Item 1: Populations in primary care	0.29 (0.16-0.42)	0.48 (0.35-0.62)	0.43 (0.34-0.52)
Item 2: Less stringent eligibility criteria	0.37 (0.24-0.50)	0.27 (0.14-0.40)	0.33 (0.23-0.42)
Item 3: Health outcomes	0.57 (0.45-0.70)	0.63 (0.50-0.76)	0.64 (0.55-0.73)
Item 4: Long study duration; clinically relevant treatment modalities	0.00 (-0.13-0.13)	0.21 (0.08-0.34)	0.11 (0.02-0.20)
Item 5: Assessment of adverse events	0.69 (0.57-0.82)	0.79 (0.66-0.92)	0.81 (0.71-0.90)
Item 6: Adequate sample size to assess a minimally important difference from a patient perspective	0.37 (0.24-0.50)	0.30 (0.17-0.44)	0.34 (0.25-0.43)
Item 7: Intention-to-treat analysis	0.85 (0.73-0.98)	0.76 (0.62-0.89)	0.82 (0.73-0.91)

3.3.4 Instrument validity

Table 3.6 shows the results of the tested validation hypotheses. The following variable levels were hypothesized to be more likely present in studies using the effectiveness trial design: non-pharmacological intervention, active comparator therapy, multi-center, full-scale or follow-up RCT, and low quality rating. A significantly greater proportion of studies scoring 3 or more (the "high" effectiveness score category) was found in studies with active compared to non-active treatment comparison groups (65% vs. 42%, p=0.041); multi-center compared to single-center trials (67% vs. 42%, p=0.007); and full-scale and follow-up RCTs compared to pilot studies (53% vs. 21%, p=0.018) (Table 3.6). Since these associations were all in the hypothesized directions, the validation hypotheses for "type of comparison group", "number of participating centers", and "sequence of RCT" were upheld. The difference in proportions of high score studies in the pharmacological and non-pharmacological intervention categories was not statistically significant, and so the validation hypothesis for "type of intervention" was unsupported. The difference in proportions of high score studies in the high and low quality categories was statistically significant (65% vs. 27%, p=0.000); however, the

association was not in the hypothesized direction, leaving the validation hypothesis for "quality rating" unsupported as well.

Table 3.6 Tests for association between key study characteristics and total "effectiveness" score (dichotomized into low, <3 and high, ≥3 categories).

Variable and associated levels	High score studies within each level (%)	χ^2 value
Type of intervention		
Pharmacological	53.8	1.14
Non-pharmacological	43.8	
Type of comparison group*		
Placebo/sham/no treatment	41.6	4.17†
Active treatment	65.4	
Number of participating centers		
Single center	41.7	7.18†
Multi-center	67.4	
Quality Rating		
High PEDro score	64.8	19.49†
Low PEDro score	27.0	
Sequence of RCT		
Pilot RCT	21.1	5.58†
Full scale or follow-up RCT	53.0	

*Analysis includes only two-arm trials (n=127).
† Statistically significant at $p<0.05$.

3.4 Discussion

The scale developed by Gartlehner and colleagues produces a range of scores from zero to 7, with the higher scores indicating studies meeting more of the criteria associated with effectiveness trials. The items lacked clear operational definitions making it difficult for raters to apply thus, there was an attempt to modify and standardize the items prior to the major application of the tool to the final sample of studies.

Generally, non-pharmacological trials met each criterion more often than their pharmacological counterparts with the exceptions of adverse events assessment and intention-to-treat analysis. The criteria met most frequently by each group of trials were not surprising: health outcomes (Item 3) and adverse events assessment (Item 5) for non-pharmacological and pharmacological trials, respectively. Health outcomes tend to be primary outcomes of interest in stroke rehabilitation, especially when the intervention is

non-pharmacological in nature. Thus, the revised definition of health outcome, appears to have allowed more variation on the item than if it had not been modified as such. Macpherson and colleagues give examples of outcomes common to each trial type: joint range of motion for the explanatory, and quality of life for the pragmatic trial [13]; this is analogous to the distinction made by the operational definition employed in the present study.

Adverse events assessment was fulfilled in the majority of pharmacological trials, which suggests the criterion definition may have been too inclusive, especially because it is recognized that clinical trials of pharmaceuticals tend to be closer to the efficacy side of the spectrum [1, 41]. Assessment of side effects only, may not be adequate for pharmaceutical studies, while in the case of non-pharmacological interventions, especially the types common in stroke rehabilitation, minor side effects may be the most adverse events that could occur. Thus, the scope of assessment required to fulfill this criteria may need to differ depending on the types and seriousness of the adverse events associated with the intervention.

The adequate sample size criterion (Item 6) was not met in the majority of trials overall. This is expected, as inadequate power and sample size are not uncommon in clinical trials, regardless of intervention type [51, 52]. The primary care setting criterion (Item 1) was rarely met in trials of pharmacological interventions. Perhaps drug trials are more likely than trials of non-pharmacological interventions to be conducted in academic institutions, where the infrastructure for such trials exist, due to partnerships between academic centers and the pharmaceutical industry.

A problem with the tool, and all instruments that involve abstraction of data from a published study, is their great dependence on reporting quality. Many studies did not describe the study setting, the intervention itself or other methodological features in much detail, resulting in studies often receiving "no" ratings on certain items. The only item that did not appear to be affected was Item 3 (Health outcomes); this does not come as a surprise, as even the most poorly reported study is still likely to describe or at least list the outcomes assessed. The item most drastically affected by level of reporting detail was Populations in primary care (Item 1), with each rater noting inadequate reporting on this item for at least 30% of the studies rated. The usual care setting is an important criterion in terms of effectiveness trial methodology [14, 25, 31, 34]; however, the setting descriptions provided in published reports may frequently be inadequate.

As expected, inter-rater reliability was problematic for Items 1, 2, 4 and 6. It was hoped that much of the subjectivity involved in the assessment of these items could have been resolved by the attempted standardization; however, the items still lacked clear operational definitions. Item 4 was the most problematic, with reliability values below that which would be expected by chance. A problem is that several questions are incorporated into the same criterion (is the study period long enough; are the treatment modalities clinically relevant; do the diagnostic standards follow usual clinical practice). Moreover, even though all raters were well-versed in the stroke rehabilitation literature, it was difficult to determine what exactly constitutes long enough, clinically-relevant treatment. Thus, the non-medical professional may have a hard time judging this item.

For Items 3, 5 and 7 reliability values were more acceptable; however, the health outcome criterion (Item 3) was not as reliable as expected, considering the strict

definition that was employed. Nevertheless, the reliability values obtained in this study are not unlike those estimated for scales of similar nature, such as those that rate studies for methodological quality [53-56].

The assessment of validity revealed significant associations between the total score and all but one of the variables investigated (intervention type). Furthermore, the associations were all in the hypothesized directions with the exception that high quality ratings were associated with high "effectiveness" scores. This is an interesting finding; although the scale items were chosen on the basis that they would improve the external validity of a study, perhaps they do not imply a necessary weakening of a study's internal validity or methodological quality. The literature often frames the difference between efficacy and effectiveness in terms of internal versus external validity; however, internal validity should still be of primary concern in any study. If adhering to RCT principles, effectiveness studies need not abandon internal validity for the sake of generalizability, but achieve an optimal balance of both [12, 33].

Treating the total score as dichotomous for the validity analyses may appear questionable, and likely resulted in a loss of statistical power; however, dichotomization was necessary due to the imbalance in group sizes among the variable levels. At the outset, it was hoped that the total score could be dichotomized (for both the reliability and validity analyses) using a cutoff of six criteria, as suggested by the authors of the tool; however, this would have resulted in extremely unbalanced groups in terms of sample size, with only four studies in the entire sample reaching this cutoff. Therefore, the total score was dichotomized around three criteria to achieve balance; however, it is not suggested that this number of criteria is significant in any way.

The validation should be interpreted cautiously, as it was only meant to be a crude analysis in an attempt to provide some support for the validity of the instrument beyond face validity. Without "gold standard" definitions for the constructs of efficacy and effectiveness, validity was challenging to assess. Perhaps an alternate or additional method of validation would have been to follow the process used in the development of the tool, and see how well the tool identified nominated efficacy and effectiveness trials in stroke rehabilitation. This method may have been problematic though, as it would have been challenging to find individuals who have "expertise" in labeling a trial as one type or the other and who were not in the original expert sample.

3.5 Conclusions

The tool by Gartlehner and colleagues was developed to easily identify effectiveness studies; however, this process may not be as simple as one would hope. If reviewers and decision-makers are to use this tool to make judgments regarding a study's design and inferences regarding external validity of a study's results, then items need to be further standardized in order to improve reliability, or at least multiple raters should arrive at consensus on a final rating. Furthermore, the factors that differentiate efficacy and effectiveness studies may vary depending on whether the intervention is pharmacological or non-pharmacological in nature, and whether or not the target population receives the intervention in a primary care setting. Although further refinements may be needed, the scale is an important advance as it brings to light the importance of judging the aim, relevance, and external validity of a given study within the decision-making process.

3.6 Acknowledgments

This work was funded by the London Life Studentship in Stroke Rehabilitation and the Canadian Institutes of Health Research, Frederick Banting and Charles Best Canada Graduate Scholarships Masters Award. The external funders provided student financial support, but played no role in the design, collection, analysis or interpretation of data; in the writing of the manuscript; or in the decision to submit the manuscript for publication.

3.7 References

1 Haynes B: Can it work? Does it work? Is it worth it? The testing of healthcare interventions is evolving. BMJ 1999;319:652-653.

2 Schwartz D, Lellouch J: Explanatory and pragmatic attitudes in therapeutical trials. J Chronic Dis 1967;20:637-648.

3 Feinstein AR: An additional basic science for clinical medicine: II. The limitations of randomized trials. Ann Intern Med 1983;99:544-550.

4 Buyse M: Regulatory versus public health requirements in clinical trials. Drug Inf J 1993;27:977-984.

5 Sackett DL, Gent M: Controversy in counting and attributing events in clinical trials. N Engl J Med 1979;301:1410-1412.

6 Depp C, Lebowitz BD: Clinical trials: bridging the gap between efficacy and effectiveness. Int Rev Psychiatry 2007;19:531-539.

7 Tunis SR, Stryer DB, Clancy CM: Practical Clinical Trials: Increasing the Value of Clinical Research for Decision Making in Clinical and Health Policy. JAMA 2003;290:1624-1632.

8 Rothwell PM: External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet 2005;365:82-93.

9 Zwarenstein M, Oxman A: Why are so few randomized trials useful, and what can we do about it? J Clin Epidemiol 2006;59:1125-1126.

10 Persaud N, Mamdani MM: External validity: the neglected dimension in evidence ranking. J Eval Clin Pract 2006;12:450-453.

11 Collins J: Which Randomized Controlled Trials Are Relevant to Clinical Practice? Obstet Gynecol 2005;106:216-218.

12 Kraemer HC: "Rules" of evidence in assessing the efficacy and effectiveness of treatments. Dev Neuropsychol 2003;24:705-718.

13 Macpherson H: Pragmatic clinical trials. Complement Ther Med 2004;12:136-140.

14 Nash J, McCrory D, Nicholson R, Andrasik F: Efficacy and Effectiveness Approaches in Behavioral Treatment Trials. Headache 2005;45:507-512.

15 Streiner DL: The 2 "Es" of research: efficacy and effectiveness trials. Can J Psychiatry 2002;47:552-556.

16 Tansella M, Thornicroft G, Barbui C, Cipriani A, Saraceno B: Seven criteria for improving effectiveness trials in psychiatry. Psychol Med 2006;36:711-720.

17 Armitage P: Attitudes in clinical trials. Stat Med 1998;17:2675-2683.

18 Fuhrer MJ: Overview of clinical trials in medical rehabilitation: impetuses, challenges, and needed future directions. Am J Phys Med Rehabil 2003;82(10 Suppl):S8-S15.

19 Helms PJ: 'Real world' pragmatic clinical trials: what are they and what do they tell us? Pediatr Allergy Immunol 2002;13:4-9.

20 Roland M, Torgerson DJ: Understanding controlled trials: What are pragmatic trials? BMJ 1998;316:285.

21 Rothwell PM: Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet 2005;365:176-186.

22 Treweek S, McCormack K, Abalos E, Campbell M, Ramsay C, Zwarenstein M: The Trial Protocol Tool: The PRACTIHC software tool that supported the writing of protocols for pragmatic randomized controlled trials. J Clin Epidemiol 2006;59:1127-1133.

23 Bauer MS, Williford WO, Dawson EE, Akiskal HS, Altshuler L, Fye C, Gelenberg A, Glick H, Kinosian B, Sajatovic M: Principles of effectiveness trials and their implementation in VA Cooperative Study #430: 'Reducing the efficacy-effectiveness gap in bipolar disorder'. J Affect Disord 2001;67:61-78.

24 Davidson MH: Differences between clinical trial efficacy and real-world effectiveness. Am J Manag Care 2006;12:405-411.

25 Lagomasino IT, Dwight-Johnson M, Simpson GM: Psychopharmacology: the need for effectiveness trials to inform evidence-based psychiatric practice. Psychiatr Serv 2005;56:649-651.

26 Weisz JR, Weiss B, Donenberg GR: The lab versus the clinic. Effects of child and adolescent psychotherapy. Am Psychol 1992;47:1578-1585.

27 Farahani P, Levine M, Goeree R: A comparison between integrating clinical practice setting and randomized controlled trial setting into economic evaluation models of therapeutics. J Eval Clin Pract 2006;12:463-470.

28 Revicki DA, Frank L: Pharmacoeconomic evaluation in the real world. Effectiveness versus efficacy studies. Pharmacoeconomics 1999;15:423-434.

29 Simon G, Wagner E, Vonkorff M: Cost-effectiveness comparisons using "real world" randomized trials: the case of new antidepressant drugs. J Clin Epidemiol 1995;48:363-373.

30 Bombardier C, Maetzel A: Pharmacoeconomic evaluation of new treatments: efficacy versus effectiveness studies? Ann Rheum Dis 1999;58(Suppl I):I82-I85.

31 Alford L: On differences between explanatory and pragmatic clinical trials. New Zealand Journal of Physiotherapy 2007;35:12-16.

32 Bausewein C, Higginson IJ: Appropriate methods to assess the effectiveness and efficacy of treatments or interventions to control cancer pain. J Palliat Med 2004;7:423-430.

33 Godwin M, Ruhland L, Casson I, MacDonald S, Delva D, Birtwhistle R, Lam M, Seguin R: Pragmatic controlled clinical trials in primary care: the struggle between external and internal validity. BMC Med Res Methodol 2003;3.

34 Sekine I, Takada M, Nokihara H, Yamamoto S, Tamura T: Knowledge of Efficacy of Treatments in Lung Cancer Is Not Enough, Their Clinical Effectiveness Should Also Be Known. J Thorac Oncol 2006;1:398-402.

35 Gartlehner G, Hansen RA, Nissman D, Lohr KN, Carey TS: A simple and valid tool distinguished efficacy from effectiveness studies. J Clin Epidemiol 2006;59:1040-1048.

36 Teasell RW, Foley NC, Bhogal SK, Speechley MR: An evidence-based review of stroke rehabilitation. Top Stroke Rehabil 2003;10:29-58.

37 Evidence-based review of stroke rehabilitation [homepage on the Internet]. c2007 [updated 2007 Sep 17; cited on 2008 Jan 25]. Available from: http://www.ebrsr.com/index_home.html.

38 Foley NC, Teasell RW, Bhogal SK, Speechley MR: Stroke Rehabilitation Evidence-Based Review: methodology. Top Stroke Rehabil 2003;10:1-7.

