A Metaevaluation of Evaluations of Health Care Programs that Employ the Chronic Care Model

Jan Fields

Western Michigan University, jankevinfields@gmail.com

Follow this and additional works at: https://scholarworks.wmich.edu/dissertations

Part of the Applied Statistics Commons, and the Health Services Research Commons

Recommended Citation
Fields, Jan, "A Metaevaluation of Evaluations of Health Care Programs that Employ the Chronic Care Model" (2014). Dissertations. 241.
https://scholarworks.wmich.edu/dissertations/241
A META-EVALUATION OF EVALUATIONS OF HEALTH CARE PROGRAMS
THAT EMPLOY THE CHRONIC CARE MODEL

by

Jan Fields

A dissertation submitted to the Graduate College
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
Interdisciplinary Ph.D. in Evaluation
Western Michigan University
April 2014

Doctoral Committee:

Chris Coryn, Ph.D., Chair
Daniela Schröter, Ph.D.
Mark Ivey, M.D.
Background: The purpose of this dissertation is to explore the use of metaevaluation to evaluate the quality of healthcare studies conducted on programs that employ the Chronic Care Model (CCM) to provide chronic illness care. In this study, healthcare studies of CCM programs are regarded as program evaluations. Method: Using a non-experimental cross-sectional design, 28 healthcare studies of CCM programs were evaluated using the accuracy standards portion of the Program Evaluations Metaevaluation Checklist (Stufflebeam, 2011). The results of the metaevaluations were analyzed and compared to the HEAL grade of the same healthcare studies as determined by the Hierarchy of Evidence and Appraisal of Limitations grading system (Gugiu, et al, 2013). Results: When the quality of the 28 healthcare studies were evaluated by metaevaluation, only five (5) studies rated Good. The rest of the studies are rated either Fair (20) or Poor (3). When the quality of the same healthcare studies were assessed by the HEAL grading system, only four (4) studies received an A or A- and only four (4) studies received a B or B-. The rest of the studies received either a C- (2) or a D (18). Also, none of the metaevaluation ratings of the 28 healthcare studies showed a significant relationship with the HEAL scores of the same healthcare studies. Conclusions: Whether rated by the accuracy metaevaluation checklist or graded by the HEAL grading system, the selected CCM studies were mostly of poor to fair quality. This finding raises the level of urgency for improving the study quality of evidence-based medicine. Also, the lack of significant relationship between the two measures of study quality may indicate that the two instruments are complementary rather than substitutionary. If so, it may be better to use both the metaevaluation ratings and the HEAL scores to determine study quality rather than using one instrument over the other.
ACKNOWLEDGEMENTS

“Brevity is the soul of wit.” — William Shakespeare, Hamlet

In keeping with this quote, I will keep these acknowledgements brief.

It took me almost two years to get a proposal approved primarily because I had a topic for a dissertation stuck in my head and couldn’t get it out no matter how many times my committee told me that it simply wasn’t an evaluation dissertation topic. Finally, Chris Coryn Ph.D., my dissertation committee chair, rescued me by giving me the idea of looking into the use of metaevaluation as a means of assessing healthcare study quality. The dissertation process went fairly smoothly from that point on. Thanks, Chris!

Chris and Daniela Schroter Ph.D. were my “What’s so?” advisors on my committee, while Mark Ivey M.D, the medical director of the respiratory therapy program where I taught for 22 years, was my “So what?” advisor. They worked together well in those capacities. Daniela was especially helpful with her reviews of my drafts. She cautioned me that she had a very German approach to her critiques, but I found them very helpful and to the point.

Pedro Mateu was the other rater in my study besides me. You cannot find a man with greater integrity than Pedro. He is the ideal rater. Because of his integrity and humility, we worked well together which accounted in large part for the high inter-rater reliability between our metaevaluation ratings.

Of course, my wife deserves high praise for tolerating my long hours in my home office without any physical output to show for it. Many a home project was neglected in my quest for a Ph.D! She is just as glad to have me done as I am. It was especially gratifying to have her present at my dissertation defense.

As I enter my 60th year this month, I hope and pray that I can still contribute in some small way to the improvement of our society through this marvelous vocation we call interdisciplinary evaluation.