39 Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz KF, Simel D, Stroup DF: Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA 1996;276:637-639.

40 World Health Organization: International classification of functioning, disability and health ICF. Geneva, World Health Organization, 2001.

41 McMahon AD: Study control, violators, inclusion criteria and defining explanatory and pragmatic trials. Stat Med 2002;21:1365-1376.

42 Hotopf M, Churchill R, Lewis G: Pragmatic randomised controlled trials in psychiatry. Br J Psychiatry 1999;175:217-223.

43 Charlton BG: Understanding randomized controlled trials: explanatory or pragmatic? Fam Pract 1994;11:243-244.

44 Campbell M, Fitzpatrick R, Haines A, Kinmonth AL, Sandercock P, Spiegelhalter D, Tyrer P: Framework for design and evaluation of complex interventions to improve health. BMJ 2000;321:694-696.

45 The Physiotherapy Evidence Database [homepage on the Internet]. Sydney: Centre for Evidence-Based Physiotherapy; c2008 [updated 2008 Mar 3; cited 2008 Mar 15]. PEDro Scale; [about 6 screens]. Available from: http://www.pedro.fhs.usyd.edu.au/scale_item.html.

46 Fleiss JL: The Measurement of Interrater Agreement; in: Statistical methods for rates and proportions. New York, Wiley, 1981, pp 212-236.

47 King J E. Software solutions for obtaining a kappa-type statistic for use with multiple raters. Paper presented at the annual meeting of the Southwest Educational Research Association; 2004, February; Dallas, TX.

48 Fleiss JL, Nee JC, Landis J: Large sample variance of kappa in the case of different sets of raters. Psychol Bull 1979;86:974-977.

49 Friedman M: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. JASA 1937;32:675-701.

50 SPSS for Windows, Rel. 15.0.0. 2006. Chicago: SPSS Inc.

51 Keen HI, Pile K, Hill CL: The prevalence of underpowered randomized clinical trials in rheumatology. J Rheumatol 2005;32:2083-2088.

52 Moher D, Dulberg CS, Wells GA: Statistical power, sample size, and their reporting in randomized controlled trials. JAMA 1994;272:122-124.

53 Foley NC, Bhogal SK, Teasell RW, Bureau Y, Speechley MR: Estimates of quality and reliability with the physiotherapy evidence-based database scale to assess the methodology of randomized controlled trials of pharmacological and nonpharmacological interventions. Phys Ther 2006;86:817-824.

54 Maher CG, Sherrington C, Herbert RD, Moseley AM, Elkins M: Reliability of the PEDro scale for rating quality of randomized controlled trials. Phys Ther 2003;83:713-721.

55 Tooth L, Bennett S, McCluskey A, Hoffmann T, McKenna K, Lovarini M: Appraising the quality of randomized controlled trials: inter-rater reliability for the OTseeker evidence database. J Eval Clin Pract 2005;11:547-555.

56 Clark HD, Wells GA, Huet C, McAlister FA, Salmi LR, Fergusson D, Laupacis A: Assessing the quality of randomized trials: reliability of the Jadad scale. Control Clin Trials 1999;20:448-452.

CHAPTER 4. GENERAL DISCUSSION

4.1 Summary of key findings

One of the primary objectives of this thesis was to provide an overview of the concepts of efficacy and effectiveness approaches in research. Upon review of the literature, it was discovered that the approaches have been referred to using several different terms; that the main distinction between the approaches has been framed in various ways; and also that they have been described according to numerous archetypal methodological elements (Table 1.1). The main objective of efficacy or explanatory research, as cited by the majority in the literature, is to understand the true biological effect or mechanism of a treatment or to determine the beneficial effect of an intervention under ideal conditions. For effectiveness or pragmatic research, the objective cited by the majority is to allow a decision to be made regarding the best mode of treatment or to determine the beneficial effect of an intervention in ordinary clinical practice. A large portion of the literature also stated that efficacy trials are usually conducted in the earlier stages of a treatment's development, and effectiveness studies in the later stages, after treatment efficacy has been established.

The study design said to be used by both approaches and unanimously by the efficacy approach, is the RCT; however, it was also suggested by several authors that a non-randomized study design is often implemented with the effectiveness approach. Regarding study setting and participating clinicians, efficacy trials are noted to take place in specialized research settings, such as academic hospitals or commercial research sites, and recruit highly trained, skilled, and specialized practitioners. Effectiveness trials, on the other hand, more often take place in routine care settings with usual providers.

Other commonly cited characteristics of the approaches relate to the subject sample, and research or treatment protocol, as well as the choice of outcomes, comparative therapy, and statistical analyses. Specifically, efficacy trials were noted to implement strict inclusion criteria to achieve a homogeneous, and highly compliant, motivated and responsive sample; and follow fixed, strict and standardized protocols. They also were noted to choose objective, biomedical outcomes; use placebos as the comparative therapy; and conduct per-protocol statistical analyses. Conversely, effectiveness studies were said to implement wide inclusion criteria to achieve a heterogeneous sample with varying levels of compliance, motivation and responsiveness; and follow flexible, individualized protocols. They also were noted to choose broad health outcomes of meaning to patients and practitioners, as well as measures of harm and economic outcomes; use standard care or clinically relevant alternatives as the comparative therapy; and conduct intention-to-treat analyses. Finally, it was largely agreed that the main emphasis of efficacy trials is on maintaining internal validity, whereas effectiveness studies aim to maximize external validity.

Most of these commonly cited characteristics were incorporated into a recently published tool that claims to make the differentiation between the trial types [1]. The other primary objectives of this thesis, to evaluate the applicability, inter-rater reliability and validity of this tool, were addressed in Chapter 3. The fact that the literature corroborates the many criteria that are included in Gartlehner et al.'s scale is reassuring, as it supports the tool's face validity. Despite their relevance to the construct at hand, however, the criteria as published lacked operational definitions, making the scale difficult to apply. Thus, operational definitions were implemented in an attempt to

standardize the items prior to the major application of the tool, and subsequent evaluation of its inter-rater reliability and validity, within a sample of studies in stroke rehabilitation. Most of the tested validation hypotheses were supported, suggesting that the scale is not only face valid, but more importantly has some degree of construct validity as well. Even though operational definitions were employed, few items on the scale produced impressive inter-rater reliability estimates, suggesting that many of the criteria are inherently subjective, most notably Items 1 (Populations in primary care), 2 (Less stringent eligibility criteria), 4 (Long study duration, clinically relevant treatment modalities) and 6 (Adequate sample size to assess a minimally important difference from a patient perspective). It was not surprising to find the remaining three criteria, Items 3 (Health outcomes), 5 (Assessment of adverse events) and 7 (Intention to treat analysis), to have superior reliability estimates, as they can be more objectively defined.

There were three secondary objectives of this thesis. First, to address any problems encountered with the application of the proposed criteria; second, to discuss how the differentiation between efficacy and effectiveness may depend on both the condition under investigation and the nature of the intervention; and finally to propose how criteria that differentiate between efficacy and effectiveness research can be incorporated into the evidence-based decision-making process. Some of these objectives were points of discussion in Chapter 3, but will be delineated further in the following sections.

4.2 Identified problems with the "effectiveness" scale

A problem with the tool, and all instruments that involve abstraction of data from published studies, is its great dependence on reporting quality. *Populations in primary care* was the criterion most affected by level of reporting detail. All raters gave at least 30% of studies a "no" rating for this item, due to the fact that information regarding study setting was often inadequate. Regarding Item 4 (Long study duration, clinically relevant treatment modalities), Rater 1 judged it as absent because of inadequate information in nearly half of the studies evaluated; however, the other two raters did not follow this trend. Perhaps this rater was the most meticulous of the three for this item, possibly being unsure of the clinical relevance or appropriate length of treatment and requiring further information to justify a "yes" response. Rater 3 appeared to be the least discriminating rater in that she was usually able to arrive at a "yes" or "no" response, with the exception of Item 1, for which inadequate reporting prompted a "no" rating in 30% of studies. Rater 2 appeared to fall somewhere between Raters 1 and 3 in this regard – not as stringent as Rater 1, but moreso than Rater 3. Obviously, these differences in agreement among experienced raters on whether there was adequate information provided in the reports affected the estimates of inter-rater reliability; however, this would not be as problematic if operational definitions could be improved by outlining specific requirements for satisfying each criterion. Perhaps it would be wise to use a Yes-No- Unsure rating system, and also, to apply a scale of methodological quality or internal validity alongside such types of criteria, as mentioned by Gartlehner and colleagues.

Regarding individual items, Item 4 (Long study duration, clinically relevant treatment modalities), was the most difficult to judge as evidenced by the very low reliability

estimates and as expressed by all raters. The likely problem with this item is that several questions are incorporated into the same criterion (is the study period long enough; are the treatment modalities clinically relevant; do the diagnostic standards follow usual clinical practice). Perhaps these elements could be separate criteria, in order to avoid the conflict that arises when not all are simultaneously present or absent. The operational definition employed for this item specified that both main elements (length and clinical relevance) be satisfied in order for this criterion to be met; however, it still remained a difficult judgment. Moreover, even though all raters were well-versed in the stroke rehabilitation literature, it was often difficult for them to determine what exactly constituted long enough, clinically-relevant treatment, suggesting that non-medical professionals may struggle when rating this item.

Another problem with the author's presentation of the scale is the lack of independence among the criteria. For instance, although Item 3 already addresses the inclusion of health outcomes, Item 4 then states that the study duration must be long enough to assess health outcomes, and Item 6 says that the sample size should be large enough to assess a minimally important difference in QOL. Thus, nearly half of all scale items make reference to health outcomes. This was another issue that had to be addressed by implementing operational definitions. It was indicated that length of study and sample size, for Items 4 and 6, respectively, were to be judged in relation to the outcomes of interest for the given study as opposed to 'health outcomes'. If the construct of efficacy-effectiveness does lie on a continuum, and if hybrid studies are possible, then each item should be independent and have the potential to be present or absent for any given study.

A further problem relates to how well the scale appears to capture the efficacy-effectiveness construct. This is relevant to one item in particular. Although Item 7 was among the easiest items to assess, had superior inter-rater reliability, and is a commonly noted characteristic of effectiveness studies (Table 1.1), raters were concerned that it may not be a factor that differentiates efficacy and effectiveness trials. The intention-to-treat (ITT) principle has been the analysis standard for any RCT since the release of the CONSORT Statement in 1996 [2]. Thus, whether or not a study used ITT analysis may be more related to methodological quality, than a feature whose presence or absence indicates a study's position on the efficacy-effectiveness continuum. Other authors agree with the notion that ITT may no longer be a factor that distinguishes effectiveness from efficacy trials [3, 4].

The issue of compliance is mentioned within the authors' descriptions of the *Long study duration, clinically relevant treatment modalities* and *Assessment of adverse events* criteria (Items 4 and 5). The tool's authors note that compliance should be defined as an outcome measure in effectiveness trials and that discontinuation rates and compliance may reflect adverse events if measured as outcomes. While it is valid that the authors tried to integrate this factor into the scale, perhaps compliance with the treatment protocol should be an item on its own, so its contribution to the scale and the efficacy-effectiveness construct is more clearly emphasized. Compliance is a crucial factor in determining an intervention's success, and the literature recognizes that the assessment of compliance or adherence is a key element of effectiveness trial methodology [5-8].

Finally, the issue of the active or inert nature of the comparative therapy is also not fully addressed in Gartlehner et al.'s criteria, and is clearly an important part of the

efficacy-effectiveness construct, as highlighted in the literature. Although Item 4 emphasizes that the treatment modalities in effectiveness studies should be clinically relevant, it does not explicitly reference the treatment and comparison groups, and the raters in the present study failed to consider this issue as well. It would not be difficult to find cases in which the experimental treatment is deemed clinically relevant, while the comparator is not (ie. placebo or sham treatment), or conversely, cases where the clinical relevance of the experimental therapy is uncertain, as when the comparator therapy is a widely used 'standard care'. Thus, further elaborating on the notion of 'clinical relevance' to address the study groups separately, may prove to be a better way to operationalize this item. Moreover, it was also emphasized in the literature that effectiveness studies often allow the use of concurrent therapies, whereas efficacy trials either prohibit or limit therapy use beyond that of the experimental treatment. Perhaps this idea could be incorporated into the 'clinical relevance' criterion, in that the prohibition of other forms of treatment (unless they were to create harmful interactions for the patient) would be a deviation from normal clinical practice, rendering the treatment not clinically relevant. However, this issue could fall under less stringent eligibility criteria as well, because the use of other therapies may be listed as an exclusion criterion, as the authors of the tool mention. Thus, maybe 'allowance of concurrent or adjuvant therapies' could be a stand-alone criterion, and information on judging this item could come from the description of either the eligibility criteria or the treatment protocol.

An editorial comment on Gartlehner et al.'s scale questioned the tool's validity and applicability, and shared similar concerns regarding poor item clarity [9]. Spigt and Kotz further emphasized that more work is needed to develop the criteria and make them more

useful to researchers. Gartlehner and colleagues addressed this comment and pointed out that the criteria were defined broadly so that they could be more clearly adapted to individual research questions [10]. They further noted that the tool was merely a starting point for the development of such criteria and they welcomed further discussion on the topic.

4.3 The efficacy-effectiveness distinction: importance of the condition of interest and the nature of the intervention

It was an important objective of this thesis to apply the "effectiveness" tool not only to studies of medications (usually considered simple treatment strategies), but also to studies of non-pharmacological interventions that by nature, are more complex. The stroke rehabilitation literature was optimal for this objective due to the array of both types of interventions that are implemented. Another interesting feature of this literature for evaluating the efficacy-effectiveness construct, is that the interventions are not usually delivered in a primary care setting. Thus, the question that remains is whether the differentiation between efficacy and effectiveness studies depends on either the nature of the intervention (simple versus complex) and/or the condition under investigation (treated in primary care versus not).

Given that stroke rehabilitation was the area of focus, raters fully expected that few studies would meet the 'Populations in primary care' criterion (Item 1), if they were to follow the authors' textual description of the item. The operational definition employed allowed raters to differentiate between usual and more specialized care settings, thus allowing more variation on this item within the sample of stroke rehabilitation studies. The revised criterion was rarely met in trials of pharmacological interventions. Perhaps

drug trials are more likely than trials of non-pharmacological interventions to be conducted in academic, specialized institutions, where the infrastructure for such trials exists, due to partnerships with the pharmaceutical industry.

Although the authors note that the primary care criterion may not be applicable to rare conditions or those that require specialized services, perhaps the more appropriate terminology for this criterion could be 'usual care setting' as opposed to 'Populations in primary care.' That way, raters could judge the item regardless of the condition under investigation and whether or not it tends to be treated in a primary care setting. But, the fact that a condition is not treated in primary care may have implications for the remaining criteria on the scale. For example, a study could appear to have relatively strict inclusion criteria if the indication requires specialized services, even though usual clinical practice would also likely follow a strict process in referring patients to these services. Furthermore, the target population requiring such services could be limited resulting in a relatively small sample available for such a study. Regardless of the typical care setting for any health condition, it should be possible to demonstrate both efficacy and effectiveness; however, different criteria definitions for studies evaluating conditions treated in and outside the primary care setting may be required.

Since health outcomes tend to be of primary interest in stroke rehabilitation, especially when the intervention is non-pharmacological, it was assumed, *a priori*, that most studies would fulfill this criterion if judged according to the author's description. Therefore, similar to 'Populations in primary care', the 'Health outcome' criterion also had to be revised. The implemented definition of "health outcome," appears to have allowed more variation on the item than if it had not been operationalized as such.

Macpherson et al., give examples of outcomes common to each trial type: joint range of motion for the explanatory, and quality of life for the pragmatic trial [11] – this is analogous to the distinction made by the operational definition employed in the present study. For the purposes of this study, measures of function at the lowest level (ie. body structures and functions) were considered non-health outcomes and could be distinguished from those at the higher level (ie. activities and participation), which *were* considered health outcomes. Such a modification however, is likely not necessary when assessing studies of pharmacological treatments, where clinical indicators and biological markers are often seen as principal outcomes, even in stroke rehabilitation. In other words, the measures of function at the lower level could be considered health outcomes in pharmaceutical studies, since they are certainly more patient-relevant measures than clinical markers. Thus, what exactly constitutes a "health outcome" may differ depending on whether the intervention is pharmacological or non-pharmacological.

Regarding assessment of adverse events, the definition employed appears to have been too inclusive, as nearly all pharmacological studies met this criterion. This is likely because the definition encompassed side effects *and* adverse events. Gartlehner et al. seem to suggest that effectiveness studies should provide a more comprehensive assessment of adverse events than efficacy studies. Perhaps then, for pharmacological trials, assessment of side effects only may not be adequate to satisfy this criterion. On the other hand, with non-pharmacological interventions, especially the types common in stroke rehabilitation, minor side effects may be the most adverse events that could occur. Thus, the scope of assessment required to fulfill this criterion may need to differ

depending on the types and seriousness of the adverse events associated with the intervention.

A final point on how the differentiation between efficacy and effectiveness studies may differ depending on the nature of the intervention is the sequencing of these study designs in the evaluation of the interventions. Macpherson suggests that for complex interventions the usual efficacy to effectiveness sequence could be reversed [11]. This would allow for an understanding of the overall effectiveness of the treatment strategy, followed by a move towards understanding the active ingredients in that strategy, which he termed "unpacking the black box" [11]. Overall, it appears that the nature of the intervention and the condition of interest do have implications for how one might differentiate between efficacy and effectiveness research approaches. This concept could be helpful to those wishing to further develop the criteria proposed by Gartlehner and colleagues.

4.4 Incorporating the efficacy-effectiveness distinction into evidence-based decision-making

The authors of the tool maintain that the differentiation between efficacy and effectiveness is an important part of the critical appraisal process, and suggest that their instrument could be used by both clinicians who want to evaluate the generalizability of a study's results and by researchers who wish to distinguish between the study types when producing systematic reviews [1]. Within systematic reviews in particular, they further propose that some of the criteria could be used as eligibility criteria for the inclusion of individual studies [1]. Likewise, criteria of this nature could be considered by health economists when evaluating the suitability of studies for economic analyses. Zwarenstein

and Oxman state that "[the criteria] are an important step toward disentangling characteristics of trials that determine their usefulness and undertaking empirical methodological studies of randomized trials. They can be used to explore possible sources of heterogeneity in systematic reviews and help ensure that trials are designed to answer pragmatic questions that are important to clinicians and other decision-makers" [12]. On a similar note, the total scale score or specific criteria deemed to be most pertinent, could be incorporated into meta-regression analyses, used alongside meta-analyses to explore heterogeneity of treatment effects.

These types of criteria can also have potential applications within the grading of research evidence and subsequent establishment of levels of evidence. Just as it has been suggested that measures of external validity should be used alongside the scales that are used to rate internal validity or study quality [13], so too could the proposed criteria be used in a similar manner. Furthermore, the tool could provide insight into downgrading evidence that is not clinically relevant or applicable. For example, the GRADE working group, in their proposal of a new system to grade research evidence and strength of recommendations, note that uncertainties regarding the directness of the evidence, for example, use of surrogate outcomes and/or lack of direct treatment comparisons, would imply a decrease in the grade of evidence [14]. The same group emphasized that in order to translate levels of evidence into recommendations, one must consider the trade-offs between benefits and harms, the quality of the evidence, the ability to translate the evidence into a specific practice setting, and the uncertainty regarding baseline risk for the target population – major uncertainties with any of which may lower the confidence in the recommendation [14]. Thus, if incorporating Gartlehner et al.'s criteria into a

similar grading process, generally low scale scores or the absence of certain criteria could mean a resulting decrease in the grade of evidence or strength of a recommendation, even if the aggregated body of evidence consists of studies that are otherwise considered to be of high methodological quality. While all of these applications are indeed plausible, it should be noted that there is a recognized scarcity of effectiveness or pragmatic studies in the literature [15, 16], so the tool and other criteria like it, may not be as relevant and applicable at the present time as they will be in the future as more pragmatic studies become available for review.

It should be noted that not all authors believe effectiveness or pragmatic studies to be worthy of higher standards of evidence. McMahon states that the distinction between explanatory and pragmatic is not as relevant as it was in Schwartz's time [4]. He maintains that this distinction was more valuable when the RCT design was implemented solely for the evaluation of pharmaceutical (therapeutic) treatments; however as the design has started being used for the assessment of non-pharmaceutical interventions, the concept of pragmatism has been used to excuse poorly controlled studies in these complex treatment areas. He emphasizes that having less control over study conditions, in an attempt to make a trial emulate 'real life' should not be desired or welcomed as pragmatic, and this lack of control may make pragmatic studies the least generalizable of all. He also refutes the notion of achieving homogeneity in human populations. Finally, he states that most RCTs (especially of drugs) should be labeled explanatory and that poorly controlled pragmatic studies will dominate in the area of "non-therapeutic" complex treatment strategies [4]. Ernst and Canter assert that pragmatic trials are limited due to their neglect of causal proof and believe that the link between treatment and

response weakens as more elements of the efficacy trial design are omitted; however, they do emphasize that the results of pragmatic trials can compliment those from efficacy trials, but only after efficacy has been demonstrated [17]. Their main argument is that pragmatic trials are comparatively weak research tools compared to efficacy studies and that the former should never be a substitute for the latter [17].

4.5 Strengths and limitations of the thesis

A main strength of this thesis was the large sample of studies that was obtained that encompassed many different interventions. Furthermore, since the sample was restricted to one medical condition, stroke rehabilitation, it was possible to evaluate reliability and validity within a single 'population', so to speak, which is important for this type of evaluation. Moreover, multiple, experienced raters were involved in the application of the tool, there was an attempt to standardize the items to achieve acceptable reliability, and studies were scored in an independent manner.

A limitation of this thesis concerns the statistical methods employed. Dichotomization of the total scale scores for evaluations of both the reliability and validity of the instrument likely led to a loss of statistical power; however, constraints in the data motivated this decision. It would have been inappropriate to treat the scale as a continuous measure, when at best it may be considered to produce ordinal data. At the outset, it was hoped that the total score could be dichotomized using a cutoff of six criteria, which (as suggested by the authors of the tool) identifies a study as the effectiveness type; however, this would have resulted in extremely unbalanced groups in terms of sample size, with only four studies in the entire sample reaching this cutoff. Because the distribution of total scores was skewed in this sample, the total score was

dichotomized around three criteria to achieve balance; however, it is not suggested that this number of criteria is significant in any way.

Also, the analyses in this thesis were not very advanced. It would have been optimal to perform a regression analyses in an attempt to predict total scores to support the validation hypotheses; however, to attempt this would have resulted in several violations of statistical assumptions, considering the limitations of the available dataset. With such a large sample of studies, it was a surprise not to find a wider variation in scores. This could be because studies on the effectiveness end of the spectrum are rare in the field of stroke rehabilitation, much like in other treatment areas, or it could be due to an inability of the tool to produce a great deal of variance in scores.

The validation should be interpreted cautiously, as it was only meant to be a crude analysis in an attempt to provide some support for the validity of the instrument beyond face validity. Without "gold standard" definitions for the constructs of efficacy and effectiveness, validity was challenging to assess. Perhaps an alternate or additional method of validation would have been to follow the process used in the development of the tool, and see how well the tool identified nominated efficacy and effectiveness trials in stroke rehabilitation. This method may have been problematic though, as it would have been challenging to find individuals who have "expertise" in labeling a trial as one type or the other.

Lastly, raters hoped that the operational definitions employed could only improve the tool, but it is possible that the scale may have inadvertently been changed to what Gartlehner and colleagues did not intend. Furthermore, it may have been better to evaluate the reliability of the instrument prior to standardization to see if any

improvement was gained. The pilot application of the tool to 10 studies from the EBRSR database brought about many concerns, and the raters struggled to score the items, thus it would have been an arduous task to evaluate the reliability of the tool in its original format.

4.6 Directions for future research

Criteria such as those proposed by Gartlehner and colleagues are an important part of the movement towards recognition of research that is not only methodologically sound but that also provides relevant information to decision-makers and results that can be generalized to the intended target population. With the demonstrated difficulties in applying the scale, and the low inter-rater reliability, however, it is evident that the criteria require further refinement. Possible revisions of the tool, as mentioned previously, could include establishing clear operational definitions for the criteria considering the intervention and patient population of interest; addressing the multiple question nature of some of the items; ensuring each item is independent of the others; excluding items that are not as relevant (such as ITT analysis); and adding items that have not been wholly addressed (such as nature of the comparative therapy and issues of compliance or adherence).

Furthermore, it remains unclear exactly which items should comprise the efficacy-effectiveness construct. Rothwell has proposed a checklist for assessing a study's external validity, the elements of which are strikingly similar to Gartlehner et al.'s criteria, including: setting, patient selection, patient characteristics, differences between trial protocol and routine practice, outcome measures and follow-up, and adverse effects of treatment [18]. The items on the effectiveness scale were chosen on the basis that they

improve a study's external validity, thus, it is unclear whether this published scale is more a measure of external validity than of effectiveness. This confusion lies in the fact that the literature has offered many different, yet no "universally accepted" definitions of efficacy and effectiveness. Thus, it needs to be discussed whether it is correct to equate efficacy with internal validity and effectiveness with external validity, as so many have suggested, or if the constructs can be separated and more correctly conceptualized as i) aiming to isolate and understand the true effect of treatment versus ii) comparing treatments in their entire form to enable a decision to be made regarding the best alternative. Even further, perhaps it was inaccurate for the literature to use efficacy synonymously with explanatory and effectiveness with pragmatic. The terms explanatory and pragmatic are often used in reference to types of clinical trials, whereas effectiveness research often implies non-randomized study designs or simply an evaluation of patient or system outcomes in normal clinical practice. Thus, it is obvious that more groundwork is needed to properly define these constructs.

Further research should include, as mentioned in section 4.4, incorporating these types of criteria into previously conducted meta-analyses to see if they can help explain heterogeneity of treatment effects across individual studies. On a similar note, Rothwell recommends further research focused on the external validity of RCTs in relation to measured treatment effect [18], much like other research that has focused on elements of internal validity or study quality, such as concealed allocation and blinding, and their association with treatment effect size [19-22].

Moreover, several authors have recognized the lack of information on the conduct and reporting of pragmatic trials [23], as well as the lack of infrastructure and funding

options for these types of research designs [16]. In order for criteria such as those proposed by Gartlehner and colleagues to be truly beneficial, more pragmatic trials need to be conducted and they need to be subjected to higher reporting standards so that reviewers can truly judge their quality as well as the relevance and generalizability of their results. There has been a suggestion to take greater consideration of external validity in the design and reporting of RCTs [18]. Moreover, it has been suggested that while the results of large pragmatic trials are more generalizable than small explanatory trials, they also are often more difficult to apply due to the various subgroups within larger samples [24]. Possibly, more informational studies on the proper conduct and interpretation of subgroup analyses is required in order to help decision-makers with the application of research results from effectiveness studies.

Lastly, further commentary and emphasis on the threat of publication bias in systematic reviews is necessary as it could be argued that this is one of the reasons for effectiveness RCTs being largely absent from systematic reviews. Specifically, the notion that the findings of effectiveness trials may not be as significant and/or positive as efficacy trials, due to the introduction of more noise around the treatment effect, may hamper their likeliness to be published and hence their potential to be included in systematic reviews. Thus, even if these types of studies are being conducted, if they fail to be published then it is very difficult to incorporate their findings into evidence-based decision-making.

4.7 Conclusions

The tool by Gartlehner et al. was developed to easily distinguish effectiveness from efficacy studies; however, this process may not be as simple as one would hope. The literature has offered various archetypal characteristics of the trial types, but uncertainty remains regarding the most important elements of the efficacy-effectiveness construct. If reviewers and decision-makers are to use the proposed criteria to make judgments regarding a study's design and inferences regarding external validity of a study's results, then items need to be further standardized in order to improve reliability, or multiple raters should arrive at consensus on a final rating. Furthermore, the factors that differentiate efficacy and effectiveness studies may vary depending on whether the intervention is pharmacological or non-pharmacological in nature, and whether or not the target population receives the intervention in a primary care setting.

The notion of imposing a rigid hierarchy of evidence upon all levels of decision-making in healthcare may become superseded by more flexible models that take into account different study designs and place more emphasis on the generalizability and applicability of evidence. Although further refinements are needed, the proposed criteria are an important advance as they bring to light the importance of judging the aim, relevance, and external validity of a given study within the decision-making process. Increased awareness of this sort will prove to be important for establishing an infrastructure that facilitates the proper conduct of effectiveness or pragmatic trials and the appropriate dissemination of their results.

4.8 References

1 Gartlehner G, Hansen RA, Nissman D, Lohr KN, Carey TS: A simple and valid tool distinguished efficacy from effectiveness studies. J Clin Epidemiol 2006;59:1040-1048.

2 Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz KF, Simel D, Stroup DF: Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA 1996;276:637-639.

3 Collins J: Which Randomized Controlled Trials Are Relevant to Clinical Practice? Obstet Gynecol 2005;106:216-218.

4 McMahon AD: Study control, violators, inclusion criteria and defining explanatory and pragmatic trials. Stat Med 2002;21:1365-1376.

5 Conine TA, Hershler C: Effectiveness: a neglected dimension in the assessment of rehabilitation devices and equipment. Int J Rehabil Res 1991;14:117-122.

6 Feinstein AR: An additional basic science for clinical medicine: II. The limitations of randomized trials. Ann Intern Med 1983;99:544-550.

7 Godwin M, Ruhland L, Casson I, MacDonald S, Delva D, Birtwhistle R, Lam M, Seguin R: Pragmatic controlled clinical trials in primary care: the struggle between external and internal validity. BMC Med Res Methodol 2003;3.

8 Depp C, Lebowitz BD: Clinical trials: bridging the gap between efficacy and effectiveness. Int Rev Psychiatry 2007;19:531-539.

9 Spigt MG, Kotz D: Comment on: "a simple and valid tool distinguished efficacy from effectiveness studies". J Clin Epidemiol 2007;60:753-755.

10 Gartlehner G, Hansen RA, Nissman D, Lohr KN, Carey TS: Authors' reply to comment on: a simple and valid tool distinguished efficacy from effectiveness studies. J Clin Epidemiol 2007;60:752-755.

11 Macpherson H: Pragmatic clinical trials. Complement Ther Med 2004;12:136-140.

12 Zwarenstein M, Oxman A: Why are so few randomized trials useful, and what can we do about it? J Clin Epidemiol 2006;59:1125-1126.

13 Green LW, Glasgow RE: Evaluating the relevance, generalization, and applicability of research: issues in external validation and translation methodology. Eval Health Prof 2006;29:126-153.

14 Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, Guyatt GH, Harbour RT, Haugh MC, Henry D, Hill S, Jaeschke R, Leng G, Liberati A, Magrini

N, Mason J, Middleton P, Mrukowicz J, O'Connell D, Oxman AD, Phillips B, Schunemann HJ, Edejer TT, Varonen H, Vist GE, Williams JW, Jr., Zaza S: Grading quality of evidence and strength of recommendations. BMJ 2004;328:1490.