Jan Fields
# TABLE OF CONTENTS

ACKNOWLEDGMENTS ................................................................................................................................. ii

LIST OF TABLES ........................................................................................................................................... v

LIST OF FIGURES ......................................................................................................................................... vi

CHAPTER

I. INTRODUCTION ............................................................................................................................................ 1
   Background of the Problem .................................................................................................................... 1
   Purpose of the Study ........................................................................................................................... 3
   Key Terms ........................................................................................................................................... 4
   Research Questions ............................................................................................................................. 6
   Theoretical Framework ....................................................................................................................... 6
   Need for the Study ............................................................................................................................... 8

II. REVIEW OF LITERATURE .......................................................................................................................... 10
   Study Quality ........................................................................................................................................ 10
      Research and Evaluation Methodology ......................................................................................... 10
      Study Quality of Research and Evaluation .................................................................................. 15
   Chronic Care Model ............................................................................................................................ 37
      Development of the Chronic Care Model ....................................................................................... 37
      Application of the Chronic Care Model ......................................................................................... 44
   Implementation Science ...................................................................................................................... 52

III. METHODOLOGY ....................................................................................................................................... 60
   Research Design .................................................................................................................................... 60
   Selection of Studies ............................................................................................................................ 60
   Instrumentation ..................................................................................................................................... 61
   Data Processing and Analysis ............................................................................................................ 63
   Research Question #1 ......................................................................................................................... 63
   Research Question #2 ......................................................................................................................... 63
Table of Contents - Continued

IV. RESULTS ............................................................................................................................................... 65
    Results Pertaining to Research Question #1 .................................................................................. 65
    Results Pertaining to Research Question #2 .............................................................................. 76
    Discussion of Results ................................................................................................................... 78

V. CONCLUSIONS .................................................................................................................................... 89
    Summary of Dissertation Findings ............................................................................................... 89
    Limitations to Dissertation Findings ............................................................................................ 90
    Implications for Implementation Science ..................................................................................... 90
    Contribution to Evaluation Practice ............................................................................................. 92
    Need for Further Study .................................................................................................................. 94

STUDIES REVIEWED ........................................................................................................................... 96

REFERENCES .......................................................................................................................................... 99

APPENDICES .......................................................................................................................................... 109
    A. Hierarchy of Evidence and Appraisal of Limitations (HEAL) Grading System .................. 109
    B. Human Subjects Institutional Review Board Letter ............................................................ 111
LIST OF TABLES

1. HEAL Scores for Selected CCM Studies................................................................. 18
2. Proposals Creating a High-Performing Chronic Care System.................................... 51
3. Evaluative Questions for the RE-AIM Framework.................................................. 56
4. Exclusion Criteria for Relevant CCM Studies......................................................... 61
5. Breakdown of CCM Studies by Research Design.................................................. 61
6. Accuracy Standard Ratings of Selected CCM Studies........................................... 66
7. Correlations between Metaevaluation Ratings and HEAL Grades for Selected CCM Studies...... 77
8. Summary of A1 Checkpoints for Selected CCM Studies........................................ 78
9. Summary of A2 Checkpoints for Selected CCM Studies........................................ 80
10. Summary of A3 Checkpoints for Selected CCM Studies......................................... 81
11. Summary of A4 Checkpoints for Selected CCM Studies......................................... 82
12. Summary of A5 Checkpoints for Selected CCM Studies......................................... 83
13. Summary of A6 Checkpoints for Selected CCM Studies......................................... 85
14. Summary of A7 Checkpoints for Selected CCM Studies......................................... 86
15. Summary of A8 Checkpoints for Selected CCM Studies......................................... 87
16. Comparisons of Metaevaluation Ratings and HEAL Grades for Selected CCM Studies........... 88
LIST OF FIGURES

1. Overview of the Chronic Care Model........................................................................................................... 7

2. Overview of the Innovative Care for Chronic Conditions (ICCC) framework........................................ 48

3. The Pathway of Harnessing Science to Promote Health............................................................................. 53

4. Example of an Accuracy Standard in the Program Evaluations Metaevaluation Checklist.................. 62

5. Analysis Process of the Total Score for Accuracy Standards................................................................. 62
CHAPTER I

INTRODUCTION

Background of the Problem

Chronic illnesses – such as heart disease, stroke, cancer, diabetes, and arthritis – are the most costly health problems in the United States. Almost 70% of deaths among Americans each year are from chronic diseases (Center for Disease Control, 2007). Heart disease, cancer and stroke account for more than 50% of all deaths each year (Kung, Hoyert, Xu, & Murphy, 2008). In 2005, 133 million Americans had at least one chronic illness (CDC, 2007). By 2020, 157 million American citizens are predicted to have multiple chronic disorders (Lancet, 2009). The total economic impact of chronic illness in the U.S. is $1.3 trillion annually. Of this amount, lost productivity totals $1.1 trillion per year, while another $277 billion is spent annually on treatment, not including costs to treat the consequences of these diseases. The total economic impact is projected to increase to $4.2 trillion by 2023 (DeVol & Bedroussian, 2007).