15 Haynes B: Can it work? Does it work? Is it worth it? The testing of healthcare interventions is evolving. BMJ 1999;319:652-653.

16 Tunis SR, Stryer DB, Clancy CM: Practical Clinical Trials: Increasing the Value of Clinical Research for Decision Making in Clinical and Health Policy. JAMA 2003;290:1624-1632.

17 Ernst E, Canter PH: Limitations of "pragmatic" trials. Postgrad Med J 2005;81:203.

18 Rothwell PM: External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet 2005;365:82-93.

19 Schulz KF, Chalmers I, Hayes RJ, Altman DG: Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995;273:408-412.

20 Kunz R, Oxman AD: The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. BMJ 1998;317:1185-1190.

21 Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P, Klassen TP: Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 1998;352:609-613.

22 Pildal J, Hrobjartsson A, Jorgensen K, Hilden J, Altman D, Gotzsche P: Impact of allocation concealment on conclusions drawn from meta-analyses of randomized trials. Int J Epidemiol 2007.

23 Glasgow RE, Magid DJ, Beck A, Ritzwoller D, Estabrooks PA: Practical clinical trials for translating research to practice: design and measurement recommendations. Med Care 2005;43:551-557.

24 Rothwell PM: Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet 2005;365:176-186.

APPENDICES

Appendix A. Scoring sheet and notes used by raters

Article ID #:

Item	Yes	No	No/Inadequate Description
1. Populations in primary care			
2. Less stringent eligibility criteria			
3. Health outcomes			
4. Long study duration; clinically relevant treatment modalities			
5. Assessment of adverse events			
6. Adequate ample size to assess a minimally important difference from a patient perspective			
7. Intention to treat analysis			

Total Score (total # "Yes" responses): /7

Adapted from: *Gartlehner G, Hansen RA, Nissman D, Lohr KN, Carey TS: A simple and valid tool distinguished efficacy from effectiveness studies. J Clin Epidemiol 2006;59:1040-1048.*

Notes:

- Make note whether there was an inadequate description in the report to allow a yes or no response for an item; these cases will be treated as absent (will receive a "No" rating) in analyses

- The following explanations of the items incorporate the author's textual descriptions as well as the operational definitions we discussed after piloting the tool

1. Populations in primary care
The setting of the study should reflect the initial care facilities available to a diverse population with the condition of interest. Specialized inpatient stroke rehabilitation units/facilities; rehabilitation hospitals/centers; academic/teaching hospitals; research laboratories; would not be considered primary care for stroke rehabilitation. Hospitals; general medical wards; general rehabilitation departments/wards; outpatient rehabilitation departments; home; would be considered primary care. Setting refers to where the intervention is taking place – if assessments take place at a research center that's OK as long as the intervention takes place in the primary care setting.

2. Less stringent eligibility criteria
Eligibility criteria must allow the sample to reflect the heterogeneity (comorbidities, variable compliance rates, and use of other interventions) of the general population affected by the condition under consideration. Applying both recurrent stroke and mild and/or moderate cognitive impairment as exclusion criteria, place great limits on the eligible study population. Comorbidities and other medication/therapy use should not be general exclusion criteria unless they would contraindicate participation in the intervention in ordinary clinical practice. Most studies in stroke rehabilitation acquire subjects as they are admitted into hospital or rehab facilities; however, if the study needs to recruit former stroke patients that are no longer receiving care, look at the recruitment methods to ensure that they are attempting to capture a diverse sample of patients. Regarding drug studies, if there is a pre-randomization run-in period, it is likely that the eligibility criteria are very stringent as they may be looking to exclude placebo-responders or poor compliers. It may be helpful to look at the study's flowchart, if available, to compare how many subjects were initially screened versus how many were actually included in the final sample.

3. Health outcomes
Looking for functional and health status outcomes (activity and participation outcomes according to the ICF). Any clinical indicators (eg. BMD), event rates (mortality), or measures of impairment (body function according to the ICF) would not be considered health outcomes in the case of rehabilitation. Since most rehab studies are assessing

function in some way, we are considering body function/impairment measures as non-health outcomes because they are measuring at the lowest level of function. The health outcome does not have to be the primary outcome but must be a main outcome that is emphasized in the results.

4. Long study duration; clinically relevant treatment modalities
Study must have both features to get a "yes". Long meaning an *appropriate* length for the given intervention (reflecting the minimum length of therapy in clinical practice), including an appropriate length of follow-up to assess the outcome(s) of interest. Treatment modalities should be clinically relevant (therapy is provided as it would in regular clinical practice – intensity, usual resources – and diagnoses rely on diagnostic standards used in clinical practice. If the intervention consists of a novel modality (eg. a new hi-tech device) it cannot be clinically relevant and we cannot know an appropriate study duration. Regarding drug studies, the administration/prescribed dosage of medication should reflect clinical practice as well (no fixed-dose designs or equivalent dosages when comparing active drugs). For both therapy and drug studies there should be some flexibility in the "dosage" – looking for more individualized therapy that is dependent on the subject's progress – not imposing the exact same therapy on everyone in the study.

5. Assessment of adverse events
Study must either outline any adverse events or side effects that occurred over the course of therapy or state that none occurred. Assessment of compliance/discontinuation rates is not adequate on its own.

6. Adequate sample size to assess a minimally clinically important difference
Regardless of the outcome on which the study has based its sample size, the study must make some reference (either to their own previous pilot work or to other background literature) that the difference they are trying to detect is the smallest clinically meaningful/important difference on the specified outcome measure – they cannot just say they were looking to detect a __ difference between groups. We are not assessing this item in terms of patient perspective but in terms of whether the study was looking to detect small, clinically relevant differences between groups.

7. ITT analysis
Study must state to some effect that data was analyzed according to the intention-to-treat principle.

Appendix B. PEDro quality rating scale items and requirements used in the Evidence-based Review of Stroke Rehabilitation (EBRSR)

PEDro Item	Requirement
1. Subjects were randomly allocated to groups (in a cross-over study, participants were randomly allocated an order in which treatments were received).	Point was awarded if random allocation of patients was stated in the methods, but the method of randomization need not be specified. Coin-tossing and dice-rolling procedures were considered random. Quasi-randomization allocation procedures, such as allocation by bed availability, did not satisfy this criterion.
2. Allocation was concealed.	Point was awarded if this was explicitly stated in the methods or if there was reference made to the fact that allocation involved sealed opaque envelopes or contacting an "off-site" holder of the allocation schedule.
3. The groups were similar at baseline on important prognostic factors.	Point was awarded if at least one key outcome measure at baseline was reported for the study and control groups. This criterion was satisfied even if only baseline data of study completers were presented.
There was blinding of: 4. all subjects; 5. all therapists who administered the therapy; and 6. all assessors who measured at least one key outcome.	The person in question was considered blind if he/she did not know to which group the participant was allocated. Participants and therapists were only considered blind if it could be expected that they would be unable to distinguish between the applied treatments. In drug therapy trials, the drug administrator (the therapist) was considered blind if he/she did not prepare the drug and was unaware of which drug was being administered.
7. Adequacy of follow-up.	Point was awarded if all of the originally randomized participants were accounted for by study end – this interpretation differs from PEDro, where adequacy is defined as the measurement of the main outcome in > 85% of participants.
8. All participants for whom outcome measures were available received their allocated treatments or, where this was not the case, data for at least one key outcome were analyzed by "intention to treat".	Point was awarded if the trial explicitly stated that analysis was by intention-to-treat.
9. The results of between-group statistical comparisons are reported for at least one key outcome.	Scoring of this criterion was design dependent. The analysis was considered a between-groups analysis if either a simple comparison of post-treatment outcomes or a comparison of the change in one group with that in another was made. The comparison may take the form of significance testing or confidence intervals. Point was awarded if between-groups comparison on at least one outcome measure was made and its analysis of comparison was provided.
10. The study provides both point measures and measures of variability for at least one key outcome.	A point measure was defined as the measure of the treatment effect size. The treatment effect was described as being either a difference in group outcomes or as the outcome in (each of) all groups. Measures of variability included standard deviations, standard errors, confidence intervals, and ranges. Point measures and/or measures of variability that were provided graphically were awarded a point if it was clear what was being graphed. For categorical outcomes, the number of participants in each category had to be given for each group.

Note: Identification of eligibility criteria (an additional PEDro item) was not evaluated for studies in the EBRSR because participant selection influences the external validity, not the internal or statistical validity, of a study.

Adapted from: *Foley NC, Teasell RW, Bhogal SK, Speechley MR: Stroke Rehabilitation Evidence-Based Review: methodology. Top Stroke Rehabil 2003;10:1-7.*

Appendix C. Validation variable definitions

Variable	Definition
Type of intervention	
Pharmacological	Included pharmaceutical preparations (tablets, injections, etc.) as well as vitamin therapy, other nutritional supplement therapies, and sunlight exposure therapy.
Non-pharmacological	Included all other forms of therapy as well as models of care which may or may not have included delivery of pharmaceutical therapies within the context of care delivery (eg. stroke unit care).
Type of control group	
Placebo	The comparator group(s) received an inert substance, indistinguishable from the experimental pharmacological therapy. The report stated that the comparator was a placebo.
Sham	The comparator group(s) received a mock therapy of the experimental non-pharmacological therapy (eg. sham acupuncture where non-acupuncture points are stimulated) OR another pseudo-therapy (potentially active) in which the subject would be unaware of experimental therapy (eg. therapy aimed at the non-affected side). The report did not necessarily state that it was a sham therapy – other terms used were pseudo-therapy, placebo-like intervention, or else the sham nature was inferred from the author's description of the intervention.
One factor varied	The comparator group(s) received a therapy that differed from the experimental therapy on only one factor (eg. intensity, therapist, equipment). This was inferred from the author's description of the intervention.
No treatment	The comparator groups(s) either received no form of treatment OR, in the cases where both treatment groups received the same "other" or "standard" rehabilitation care, no *additional* treatment.
Active treatment	The comparator group(s) received a different active form of therapy than the experimental group (eg. another pharmaceutical agent or model of care).
Number of centers	
Single-center	Report stated the study was conducted at one center only. In cases where the report lacked information on the number of participating centers, it was assumed to be a single-center study.
Multi-center	Report stated that the study took place at more than one center.
Sequence of study	
Pilot	Report stated that it was a "pilot" or "preliminary" study.
Full-scale RCT	Study was not specifically referred to as a "pilot" "preliminary" or "follow-up" RCT.
Follow-up RCT	Report stated that it was a follow-up study of an RCT
Quality rating	
High Quality	Study had been assigned a PEDro score greater than or equal to 7, as assessed by reviewers for the EBRSR
Low Quality	Study had been assigned a PEDro score less than 7, as assessed by reviewers for the EBRSR

Appendix D. Description of inter-rater reliability statistics used

- **Multiple-Rater Kappa Statistic**

Fleiss's multiple-rater kappa statistic (**κ**) was used to assess the inter-rater reliability of the seven individual scale items. Fleiss (1981) adapted the kappa statistic (a chance-corrected measure of agreement) devoted to the case of two raters, to the case of multiple raters.

For the case of a dichotomous variable:

$$\kappa = \frac{BMS - WMS}{BMS + (\bar{m} - 1)WMS} = \frac{1 - \sum \frac{x_i(m_i - x_i)}{m_i}}{n(\bar{m} - 1)\bar{p}\,\bar{q}}$$

where

$$BMS = \frac{1}{n} \sum \frac{(x_i - m_i \bar{p})^2}{m_i}$$

(BMS = between-subject mean square)

$$WMS = \frac{1}{n(\bar{m} - 1)} \sum \frac{(m_i - x_i)}{m_i}$$

(WMS= within-subject mean square)

n = number of subjects (trials)
x_i = number of positive ratings on subject i
m_i = number of ratings on the i^{th} subject
$m_i - x_i$ = number of negative ratings on subject i

$$\bar{p} = \sum x_i / n\bar{m} \quad \bar{m} = \sum m_i / n \quad \bar{q} = 1 - \bar{p}$$

Since the number of ratings per subject were equal in this case, and the sample size was relatively large, the large sample variance of kappa, derived by Fleiss, Nee and Landis (1979) was employed:

$$Var(\kappa) = \frac{2}{Nn(n-1)(\sum p_j q_j)^2} \times [(\sum p_j q_j)^2 - \sum p_j q_j (q_j - p_j)]$$

where n = number of ratings per subject
N = number of subjects
j = number of rating categories
p_j = proportion of all assignments to the jth category
$q_j = 1 - p_j$

Values for kappa range from -1.0 to 1.0 with values above zero representing agreement levels higher than chance, values of zero equivalent to chance agreement, and values below zero representing agreement levels less than chance. The calculation of the multiple-rater kappa values and their associated 95% confidence intervals were performed using an Excel template developed by King (2004).

- **Friedman test**

The Friedman test was used to assess the inter-rater reliability of the total scale score (un-dichotomized). Friedman (1937) developed this statistical method as a non-parametric alternative to the analysis of variance (ANOVA). This test makes use of ranked data and tests the hypothesis that the mean ranks from each category come from a "single homogeneous normal universe." In this case, the mean ranks among raters are being compared to assess whether there are any distributional differences in total scores across raters.

The test statistic follows a chi-square distribution and is calculated as follows:

$$\chi^2_r = \frac{p-1}{p\sigma^2} \sum(\bar{r}_i - p)^2$$

$$= \frac{12n}{p(p+1)} \sum\{\bar{r}_i - \tfrac{1}{2}(p+1)\}^2$$

where

p = the number of ranks

$\bar{r}_i$ = the mean rank of the j^{th} column
n = the number of rows (the number of ranks averaged)

..... and is distributed with a mean (ρ) and variance (σ^2):

$$\rho = \tfrac{1}{2}(p+1)$$

$$\sigma^2 = (p^2-1)/(12n)$$

Appendix E. Power analysis: precision of inter-rater reliability estimates

For the assessment of inter-rater agreement among 3 raters, with an alpha level of 0.05:

W = width of 95% confidence interval for multiple-rater kappa coefficient
= 2 (1.96) {var(P)}^0.5,

where {var(P)}= (1-P) / 3N [(1-8P) (1-P) + P (3-2P) /π (1-π)]

and N = number of subjects (papers)
P = planning value of multiple-rater kappa coefficient
π = probability of successful rating (study classified as an effectiveness study)

Reference: *Altaye M, Donner A, Klar, N. Inference Procedures for Assessing Interobserver Agreement among Multiple Raters. Biometrics, 2001;57:584-8.*

Estimated width of 95% confidence interval (CI) for multiple-rater kappa coefficient for given values of N, P, and π.

N	P	π	W	N	P	π	W
50	0.3	0.1	0.709509	70	0.3	0.1	0.599645
		0.3	0.419031			0.3	0.354146
		0.5	0.369119			0.5	0.311962
	0.4	0.1	0.721015		0.4	0.1	0.609369
		0.3	0.420042			0.3	0.355001
		0.5	0.367729			0.5	0.310787
	0.5	0.1	0.701637		0.5	0.1	0.592991
		0.3	0.408753			0.3	0.345459
		0.5	0.357845			0.5	0.302435
	0.6	0.1	0.655316		0.6	0.1	0.553843
		0.3	0.385297			0.3	0.325636
		0.5	0.338727			0.5	0.286276
	0.7	0.1	0.583131		0.7	0.1	0.492835
		0.3	0.348564			0.3	0.294591
		0.5	0.308661			0.5	0.260866
60	0.3	0.1	0.64769	80	0.3	0.1	0.560916
		0.3	0.382521			0.3	0.331273
		0.5	0.336958			0.5	0.291814
	0.4	0.1	0.658193		0.4	0.1	0.570012
		0.3	0.383444			0.3	0.332073
		0.5	0.335689			0.5	0.290715
	0.5	0.1	0.640504		0.5	0.1	0.554692
		0.3	0.373139			0.3	0.323148
		0.5	0.326667			0.5	0.282902
	0.6	0.1	0.598219		0.6	0.1	0.518073
		0.3	0.351727			0.3	0.304604
		0.5	0.309214			0.5	0.267787
	0.7	0.1	0.532323		0.7	0.1	0.461005
		0.3	0.318194			0.3	0.275564
		0.5	0.281768			0.5	0.244018

Appendix F. Table of included studies

Study author and year of publication	General condition under investigation	Brief description of treatment and control condition(s)	Item number*							Total score
			1	2	3	4	5	6	7	
RCTs of pharmacological interventions										
Almeida et al. (2006) [1]	Depression	Sertraline vs. placebo	-	+	-	-	-	-	+	2
Ashtary et al. (2006) [2]	Aphasia	Bromocriptine vs. placebo	-	-	-	-	+	-	+	2
Attal et al. (2000) [3]	Central pain	Lidocaine vs. placebo	-	+	-	-	+	-	-	2
Attal et al. (2002) [4]	Central pain	Morphine vs. placebo	-	+	-	+	+	-	-	3
Bakheit et al. (2000) [5]	Upper limb spasticity	Botox (1 of 3 doses) vs. placebo	-	+	+	-	+	-	+	4
Bakheit et al. (2001) [6]	Upper limb spasticity	Botox vs. placebo	-	+	-	+	+	-	+	4
Bath et al. (2001) [7]	Acute neurological/functional recovery	Low molecular weight heparin, tinzaparin (high vs. medium-dose) vs. aspirin	-	-	+	+	+	+	+	5
Berge et al. (2000) [8]	Acute thromboembolic complications	Low molecular weight heparin, dalteparin vs. control (aspirin), both with matching placebos	-	-	+	-	-	-	+	2
Berthier et al. (2006) [9]	Aphasia	Donepezil vs. placebo	-	+	-	+	+	+	-	4
Bhakta et al. (2000) [10]	Upper limb spasticity	Botox vs. placebo	-	+	+	+	+	+	+	6
Black et al. (2003) [11]	Cognition/dementia	Donepezil (1 of 2 doses) vs. placebo	-	+	+	-	+	+	+	5
Brashear et al. (2002) [12]	Upper limb spasticity	Botox vs. placebo	-	+	+	+	+	-	-	4
Brashear et al. (2004) [13]	Upper limb spasticity	Botox vs. placebo	+	-	-	+	+	-	-	3
Brown et al. (1998) [14]	Depression	Fluoxetine vs. placebo	+	+	-	-	+	-	-	3

Appendix F. Table of included studies

Study author and year of publication	General condition under investigation	Brief description of treatment and control condition(s)	Item number*							Total score
			1	2	3	4	5	6	7	
Burns et al. (1999) [15]	Depression	Sertraline vs. placebo	-	+	+	-	+	-	+	4
Chemerinski et al. (2001) [16]	Depression	Nortriptyline vs. placebo	-	+	+	-	-	-	-	2
Childers et al. (2004) [17]	Upper limb spasticity	Botox (1 of 3 doses) vs. placebo	+	-	+	+	+	-	-	4
Choi-Kwon et al. (2006) [18]	Depression	Fluoextine vs. placebo	+	-	-	+	+	-	+	4
Cohen et al. (2003) [19]	Cognition/dementia	Citicoline vs. placebo	-	-	-	-	-	-	-	0
De Deyn et al. (1997) [20]	Acute neurological/functional recovery	Piracetam vs. placebo	-	-	+	-	+	+	+	4
Diener et al. (2006) [21]	Acute thromboembolic complications	Low molecular weight heparin, certoparin vs. unfractionated heparin	-	-	-	+	+	-	+	3
Fogari et al. (2004) [22]	Cognition/dementia	Valsartan vs. enalpril	+	-	-	+	+	-	-	3
Forette et al. (2002) [23]	Cognition/dementia	Nitrendipine, with possible addition of other antihypertensives vs. placebo, with open label therapy as needed	-	+	-	+	-	-	+	3
Francisco et al. (2002) [24]	Upper limb spasticity	High vs. low volume botox preparations	-	+	-	+	+	+	-	4
Fruehwald et al. (2003) [25]	Depression	Fluoxetine vs. placebo	-	+	+	+	+	-	-	4
Gariballa et al. (1998) [26]	Nutritional status	Nutritional supplementation (via enteral sip feeding) vs. control (hospital diet only†)	+	-	-	+	-	-	-	2
Ginsberg et al. (2002) [27]	Acute thromboembolic complications	Warfarin vs. placebo	-	-	-	-	+	-	+	2
Grade et al. (1998) [28]	Paresis/hemiplegia	Methylphenidate vs. placebo	-	-	+	-	+	-	-	2

Appendix F. Table of included studies

Study author and year of publication	General condition under investigation	Brief description of treatment and control condition(s)	Item number*							Total score
			1	2	3	4	5	6	7	
Hillbom et al. (2002) [29]	Acute thromboembolic complications	Enoxaparin vs. heparin	-	-	-	+	+	-	-	2
Huber et al. (1997) [30]	Aphasia	Piracetam vs. placebo	-	-	-	-	+	-	-	1
Johnson et al. (2002) [31]	Lower limb spasticity	Botox and functional ES vs. control (PT only†)	+	-	-	-	-	-	-	1
Johnson et al. (2004) [32]	Lower limb spasticity	Botox and functional ES vs. control (PT only†)	-	-	+	+	-	-	-	2
Jorge et al. (2003) [33]	Depression	Fluoxetine vs. nortriptyline, both with matching placebos	-	+	-	+	-	-	+	3
Kessler et al. (2000) [34]	Aphasia	Piracetam vs. placebo	-	-	-	-	-	-	-	0
Kimura et al. (2000) [35]	Depression	Nortriptyline vs. placebo	-	+	-	-	-	-	-	1
Kirazli et al. (1998) [36]	Lower limb spasticity	Botox vs. phenol	-	+	-	+	+	-	-	3
Kong et al. (2007) [37]	Hemiplegic/spastic shoulder pain	Botox vs. placebo	+	+	-	+	+	-	-	4
Lampl et al. (2002) [38]	Central pain	Amitriptyline vs. placebo	-	+	-	+	+	-	-	3
Laska et al. (2005) [39]	Aphasia	Moclobemide vs. placebo	-	-	-	-	+	-	+	2
Lithell et al. (2003) [40]	Cognition/dementia	Antihypertensive, candesartan vs. placebo, with open-label therapy as needed	-	-	-	+	+	-	+	3
Mancini et al. (2005) [41]	Lower limb spasticity	1 of 3 botox doses	-	-	-	-	+	-	-	1
Marco et al. (2007) [42]	Hemiplegic/spastic shoulder pain	Botox vs. placebo	+	+	-	+	+	-	-	4

Appendix F. Table of included studies

Study author and year of publication	General condition under investigation	Brief description of treatment and control condition(s)	Item number*							Total score
			1	2	3	4	5	6	7	
Martinsson & Wahlgren (2003) [43]	Acute neurological/functional recovery	Dexamphetamine (1 of 3 doses) vs. placebo	-	+	+	-	+	-	-	3
Meythaler et al. (2001) [44]	Spasticity	Intrathecal baclofen vs. placebo	-	+	-	+	+	-	-	3
Niedermaier et al. (2004) [45]	Depression	Mirtazapine vs. control (no treatment)	-	-	+	+	+	-	-	3
On et al. (1999) [46]	Lower limb spasticity	Botox vs. phenol	-	+	-	-	-	-	-	1
Palomaki et al. (1999) [47]	Depression	Mianserin vs. placebo	-	+	+	+	+	-	+	5
Pantoni et al. (2000) [48]	Cognition/dementia	Nimodipine vs. placebo	-	-	+	-	+	-	+	3
Perez et al. (1998) [49]	Dysphagia	Nifedipine vs. placebo	-	-	-	-	-	-	-	0
Pittock et al. (2003) [50]	Lower limb spasticity	1 of 3 botox doses	-	+	-	+	+	-	+	4
Poole et al. (2007) [51]	Acute thromboembolic complications	Low molecular weight heparin, enoxaparin vs. unfractionated heparin	-	-	+	-	+	+	+	4
Rasmussen et al. (2003) [52]	Depression	Sertraline vs. placebo	-	-	-	+	+	-	-	2
Reiter et al. (1998) [53]	Lower limb spasticity	Low-dosage botox and ankle taping vs. high-dosage botox	-	-	-	+	+	-	-	2
Robinson et al. (2000) [54]	Depression	Fluoxetine vs. nortriptyline, both with matching placebos	-	+	+	-	+	-	+	4
Rowbotham et al. (2003) [55]	Central pain	High vs. low strength oral opioid therapy, levorphanol	+	-	-	+	+	-	-	3
Sato et al. (1997) [56]	Osteoporosis	1 alpha-hydroxyvitamin D3 vs. placebo	+	-	-	-	-	-	-	1

Appendix F. Table of included studies

Study author and year of publication	General condition under investigation	Brief description of treatment and control condition(s)	Item number*							Total score
			1	2	3	4	5	6	7	
Sato et al. (1999) [57]	Osteoporosis	Ipriflavone vs. 1 alpha-hydroxyvitamin D3 vs. control (no treatment)	+	-	-	-	+	-	-	2
Sato et al. (2000) [58]	Osteoporosis	Intermittent cyclical etidronate therapy vs. placebo	-	-	-	-	+	-	-	1
Sato et al. (2003) [59]	Osteoporosis	Sunlight exposure vs. sunlight deprivation	-	-	-	+	-	-	-	1
Sato et al. (2005a) [60]	Osteoporosis	Risedronate sodium therapy vs. placebo	-	-	-	-	+	-	+	2
Sato et al. (2005b) [61]	Osteoporosis	Risedronate sodium therapy vs. placebo	-	-	+	+	+	-	-	3
Scheidtmann et al. (2001) [62]	Paresis/hemiplegia	Levodopa vs. placebo	-	+	-	-	+	-	-	2
Sherman et al. (2007) [63]	Osteoporosis	Zoledronate vs. placebo	-	-	-	-	+	-	-	1
Smith et al. (2000) [64]	Upper limb spasticity	Botox (1 of 3 doses) vs. placebo	-	+	-	+	+	-	-	3
Snels et al. (2000) [65]	Hemiplegic/spastic shoulder pain	Triamcinolone acetonide vs. placebo	-	+	+	-	+	-	-	3
Sonde et al. (2001) [66]	Paresis/hemiplegia	Amphetamine vs. placebo	-	-	+	-	-	+	-	2
Sonde & Lokk (2007) [67]	Paresis/hemiplegia	Amphetamine vs. levodopa vs. amphetamine and levodopa vs. control (PT only†), with placebo given in absence of one or both active drugs	-	-	+	-	+	+	+	4
Stamenova et al. (2005) [68]	Spasticity	Tolperisone vs. placebo	-	-	+	+	+	-	+	4
Sze et al. (1998) [69]	Memory	Nimodipine vs. control (no treatment)	-	+	-	-	-	-	-	1
Szelies et al. (2001) [70]	Aphasia	Piracetam vs. placebo	-	-	-	-	+	-	-	1

Appendix F. Table of included studies

Study author and year of publication	General condition under investigation	Brief description of treatment and control condition(s)	Item number*							Total score
			1	2	3	4	5	6	7	
Tanaka et al. (1997) [71]	Aphasia	Cholinergic agent, bifemelane vs. control (no treatment)	-	-	-	-	-	-	-	0
Tardy et al. (2006) [72]	Paresis/hemiplegia	Methylphenidate vs. placebo	-	-	-	-	+	-	+	2
TOAST Investigators (1998) [73]	Acute neurological/functional recovery	Low molecular weight heparanoid, danaparoid sodium vs. placebo	-	-	+	+	+	-	+	4
Vestergaard et al. (2001) [74]	Central pain	Lamotrigine vs. placebo	-	-	-	-	+	-	+	2
Walker-Batson et al. (2001) [75]	Aphasia	Dextroamphetamine vs. placebo	-	-	-	+	+	-	-	2
Wiart et al. (2000) [76]	Depression	Fluoxetine vs. placebo	-	-	+	-	+	-	+	3
Wilkinson et al. (2003) [77]	Cognition/dementia	Donepezil (1 of 2 doses) vs. placebo	-	+	+	-	+	-	+	4
Yelnik et al. (2006) [78]	Hemiplegic/spastic shoulder pain	Botox vs. placebo	-	-	-	+	+	-	-	2
RCTs of non-pharmacological interventions										
Ada et al. (2005) [79]	Upper extremity function	Contracture preventive shoulder positioning vs. control (shoulder exercises and standard upper limb care only†)	+	+	-	-	-	-	-	2
Alberts et al. (2004) [80]	Upper extremity function	Immediate vs. delayed constraint-induced therapy	-	-	+	-	-	-	-	1
Alon et al. (2007) [81]	Upper extremity function	Functional ES vs. control (standardized task-specific PT and OT only†)	-	-	+	-	+	-	-	2
Altschuler et al. (1999) [82]	Upper extremity function	Mirror therapy vs. same therapy using	-	-	-	-	-	-	-	0

Appendix F. Table of included studies

Study author and year of publication	General condition under investigation	Brief description of treatment and control condition(s)	Item number*							Total score
			1	2	3	4	5	6	7	
		transparent plastic								
Armagan et al. (2003) [83]	Upper extremity function	EMG biofeedback vs. sham therapy	-	-	-	-	-	-	-	0
Barrett et al. (2001) [84]	General physical function	Encouragement vs. discouragement of wheelchair self-propulsion	-	+	+	+	+	-	+	5
Bonan et al. (2004) [85]	Lower extremity function	Balance rehab with visual cue deprivation vs. free vision	+	-	-	+	-	-	-	2
Boter & HESTIA Study Group (2004) [86]	Models of care delivery	Outreach nursing support program vs. control (standard care only†)	+	-	+	+	-	+	+	5
Carnaby (2006) [87]	Dysphagia	Standard behavioral intervention (high vs. low intensity), consisting of swallowing compensation strategies and dietary prescription vs. usual care	-	+	-	-	-	+	+	3
Chen, I. et al. (2002) [88]	Lower extremity function	Visual feedback balance training with the "SMART Balance Master" vs. control (conventional therapy only†)	+	-	+	-	-	-	-	2
Chen, J. et al. (2005) [89]	Upper extremity function	Thermal stimulation vs. control (standard rehab only†)	+	-	+	-	+	-	-	3
Chen, S. et al. (2005) [90]	Lower limb spasticity	Surface ES vs. sham therapy	-	-	-	-	-	-	-	0
Clark et al. (2003) [91]	Education or reintegration	Family education and counseling vs. control (no treatment)	+	+	+	+	-	+	-	5
da Cunha, Jr. et al. (2002) [92]	Lower extremity function	STAT vs. control (regular rehab only†)	-	-	+	-	-	-	-	1
Daly et al. (2004) [93]	Lower extremity function	FNS vs. control (gait training only†)	-	+	-	-	-	-	-	1

Appendix F. Table of included studies

Study author and year of publication	General condition under investigation	Brief description of treatment and control condition(s)	Item number*							Total score
			1	2	3	4	5	6	7	
de Jong et al. (2006) [94]	Upper extremity function	Contracture preventive arm positioning vs. control (conventional rehab only†)	-	-	-	-	-	-	-	0
de Kroon et al. (2004) [95]	Upper extremity function	ES of hand extensors and flexors vs. ES of extensors only	-	-	+	-	+	-	-	2
de Seze et al. (2001) [96]	Lower extremity function	Trunk control retraining using the "Bon Saint Come" device vs. conventional rehab	-	-	-	-	+	-	-	1
Dean & Shepherd (1997) [97]	General physical function	Task-related reach training vs. sham training	+	+	-	-	-	-	-	2
Desrosiers et al. (2005) [98]	Upper extremity function	Arm therapy programme (repetitive unilateral and symmetrical bilateral tasks) vs. control (usual arm rehab only†)	-	-	+	+	-	-	-	2
Doornhein & De Haan (1998) [99]	Memory	Memory training strategy vs. pseudo-memory training	-	+	-	-	-	-	-	1
Duncan et al. (2003) [100]	General physical function	Structured, physiologically-based, in-home exercise program vs. usual care	+	-	+	+	+	-	+	5
Edmans et al. (2000) [101]	Perception	Transfer-of-training vs. functional approach	-	+	+	+	-	-	-	3
Ertelt et al. (2007) [102]	Upper extremity function	Action observation therapy vs. placebo-like intervention (therapy without viewing action sequences)	-	-	+	-	-	-	-	1
Fagerberg et al. (2000) [103]	Models of care delivery	Stroke unit care vs. general medical ward	-	-	+	-	+	-	+	3
Fink et al. (2004) [104]	Lower limb spasticity	Acupuncture vs. sham therapy	-	+	-	+	-	-	-	2

Appendix F. Table of included studies

Study author and year of publication	General condition under investigation	Brief description of treatment and control condition(s)	Item number*							Total score
			1	2	3	4	5	6	7	
Goulding & Bakheit (2000) [105]	Dysphagia	Determination of prescribed fluid viscosity with a viscometer vs. standard procedure (subjective judgment)	-	+	-	-	-	+	-	2
Grant et al. (2002) [106]	Education or reintegration	SPTP vs. sham intervention vs. control (usual discharge planning only†)	+	+	+	+	-	-	-	4
Green et al. (2002) [107]	General physical function	Community PT vs. control (no treatment)	+	+	+	+	-	+	+	6
Hemmen & Seelen (2007) [108]	Upper extremity function	EMG-triggered feedback and movement imagery vs. control (conventional ES and conventional rehab only†)	-	-	+	-	-	-	-	1
Howe et al. (2005) [109]	Lower extremity function	Lateral weight transference exercises vs. control (usual care only†)	-	-	-	+	-	-	-	1
Indredavik et al. (2000) [110]	Models of care delivery	Extended stroke unit service with ESD vs. ordinary stroke unit service	-	-	+	+	-	-	-	2
Indredavik et al. (1998) [111]	Models of care delivery	Stroke unit care vs. general medical ward	-	+	+	-	-	-	+	3
Jorge et al. (2003) [112]	Depression	Active rTMS vs. sham therapy	-	-	-	-	+	-	-	1
Kalra et al. (2004) [113]	Education or reintegration	Caregiver training vs. conventional care	-	+	+	+	-	+	+	5
Khedr et al. (2005) [114]	General physical function	Active rTMS vs. sham therapy	-	-	+	-	+	-	-	2
Kim et al. (2001) [115]	Lower extremity function	Maximal concentric isokinetic strength training vs. passive range of motion	-	+	-	-	-	-	-	1
Kjendahl et al. (1997) [116]	General physical function	Acupuncture vs. control (multidisciplinary rehab only†)	-	-	+	+	-	-	-	2

Appendix F. Table of included studies

Study author and year of publication	General condition under investigation	Brief description of treatment and control condition(s)	Item number*							Total score
			1	2	3	4	5	6	7	
Lam et al. (2006) [117]	Education or reintegration	Virtual-reality-based transportation skills program vs. video-based psycho-educational program vs. control (no treatment)	-	+	+	-	-	-	-	2
Langhammer et al. (2007) [118]	General physical function	Intensive PT vs. individualized regular exercise	+	-	+	-	-	-	+	3
Lincoln et al. (1999) [119]	Upper extremity function	Routine PT vs. additional PT with qualified therapist vs. additional PT with trained assistant	-	+	+	+	-	-	-	3
Lincoln et al. (2000) [120]	Models of care delivery	Stroke unit care vs. general medical ward	-	+	+	+	-	-	-	3
Logan et al. (1997) [121]	General physical function	Enhanced social services OT vs. routine social services OT	+	+	+	-	-	-	+	4
Lowe et al. (2007) [122]	Education or reintegration	"CareFile" stroke information booklet vs. control (usual pamphlets only†)	-	+	+	+	-	-	-	3
Macko et al. (2005) [123]	Lower extremity function	Progressive treadmill aerobic training vs. conventional rehab program	-	-	+	+	-	-	-	2
Mayo et al. (2000) [124]	ü ü	ESD with home rehab vs. usual discharge and follow-up services	-	-	+	+	-	+	-	3
McDowell et al. (1999) [125]	Urinary incontinence	Behavioral home therapy vs. control (no treatment)	+	+	+	+	-	-	-	4
Moon et al. (2003) [126]	Upper limb spasticity	Electroacupuncture vs. moxibustion vs. routine acupuncture	-	+	-	-	-	-	-	1
Moseley (2004) [127]	Complex regional pain	Motor imagery program vs. waiting-list	-	-	-	-	-	-	-	0

Appendix F. Table of included studies

Study author and year of publication	General condition under investigation	Brief description of treatment and control condition(s)	Item number*							Total score
			1	2	3	4	5	6	7	
		control (ongoing medical management)								
Ouellette et al. (2004) [128]	Lower extremity function	High-intensity progressive resistance training vs. placebo-like intervention (upper extremity stretching)	-	-	+	+	+	-	+	4
Page et al. (2004) [129]	Upper extremity function	Modified constraint-induced movement therapy vs. traditional rehab vs. control (no treatment)	+	+	+	-	-	+	-	4
Pang et al. (2005) [130]	Lower extremity function	Community-based fitness and mobility exercise program vs. placebo-like intervention (upper extremity exercise)	+	-	+	+	+	-	+	5
Pang et al. (2006) [131]	Upper extremity function	Community-based group exercise program vs. placebo-like intervention (lower extremity exercise)	+	-	+	+	+	-	-	4
Parker et al. (2001) [132]	General physical function	Leisure-based OT vs. ADL-based OT vs. control (no OT)	+	+	+	+	-	+	+	6
Peurala et al. (2005) [133]	Lower extremity function	BWS gait training and functional ES vs. gait training only vs. over-ground walking only	-	-	+	-	-	-	-	1
Ploughman & Corbett (2004) [134]	Upper extremity function	Forced-use therapy vs. control (conventional therapy only†)	-	-	+	+	+	-	-	3
Ring & Rosenthal (2005) [135]	Upper extremity function	Home functional ES via neuroprosthesis vs. control (outpatient rehab only†)	+	+	+	+	+	-	-	5
Rodgers et al. (1997) [136]	Models of care delivery	ESD vs. conventional care	+	+	+	+	-	-	+	5
Rodgers et al. (1999) [137]	Education or reintegration	Stroke education program vs. control	-	+	+	+	-	-	+	4

Appendix F. Table of included studies

Study author and year of publication	General condition under investigation	Brief description of treatment and control condition(s)	Item number*							Total score
			1	2	3	4	5	6	7	
		(conventional stroke unit care only†)								
Salbach et al. (2004) [138]	Lower extremity function	Functional task-oriented lower extremity intervention vs. placebo-like intervention (upper extremity activity)	+	-	+	+	+	+	+	6
Sulch et al. (2002) [139]	Models of care delivery	Integrated care pathway vs. conventional multidisciplinary team care	-	+	+	+	-	-	-	3
Sullivan et al. (2002) [140]	Lower extremity function	BWS treadmill training (1 of 3 treadmill speeds)	-	+	-	+	-	-	-	2
Suwanwela et al. (2002) [141]	Models of care delivery	Short hospitalization with home care follow-up vs. conventional hospital stay	+	-	+	+	+	-	-	4
Teixeira da Cunha Filho et al. (2001) [142]	Lower extremity function	STAT vs. control (regular rehab only†)	-	-	+	-	-	-	-	1
Tekeoglu et al. (1998) [143]	General physical function	TENS vs. placebo TENS	-	+	+	-	-	-	-	2
Thorsen [144]	Models of care delivery	ESD and home rehab vs. conventional hospital rehab	-	-	+	+	-	-	-	2
Tibaek et al. (2007) [145]	Urinary incontinence	Pelvic floor muscle training vs. control (standard rehab only†)	+	-	+	+	-	-	-	3
Turton & Britton (2005) [146]	Upper extremity function	Stretch regime vs. control (usual care only†)	-	-	-	+	+	-	-	2
Volpe et al. (2000) [147]	Upper extremity function	Robotic training vs. sham therapy (exposure to device without training)	-	+	-	-	-	-	-	1
Wittenberg et al. (2003) [148]	Upper extremity function	Constraint-induced movement therapy vs. placebo-like intervention (therapy aimed at non-affected side)	-	+	+	-	-	-	-	2

Appendix F. Table of included studies

Study author and year of publication	General condition under investigation	Brief description of treatment and control condition(s)	Item number*							Total score
			1	2	3	4	5	6	7	
Wu et al. (2001) [149]	Upper extremity function	4 experimental conditions created by the crossing of functional goals (high vs. low) and personal preferences (high vs. low)	-	-	-	-	-	-	-	0
Yang et al. (2005) [150]	Lower extremity function	Additional backward walking vs. control (conventional training only†)	+	-	-	-	-	-	-	1
Yavuzer et al. (2007) [151]	Lower extremity function	Sensory-amplitude ES vs. sham therapy	-	+	-	+	-	-	-	2

* Item title
1: Populations in primary care
2: Less stringent eligibility criteria
3: Health outcomes
4: Long study duration, clinically relevant treatment modalities
5: Assessment of adverse events
6: Adequate sample size to assess a minimally important difference from a patient perspective
7: Intention-to-treat analysis

†The treatment group(s) received the control condition in addition to the experimental intervention.

Abbreviations:
BWS: body-weight supported;
EMG: electromyographic;
ES: electrical stimulation;
ESD: early-supported discharge;
FNS: functional neuromuscular stimulation;
OT: occupational therapy;
PT: physiotherapy;
rTMS: repetitive transcranial magnetic stimulation;
SPTP: social problem-solving telephone partnerships;
STAT: supported treadmill ambulation training;
TENS: transcutaneous electrical nerve stimulation

References

1 Almeida OP, Waterreus A, Hankey GJ: Preventing depression after stroke: Results from a randomized placebo-controlled trial. J Clin Psychiatry 2006;67:1104-1109.

2 Ashtary F, Janghorbani M, Chitsaz A, Reisi M, Bahrami A: A randomized, double-blind trial of bromocriptine efficacy in nonfluent aphasia after stroke. Neurology 2006;66:914-916.

3 Attal N, Gaude V, Brasseur L, Dupuy M, Guirimand F, Parker F, Bouhassira D: Intravenous lidocaine in central pain: a double-blind, placebo-controlled, psychophysical study. Neurology 2000;54:564-574.

4 Attal N, Guirimand F, Brasseur L, Gaude V, Chauvin M, Bouhassira D: Effects of IV morphine in central pain: a randomized placebo-controlled study. Neurology 2002;58:554-563.

5 Bakheit AM, Thilmann AF, Ward AB, Poewe W, Wissel J, Muller J, Benecke R, Collin C, Muller F, Ward CD, Neumann C: A randomized, double-blind, placebo-controlled, dose-ranging study to compare the efficacy and safety of three doses of botulinum toxin type A (Dysport) with placebo in upper limb spasticity after stroke. Stroke 2000;31:2402-2406.

6 Bakheit AM, Pittock S, Moore AP, Wurker M, Otto S, Erbguth F, Coxon L: A randomized, double-blind, placebo-controlled study of the efficacy and safety of botulinum toxin type A in upper limb spasticity in patients with stroke. Eur J Neurol 2001;8:559-565.

7 Bath PM, Lindenstrom E, Boysen G, De Deyn P, Friis P, Leys D, Marttila R, Olsson J, O'Neill D, Orgogozo J, Ringelstein B, van der SJ, Turpie AG: Tinzaparin in acute ischaemic stroke (TAIST): a randomised aspirin-controlled trial. Lancet 2001;358:702-710.

8 Berge E, Abdelnoor M, Nakstad PH, Sandset PM: Low molecular-weight heparin versus aspirin in patients with acute ischaemic stroke and atrial fibrillation: a double-blind randomised study. HAEST Study Group. Heparin in Acute Embolic Stroke Trial. Lancet 2000;355:1205-1210.

9 Berthier ML, Green C, Higueras C, Fernandez I, Hinojosa J, Martin MC: A randomized, placebo-controlled study of donepezil in poststroke aphasia. Neurology 2006;67:1687-1689.

10 Bhakta BB, Cozens JA, Chamberlain MA, Bamford JM: Impact of botulinum toxin type A on disability and carer burden due to arm spasticity after stroke: a randomised double blind placebo controlled trial. J Neurol Neurosurg Psychiatry 2000;69:217-221.

11 Black S, Roman GC, Geldmacher DS, Salloway S, Hecker J, Burns A, Perdomo C, Kumar D, Pratt R: Efficacy and tolerability of donepezil in vascular dementia: positive results of a 24-week, multicenter, international, randomized, placebo-controlled clinical trial. Stroke 2003;34:2323-2330.

12 Brashear A, Gordon MF, Elovic E, Kassicieh VD, Marciniak C, Do M, Lee CH, Jenkins S, Turkel C: Intramuscular injection of botulinum toxin for the treatment of wrist and finger spasticity after a stroke. N Engl J Med 2002;347:395-400.

13 Brashear A, McAfee AL, Kuhn ER, Fyffe J: Botulinum toxin type B in upper-limb poststroke spasticity: a double-blind, placebo-controlled trial. Arch Phys Med Rehabil 2004;85:705-709.

14 Brown KW, Sloan RL, Pentland B: Fluoxetine as a treatment for post-stroke emotionalism. Acta Psychiatr Scand 1998;98:455-458.

15 Burns A, Russell E, Stratton-Powell H, Tyrell P, O'Neill P, Baldwin R: Sertraline in stroke-associated lability of mood. Int J Geriatr Psychiatry 1999;14:681-685.

16 Chemerinski E, Robinson RG, Arndt S, Kosier JT: The effect of remission of poststroke depression on activities of daily living in a double-blind randomized treatment study. J Nerv Ment Dis 2001;189:421-425.

17 Childers MK, Brashear A, Jozefczyk P, Reding M, Alexander D, Good D, Walcott JM, Jenkins SW, Turkel C, Molloy PT: Dose-dependent response to intramuscular botulinum toxin type A for upper-limb spasticity in patients after a stroke. Arch Phys Med Rehabil 2004;85:1063-1069.

18 Choi-Kwon S, Han SW, Kwon SU, Kang DW, Choi JM, Kim JS: Fluoxetine treatment in poststroke depression, emotional incontinence, and anger proneness: a double-blind, placebo-controlled study. Stroke 2006;37:156-161.

19 Cohen RA, Browndyke JN, Moser DJ, Paul RH, Gordon N, Sweet L: Long-term citicoline (cytidine diphosphate choline) use in patients with vascular dementia: neuroimaging and neuropsychological outcomes. Cerebrovasc Dis 2003;16:199-204.

20 De Deyn PP, Reuck JD, Deberdt W, Vlietinck R, Orgogozo JM: Treatment of acute ischemic stroke with piracetam. Members of the Piracetam in Acute Stroke Study (PASS) Group. Stroke 1997;28:2347-2352.

21 Diener HC, Ringelstein EB, von Kummer R, Landgraf H, Koppenhagen K, Harenberg J, Rektor I, Csanyi A, Schneider D, Klingelhofer J, Brom J, Weidinger G: Prophylaxis of thrombotic and embolic events in acute ischemic stroke with the low-molecular-weight heparin certoparin: results of the PROTECT Trial. Stroke 2006;37:139-144.

22 Fogari R, Mugellini A, Zoppi A, Marasi G, Pasotti C, Poletti L, Rinaldi A, Preti P: Effects of valsartan compared with enalapril on blood pressure and cognitive function in elderly patients with essential hypertension. Eur J Clin Pharmacol 2004;59:863-868.

23 Forette F, Seux ML, Staessen JA, Thijs L, Babarskiene MR, Babeanu S, Bossini A, Fagard R, Gil-Extremera B, Laks T, Kobalava Z, Sarti C, Tuomilehto J, Vanhanen H, Webster J, Yodfat Y, Birkenhager WH: The prevention of dementia with antihypertensive treatment: new evidence from the Systolic Hypertension in Europe (Syst-Eur) study. Arch Intern Med 2002;162:2046-2052.

24 Francisco GE, Boake C, Vaughn A: Botulinum toxin in upper limb spasticity after acquired brain injury: a randomized trial comparing dilution techniques. Am J Phys Med Rehabil 2002;81:355-363.

25 Fruehwald S, Gatterbauer E, Rehak P, Baumhackl U: Early fluoxetine treatment of post-stroke depression--a three-month double-blind placebo-controlled study with an open-label long-term follow up. J Neurol 2003;250:347-351.

26 Gariballa SE, Parker SG, Taub N, Castleden CM: A randomized, controlled, a single-blind trial of nutritional supplementation after acute stroke. JPEN 1998;22:315-319.

27 Ginsberg JS, Bates SM, Oczkowski W, Booker N, Magier D, MacKinnon B, Weitz J, Kearon C, Cruickshank M, Julian JA, Gent M: Low-dose warfarin in rehabilitating stroke survivors. Thromb Res 2002;107:287-290.

28 Grade C, Redford B, Chrostowski J, Toussaint L, Blackwell B: Methylphenidate in early poststroke recovery: a double-blind, placebo-controlled study. Arch Phys Med Rehabil 1998;79:1047-1050.

29 Hillbom M, Erila T, Sotaniemi K, Tatlisumak T, Sarna S, Kaste M: Enoxaparin vs. heparin for prevention of deep-vein thrombosis in acute ischaemic stroke: a randomized, double-blind study. Acta Neurol Scand 2002;106:84-92.

30 Huber W, Willmes K, Poeck K, Van Vleymen B, Deberdt W: Piracetam as an adjuvant to language therapy for aphasia: a randomized double-blind placebo-controlled pilot study. Arch Phys Med Rehabil 1997;78:245-250.

31 Johnson CA, Wood DE, Swain ID, Tromans AM, Strike P, Burridge JH: A pilot study to investigate the combined use of botulinum neurotoxin type a and functional electrical stimulation, with physiotherapy, in the treatment of spastic dropped foot in subacute stroke. Artif Organs 2002;26:263-266.

32 Johnson CA, Burridge JH, Strike PW, Wood DE, Swain ID: The effect of combined use of botulinum toxin type A and functional electric stimulation in the treatment of spastic drop foot after stroke: a preliminary investigation. Arch Phys Med Rehabil 2004;85:902-909.

33 Jorge RE, Robinson RG, Arndt S, Starkstein S: Mortality and poststroke depression: A placebo-controlled trial of antidepressants. Am J Psychiatry 2003;160:1823-1829.

34 Kessler J, Thiel A, Karbe H, Heiss WD: Piracetam improves activated blood flow and facilitates rehabilitation of poststroke aphasic patients. Stroke 2000;31:2112-2116.

35 Kimura M, Robinson RG, Kosier JT: Treatment of cognitive impairment after poststroke depression : a double-blind treatment trial. Stroke 2000;31:1482-1486.

36 Kirazli Y, On AY, Kismali B, Aksit R: Comparison of phenol block and botulinus toxin type A in the treatment of spastic foot after stroke: a randomized, double-blind trial. Am J Phys Med Rehabil 1998;77:510-515.

37 Kong KH, Neo JJ, Chua KS: A randomized controlled study of botulinum toxin A in the treatment of hemiplegic shoulder pain associated with spasticity. Clin Rehabil 2007;21:28-35.

38 Lampl C, Yazdi K, Roper C: Amitriptyline in the prophylaxis of central poststroke pain. Preliminary results of 39 patients in a placebo-controlled, long-term study. Stroke 2002;33:3030-3032.

39 Laska AC, von Arbin M, Kahan T, Hellblom A, Murray V: Long-term antidepressant treatment with moclobemide for aphasia in acute stroke patients: a randomised, double-blind, placebo-controlled study. Cerebrovasc Dis 2005;19:125-132.

40 Lithell H, Hansson L, Skoog I, Elmfeldt D, Hofman A, Olofsson B, Trenkwalder P, Zanchetti A: The Study on Cognition and Prognosis in the Elderly (SCOPE): principal results of a randomized double-blind intervention trial. J Hypertens 2003;21:875-886.

41 Mancini F, Sandrini G, Moglia A, Nappi G, Pacchetti C: A randomised, double-blind, dose-ranging study to evaluate efficacy and safety of three doses of botulinum toxin type A (Botox) for the treatment of spastic foot. Neurol Sci 2005;26:26-31.

42 Marco E, Duarte E, Vila J, Tejero M, Guillen A, Boza R, Escalada F, Espadaler JM: Is botulinum toxin type A effective in the treatment of spastic shoulder pain in patients after stroke? A double-blind randomized clinical trial. J Rehabil Med 2007;39:440-447.

43 Martinsson L, Wahlgren NG: Safety of dexamphetamine in acute ischemic stroke: a randomized, double-blind, controlled dose-escalation trial. Stroke 2003;34:475-481.

44 Meythaler JM, Guin-Renfroe S, Brunner RC, Hadley MN: Intrathecal baclofen for spastic hypertonia from stroke. Stroke 2001;32:2099-2109.

45 Niedermaier N, Bohrer E, Schulte K, Schlattmann P, Heuser I: Prevention and treatment of poststroke depression with mirtazapine in patients with acute stroke. J Clin Psychiatry 2004;65:1619-1623.

46 On AY, Kirazli Y, Kismali B, Aksit R: Mechanisms of action of phenol block and botulinus toxin Type A in relieving spasticity: electrophysiologic investigation and follow-up. Am J Phys Med Rehabil 1999;78:344-349.

47 Palomaki H, Kaste M, Berg A, Lonnqvist R, Lonnqvist J, Lehtihalmes M, Hares J: Prevention of poststroke depression: 1 year randomised placebo controlled double blind trial of mianserin with 6 month follow up after therapy. J Neurol Neurosurg Psychiatry 1999;66:490-494.

48 Pantoni L, Bianchi C, Beneke M, Inzitari D, Wallin A, Erkinjuntti T: The Scandinavian Multi-Infarct Dementia Trial: a double-blind, placebo-controlled trial on nimodipine in multi-infarct dementia. J Neurol Sci 2000;175:116-123.

49 Perez I, Smithard DG, Davies H, Kalra L: Pharmacological treatment of dysphagia in stroke. Dysphagia 1998;13:12-16.

50 Pittock SJ, Moore AP, Hardiman O, Ehler E, Kovac M, Bojakowski J, Al K, I, Brozman M, Kanovs.ky P, Skorometz A, Slawek J, Reichel G, Stenner A, Timerbaeva S, Stelmasiak Z, Zifko UA, Bhakta B, Coxon E: A double-blind randomised placebo-controlled evaluation of three doses of botulinum toxin type A (Dysport) in the treatment of spastic equinovarus deformity after stroke. Cerebrovasc Dis 2003;15:289-300.

51 Poole KE, Loveridge N, Rose CM, Warburton EA, Reeve J: A single infusion of zoledronate prevents bone loss after stroke. Stroke 2007;38:1519-1525.

52 Rasmussen A, Lunde M, Poulsen DL, Sorensen K, Qvitzau S, Bech P: A double-blind, placebo-controlled study of sertraline in the prevention of depression in stroke patients. Psychosomatics 2003;44:216-221.

53 Reiter F, Danni M, Lagalla G, Ceravolo G, Provinciali L: Low-dose botulinum toxin with ankle taping for the treatment of spastic equinovarus foot after stroke. Arch Phys Med Rehabil 1998;79:532-535.

54 Robinson RG, Schultz SK, Castillo C, Kopel T, Kosier JT, Newman RM, Curdue K, Petracca G, Starkstein SE: Nortriptyline versus fluoxetine in the treatment of depression and in short-term recovery after stroke: a placebo-controlled, double-blind study. Am J Psychiatry 2000;157:351-359.

55 Rowbotham MC, Twilling L, Davies PS, Reisner L, Taylor K, Mohr D: Oral opioid therapy for chronic peripheral and central neuropathic pain. N Engl J Med 2003;348:1223-1232.

56 Sato Y, Maruoka H, Oizumi K: Amelioration of hemiplegia-associated osteopenia more than 4 years after stroke by 1 alpha-hydroxyvitamin D3 and calcium supplementation. Stroke 1997;28:736-739.

57 Sato Y, Kuno H, Kaji M, Saruwatari N, Oizumi K: Effect of ipriflavone on bone in elderly hemiplegic stroke patients with hypovitaminosis D. Am J Phys Med Rehabil 1999;78:457-463.

58 Sato Y, Asoh T, Kaji M, Oizumi K: Beneficial effect of intermittent cyclical etidronate therapy in hemiplegic patients following an acute stroke. J Bone Miner Res 2000;15:2487-2494.

59 Sato Y, Metoki N, Iwamoto J, Satoh K: Amelioration of osteoporosis and hypovitaminosis D by sunlight exposure in stroke patients. Neurology 2003;61:338-342.

60 Sato Y, Iwamoto J, Kanoko T, Satoh K: Risedronate sodium therapy for prevention of hip fracture in men 65 years or older after stroke. Arch Intern Med 2005;165:1743-1748.

61 Sato Y, Iwamoto J, Kanoko T, Satoh K: Risedronate therapy for prevention of hip fracture after stroke in elderly women. Neurology 2005;64:811-816.

62 Scheidtmann K, Fries W, Muller F, Koenig E: Effect of levodopa in combination with physiotherapy on functional motor recovery after stroke: a prospective, randomised, double-blind study. Lancet 2001;358:787-790.

63 Sherman DG, Albers GW, Bladin C, Fieschi C, Gabbai AA, Kase CS, O'Riordan W, Pineo GF: The efficacy and safety of enoxaparin versus unfractionated heparin for the prevention of venous thromboembolism after acute ischaemic stroke (PREVAIL Study): an open-label randomised comparison. Lancet 2007;369:1347-1355.

64 Smith SJ, Ellis E, White S, Moore AP: A double-blind placebo-controlled study of botulinum toxin in upper limb spasticity after stroke or head injury. Clin Rehabil 2000;14:5-13.

65 Snels IA, Beckerman H, Twisk JW, Dekker JH, Peter DK, Koppe PA, Lankhorst GJ, Bouter LM: Effect of triamcinolone acetonide injections on hemiplegic shoulder pain : A randomized clinical trial. Stroke 2000;31:2396-2401.

66 Sonde L, Nordstrom M, Nilsson CG, Lokk J, Viitanen M: A double-blind placebo-controlled study of the effects of amphetamine and physiotherapy after stroke. Cerebrovasc Dis 2001;12:253-257.

67 Sonde L, Lokk J: Effects of amphetamine and/or L-dopa and physiotherapy after stroke - a blinded randomized study. Acta Neurol Scand 2007;115:55-59.

68 Stamenova P, Koytchev R, Kuhn K, Hansen C, Horvath F, Ramm S, Pongratz D: A randomized, double-blind, placebo-controlled study of the efficacy and safety of tolperisone in spasticity following cerebral stroke. Eur J Neurol 2005;12:453-461.

69 Sze KH, Sim TC, Wong E, Cheng S, Woo J: Effect of nimodipine on memory after cerebral infarction. Acta Neurol Scand 1998;97:386-392.

70 Szelies B, Mielke R, Kessler J, Heiss WD: Restitution of alpha-topography by piracetam in post-stroke aphasia. Int J Clin Pharmacol Ther 2001;30:152-157.

71 Tanaka Y, Miyazaki M, Albert ML: Effects of increased cholinergic activity on naming in aphasia. Lancet 1997;350:116-117.

72 Tardy J, Pariente J, Leger A, Dechaumont-Palacin S, Gerdelat A, Guiraud V, Conchou F, Albucher JF, Marque P, Franceries X, Cognard C, Rascol O, Chollet F, Loubinoux I: Methylphenidate modulates cerebral post-stroke reorganization. Neuroimage 2006;33:913-922.

73 Low molecular weight heparinoid, ORG 10172 (danaparoid), and outcome after acute ischemic stroke: a randomized controlled trial. The Publications Committee for the Trial of ORG 10172 in Acute Stroke Treatment (TOAST) Investigators. JAMA 1998;279:1265-1272.

74 Vestergaard K, Andersen G, Gottrup H, Kristensen BT, Jensen TS: Lamotrigine for central poststroke pain: a randomized controlled trial. Neurology 2001;56:184-190.

75 Walker-Batson D, Curtis S, Natarajan R, Ford J, Dronkers N, Salmeron E, Lai J, Unwin DH: A double-blind, placebo-controlled study of the use of amphetamine in the treatment of aphasia. Stroke 2001;32:2093-2098.

76 Wiart L, Petit H, Joseph PA, Mazaux JM, Barat M: Fluoxetine in early poststroke depression: a double-blind placebo-controlled study. Stroke 2000;31:1829-1832.

77 Wilkinson D, Doody R, Helme R, Taubman K, Mintzer J, Kertesz A, Pratt RD: Donepezil in vascular dementia: a randomized, placebo-controlled study. Neurology 2003;61:479-486.

78 Yelnik AP, Colle FM, Bonan IV, Vicaut E: Treatment of shoulder pain in spastic hemiplegia by reducing spasticity of the subscapular muscle: A randomized, double-blind, placebo-controlled study of botulinum toxin A. J Neurol Neurosurg Psychiatry 2006.

79 Ada L, Goddard E, McCully J, Stavrinos T, Bampton J: Thirty minutes of positioning reduces the development of shoulder external rotation contracture after stroke: A randomized controlled trial. Arch Phys Med Rehabil 2005;86:230-234.

80 Alberts JL, Butler AJ, Wolf SL: The effects of constraint-induced therapy on precision grip: a preliminary study. Neurorehabil Neural Repair 2004;18:250-258.

81 Alon G, Levitt AF, McCarthy PA: Functional electrical stimulation enhancement of upper extremity functional recovery during stroke rehabilitation: a pilot study. Neurorehabil Neural Repair 2007;21:207-215.

82 Altschuler EL, Wisdom SB, Stone L, Foster C, Galasko D, Llewellyn DM, Ramachandran VS.: Rehabilitation of hemiparesis after stroke with a mirror. Lancet 1999;353:2035-2036.

83 Armagan O, Tascioglu F, Oner C: Electromyographic biofeedback in the treatment of the hemiplegic hand: a placebo-controlled study. Am J Phys Med Rehabil 2003;82:856-861.

84 Barrett JA, Watkins C, Plant R, Dickinson H, Clayton L, Sharma AK, Reston A, Gratton J, Fall S, Flynn A, Smith T, Leathley M, Smith S, Barer DH: The COSTAR wheelchair study: a two-centre pilot study of self-propulsion in a wheelchair in early stroke rehabilitation. Collaborative Stroke Audit and Research. Clin Rehabil 2001;15:32-41.

85 Bonan IV, Yelnik AP, Colle FM, Michaud C, Normand E, Panigot B, Roth P, Guichard JP, Vicaut E: Reliance on visual information after stroke. Part II: Effectiveness of a balance rehabilitation program with visual cue deprivation after stroke: a randomized controlled trial. Arch Phys Med Rehabil 2004;85:274-278.

86 Boter H, HESTIA Study Group: Multicenter randomized controlled trial of outreach nursing support program for recently discharged stroke patients. Stroke 2004;35:2867-2872.

87 Carnaby G, Hankey GJ, Pizzi J: Behavioural intervention for dysphagia in acute stroke: a randomised controlled trial. Lancet Neurol 2006;5:31-37.

88 Chen IC, Cheng PT, Chen CL, Chen SC, Chung CY, Yeh TH: Effects of balance training on hemiplegic stroke patients. Chang Gung Med J 2002;25:583-590.

89 Chen JC, Liang CC, Shaw FZ: Facilitation of sensory and motor recovery by thermal intervention for the hemiplegic upper limb in acute stroke patients: a single-blind randomized clinical trial. Stroke 2005;36:2665-2669.

90 Chen SC, Chen YL, Chen CJ, Lai CH, Chiang WH, Chen WL: Effects of surface electrical stimulation on the muscle-tendon junction of spastic gastrocnemius in stroke patients. Disabil Rehabil 2005;27:105-110.

91 Clark MS, Rubenach S, Winsor A: A randomized controlled trial of an education and counselling intervention for families after stroke. Clin Rehabil 2003;17:703-712.

92 da Cunha I, Jr., Lim PA, Qureshy H, Henson H, Monga T, Protas EJ: Gait outcomes after acute stroke rehabilitation with supported treadmill ambulation training: a randomized controlled pilot study. Arch Phys Med Rehabil 2002;83:1258-1265.

93 Daly JJ, Roenigk KL, Butler KM, Gansen JL, Fredrickson E, Marsolais EB, Rogers J, Ruff RL: Response of sagittal plane gait kinematics to weight-supported treadmill training and functional neuromuscular stimulation following stroke. J Rehabil Res Dev 2004;41:807-820.

94 de Jong LD, Nieuwboer A, Aufdemkampe G: Contracture preventive positioning of the hemiplegic arm in subacute stroke patients: a pilot randomized controlled trial. Clin Rehabil 2006;20:656-667.

95 de Kroon JR, IJzerman MJ, Lankhorst GJ, Zilvold G: Electrical stimulation of the upper limb in stroke: stimulation of the extensors of the hand vs. alternate stimulation of flexors and extensors. Am J Phys Med Rehabil 2004;83:592-600.

96 de Seze M, Wiart L, Bon SC, Debelleix X, Joseph PA, Mazaux JM, Barat M: Rehabilitation of postural disturbances of hemiplegic patients by using trunk control retraining during exploratory exercises. Arch Phys Med Rehabil 2001;82:793-800.

97 Dean CM, Shepherd RB: Task-related training improves performance of seated reaching tasks after stroke. A randomized controlled trial. Stroke 1997;28:722-728.

98 Desrosiers J, Bourbonnais D, Corriveau H, Gosselin S, Bravo G: Effectiveness of unilateral and symmetrical bilateral task training for arm during the subacute phase after stroke: a randomized controlled trial. Clin Rehabil 2005;19:581-593.

99 Doornhein K, De Haan EHF: Cognitive training for memorty deficits in stroke pateints. Neuropsychological Rehabilitation 1998;8:393-400.

100 Duncan P, Studenski S, Richards L, Gollub S, Lai SM, Reker D, Perera S, Yates J, Koch V, Rigler S, Johnson D: Randomized clinical trial of therapeutic exercise in subacute stroke. Stroke 2003;34:2173-2180.

101 Edmans JA, Webster J, Lincoln NB: A comparison of two approaches in the treatment of perceptual problems after stroke. Clin Rehabil 2000;14:230-243.

102 Ertelt D, Small S, Solodkin A, Dettmers C, McNamara A, Binkofski F, Buccino G: Action observation has a positive impact on rehabilitation of motor deficits after stroke. Neuroimage 2007;36(Suppl 2):T164-T173.

103 Fagerberg B, Claesson L, Gosman-Hedstrom G, Blomstrand C: Effect of acute stroke unit care integrated with care continuum versus conventional treatment: A randomized 1-year study of elderly patients: the Goteborg 70+ Stroke Study. Stroke 2000;31:2578-2584.

104 Fink M, Rollnik JD, Bijak M, Borstadt C, Dauper J, Guergueltcheva V, Dengler R, Karst M: Needle acupuncture in chronic poststroke leg spasticity. Arch Phys Med Rehabil 2004;85:667-672.

105 Goulding R, Bakheit AM: Evaluation of the benefits of monitoring fluid thickness in the dietary management of dysphagic stroke patients. Clin Rehabil 2000;14:119-124.

106 Grant JS, Elliott TR, Weaver M, Bartolucci AA, Giger JN: Telephone intervention with family caregivers of stroke survivors after rehabilitation. Stroke 2002;33:2060-2065.

107 Green J, Forster A, Bogle S, Young J: Physiotherapy for patients with mobility problems more than 1 year after stroke: a randomised controlled trial. Lancet 2002;359:199-203.

108 Hemmen B, Seelen HA: Effects of movement imagery and electromyography-triggered feedback on arm hand function in stroke patients in the subacute phase. Clin Rehabil 2007;21:587-594.

109 Howe TE, Taylor I, Finn P, Jones H: Lateral weight transference exercises following acute stroke: a preliminary study of clinical effectiveness. Clin Rehabil 2005;19:45-53.

110 Indredavik B, Fjaertoft H, Ekeberg G, Loge AD, Morch B: Benefit of an extended stroke unit service with early supported discharge: A randomized, controlled trial. Stroke 2000;31:2989-2994.

111 Indredavik B, Bakke F, Slordahl SA, Rokseth R, Haheim LL: Stroke unit treatment improves long-term quality of life: a randomized controlled trial. Stroke 1998;29:895-899.

112 Jorge RE, Robinson RG, Tateno A, Narushima K, Acion L, Moser D, Arndt S, Chemerinski E: Repetitive transcranial magnetic stimulation as treatment of poststroke depression: A preliminary study. Biol Psychiatry 2003;55:398-405.

113 Kalra L, Evans A, Perez I, Melbourn A, Patel A, Knapp M, Donaldson N: Training carers of stroke patients: randomised controlled trial. BMJ 2004;328:1099.

114 Khedr EM, Ahmed MA, Fathy N, Rothwell JC: Therapeutic trial of repetitive transcranial magnetic stimulation after acute ischemic stroke. Neurology 2005;65:466-468.

115 Kim CM, Eng JJ, MacIntyre DL, Dawson AS: Effects of isokinetic strength training on walking in persons with stroke: a double-blind controlled pilot study. J Stroke Cerebrovasc Dis 2001;10:265-273.

116 Kjendahl A, Sallstrom S, Osten PE, Stanghelle JK, Borchgrevink CF: A one year follow-up study on the effects of acupuncture in the treatment of stroke patients in the subacute stage: a randomized, controlled study. Clin Rehabil 1997;11:192-200.

117 Lam YS, Man DW, Tam SF, Weiss PL: Virtual reality training for stroke rehabilitation. NeuroRehabilitation 2006;21:245-253.

118 Langhammer B, Lindmark B, Stanghelle JK: Stroke patients and long-term training: is it worthwhile? A randomized comparison of two different training strategies after rehabilitation. Clin Rehabil 2007;21:495-510.

119 Lincoln NB, Parry RH, Vass CD: Randomized, controlled trial to evaluate increased intensity of physiotherapy treatment of arm function after stroke. Stroke 1999;30:573-579.

120 Lincoln NB, Husbands S, Trescoli C, Drummond AE, Gladman JR, Berman P: Five year follow up of a randomised controlled trial of a stroke rehabilitation unit. BMJ 2000;320:549.

121 Logan PA, Ahern J, Gladman JR, Lincoln NB: A randomized controlled trial of enhanced Social Service occupational therapy for stroke patients. Clin Rehabil 1997;11:107-113.

122 Lowe DB, Sharma AK, Leathley MJ: The CareFile Project: a feasibility study to examine the effects of an individualised information booklet on patients after stroke. Age Ageing 2007;36:83-89.

123 Macko RF, Ivey FM, Forrester LW, Hanley D, Sorkin JD, Katzel LI, Silver KH, Goldberg AP: Treadmill exercise rehabilitation improves ambulatory function and cardiovascular fitness in patients with chronic stroke: a randomized, controlled trial. Stroke 2005;36:2206-2211.

124 Mayo NE, Wood-Dauphinee S, Cote R, Gayton D, Carlton J, Buttery J, Tamblyn R: There's no place like home : an evaluation of early supported discharge for stroke. Stroke 2000;31:1016-1023.

125 McDowell BJ, Engberg S, Sereika S, Donovan N, Jubeck ME, Weber E, Engberg R: Effectiveness of behavioral therapy to treat incontinence in homebound older adults. J Am Geriatr Soc 1999;47:309-318.

126 Moon SK, Whang YK, Park SU, Ko CN, Kim YS, Bae HS, Cho KH: Antispastic effect of electroacupuncture and moxibustion in stroke patients. Am J Chin Med 2003;31:467-474.

127 Moseley GL: Graded motor imagery is effective for long-standing complex regional pain syndrome: a randomised controlled trial. Pain 2004;108:192-198.

128 Ouellette MM, LeBrasseur NK, Bean JF, Phillips E, Stein J, Frontera WR, Fielding RA: High-intensity resistance training improves muscle strength, self-reported function, and disability in long-term stroke survivors. Stroke 2004;35:1404-1409.

129 Page SJ, Sisto S, Levine P, McGrath RE: Efficacy of modified constraint-induced movement therapy in chronic stroke: A single-blinded randomized controlled trial. Arch Phys Med Rehabil 2004;85:14-18.

130 Pang MY, Eng JJ, Dawson AS, McKay HA, Harris JE: A community-based fitness and mobility exercise program for older adults with chronic stroke: a randomized, controlled trial. J Am Geriatr Soc 2005;53:1667-1674.

131 Pang MY, Harris JE, Eng JJ: A community-based upper-extremity group exercise program improves motor function and performance of functional activities in chronic stroke: a randomized controlled trial. Arch Phys Med Rehabil 2006;87:1-9.

132 Parker CJ, Gladman JR, Drummond AE, Dewey ME, Lincoln NB, Barer D, Logan PA, Radford KA: A multicentre randomized controlled trial of leisure therapy and conventional occupational therapy after stroke. TOTAL Study Group. Trial of Occupational Therapy and Leisure. Clin Rehabil 2001;15:42-52.

133 Peurala SH, Tarkka IM, Pitkanen K, Sivenius J: The effectiveness of body weight-supported gait training and floor walking in patients with chronic stroke. Arch Phys Med Rehabil 2005;86:1557-1564.

134 Ploughman M, Corbett D: Can forced-use therapy be clinically applied after stroke? an exploratory randomized controlled trial. Arch Phys Med Rehabil 2004;85:1417-1423.

135 Ring H, Rosenthal N: Controlled study of neuroprosthetic functional electrical stimulation in sub-acute post-stroke rehabilitation. J Rehabil Med 2005;37:32-36.

136 Rodgers H, Soutter J, Kaiser W, Pearson P, Dobson R, Skilbeck C, Bond J: Early supported hospital discharge following acute stroke: pilot study results. Clin Rehabil 1997;11:280-287.

137 Rodgers H, Atkinson C, Bond S, Suddes M, Dobson R, Curless R: Randomized controlled trial of a comprehensive stroke education program for patients and caregivers. Stroke 1999;30:2585-2591.

138 Salbach NM, Mayo NE, Wood-Dauphinee S, Hanley JA, Richards CL, Cote R: A task-orientated intervention enhances walking distance and speed in the first year post stroke: a randomized controlled trial. Clin Rehabil 2004;18:509-519.

139 Sulch D, Melbourn A, Perez I, Kalra L: Integrated care pathways and quality of life on a stroke rehabilitation unit. Stroke 2002;33:1600-1604.

140 Sullivan KJ, Knowlton BJ, Dobkin BH: Step training with body weight support: effect of treadmill speed and practice paradigms on poststroke locomotor recovery. Arch Phys Med Rehabil 2002;83:683-691.

141 Suwanwela NC, Phanthumchinda K, Limtongkul S, Suvanprakorn P: Comparison of short (3-day) hospitalization followed by home care treatment and conventional (10-day) hospitalization for acute ischemic stroke. Cerebrovasc Dis 2002;13:267-271.

142 Teixeira da Cunha Filho I, Lim PA, Qureshy H, Henson H, Monga T, Protas EJ: A comparison of regular rehabilitation and regular rehabilitation with supported treadmill ambulation training for acute stroke patients. J Rehabil Res Dev 2001;38:245-255.

143 Tekeoglu Y, Adak B, Goksoy T: Effect of transcutaneous electrical nerve stimulation (TENS) on Barthel Activities of Daily Living (ADL) index score following stroke. Clin Rehabil 1998;12:277-280.

144 Thorsen AM, Holmqvist LW, Pedro-Cuesta J, von Koch L: A randomized controlled trial of early supported discharge and continued rehabilitation at home after stroke: five-year follow-up of patient outcome. Stroke 2005;36:297-303.

145 Tibaek S, Gard G, Jensen R: Is there a long-lasting effect of pelvic floor muscle training in women with urinary incontinence after ischemic stroke? : A 6-month follow-up study. Int Urogynecol J 2007;18:281-287.

146 Turton AJ, Britton E: A pilot randomized controlled trial of a daily muscle stretch regime to prevent contractures in the arm after stroke. Clin Rehabil 2005;19:600-612.

147 Volpe BT, Krebs HI, Hogan N, Edelstein OTR, Diels C, Aisen M: A novel approach to stroke rehabilitation: robot-aided sensorimotor stimulation. Neurology 2000;54:1938-1944.

148 Wittenberg GF, Chen R, Ishii K, Bushara KO, Eckloff S, Croarkin E, Taub E, Gerber LH, Hallett M, Cohen LG: Constraint-induced therapy in stroke: magnetic-stimulation motor maps and cerebral activation. Neurorehabil Neural Repair 2003;17:48-57.

149 Wu CY, Wong MK, Lin KC, Chen HC: Effects of task goal and personal preference on seated reaching kinematics after stroke. Stroke 2001;32:70-76.

150 Yang YR, Yen JG, Wang RY, Yen LL, Lieu FK: Gait outcomes after additional backward walking training in patients with stroke: a randomized controlled trial. Clin Rehabil 2005;19:264-273.

151 Yavuzer G, Oken O, Atay MB, Stam HJ: Effect of sensory-amplitude electric stimulation on motor recovery and gait kinematics after stroke: a randomized controlled study. Arch Phys Med Rehabil 2007;88:710-714.