According to the Committee on the Quality of Health Care in America (2001), the American healthcare delivery system is in need of fundamental change. The Committee has issued a call for systemic change including a redesign of care processes based on best practices, the use of information technologies to improve access to clinical information and support clinical decision making, an enhancement of patient knowledge and skill management, the development of effective practice teams, the coordination of care across services and settings over time, and the incorporation of performance and outcome measurements for improvement and accountability (CQHCA, 2001).

A variety of changes for the management of chronic illness care have been advocated (Bodenheimer, Chen, & Bennett, 2009). Renders et al. (2001) concluded after a Cochrane review that the most effective approach to chronic disease care is a multi-pronged strategy of systemic change. The Chronic Care Model (CCM) is an example of such an approach. The CCM has been used in the design of many health care programs targeting chronic illness over the past ten years.
There have been several published systematic reviews and meta-analyses of the efficacy of the CCM (Adams et al., 2007; Bodenheimer, 2003; Bonomi, Wagner, Glasgow, & VonKorff, 2002; Coleman, Austin, Brach, & Wagner, 2009; Minkman, Ahaus, & Huijsman, 2007; Scott, 2008; Si & Bailie, 2008; Tsai, Morton, Mangione, & Keeler, 2005), but only half of these studies included an assessment of study quality (Gugiu & Gugiu, 2010). Those studies that did assess study quality used grading methods such as the United States Preventative Task Force criteria (Agency for Healthcare Research and Quality, 2008) and the Jadad Scale (Jadad et al., 1996). However, these grading methods have serious limitations that have recently been highlighted (Alperson & Berger, 2011; Berger & Alperson, 2009; Gugiu & Gugiu, 2010).

These issues form the background of the problem being addressed by this dissertation. Chronic illness in the United States is widespread, expensive, and deadly. Our current healthcare system is not designed to effectively provide chronic illness care. The adoption of a multi-pronged strategy of systemic change like the CCM should provide effective chronic illness care, but most evidence regarding its use is inconclusive because of low quality studies (Gugiu, Westine, Coryn, & Hobson, 2013). The most fundamental problem, then, is not the prevalence of chronic illness or the inadequacy of chronic illness care, but rather our inability and/or unwillingness to conduct high quality healthcare studies of programs created to address chronic illnesses. This is the problem being addressed by this dissertation.

As mentioned above, grading systems have been created to assess healthcare study quality. Currently, the most popular grading system for assessing the quality of healthcare studies is the GRADE (Grades of Recommendation, Assessment, Development, and Evaluation) method (Guyatt, Oxman, Schünemann, Tugwell, & Knottnerus, 2011). The GRADE method includes an assessment of risk of bias using the Cochrane risk of bias tool.

A new grading system, the HEAL (Hierarchy of Evidence and Appraisal of Limitations) method, builds on the strengths of the GRADE method by systematically identifying additional important threats to validity and more clearly distinguishing between strong and weak non-randomized controlled trials (Gugiu & Gugiu, 2010). Gugiu et al. (2013) demonstrated the HEAL grading system by assessing the quality of recent studies of CCM programs. The study’s authors concluded that most of the selected CCM studies
were poorly designed and/or poorly executed. They also found that the HEAL grading system appeared to be an improvement on the GRADE method (Gugiu et al., 2013). Still, while the HEAL grading system may be an improvement on other study quality grading systems, it may still be inadequate to the task of addressing the problem being addressed by this dissertation; that is, our inability and/or unwillingness to conduct high quality healthcare studies of programs created to address chronic illnesses.

McInerny (2005) states that the first step in preparing the mind for logical thinking, for sound reasoning and deep understanding, is to pay attention. Miscues and misunderstandings are often the result of not paying sufficient attention. This is especially the case when the situation being studied is familiar. Because it is familiar, we regard it as a repeat performance. “But, in the strictest sense, there are no repeat performances. Every situation is unique, and we must be alert to its uniqueness” (McInerny, 2005, p. 3). Perhaps what is needed in healthcare research is a study methodology that pays closer attention to the interface between provider and patient. If so, what is also needed is an assessment or grading system of study quality that encourages and holds accountable such a study methodology.

Purpose of the Study

The purpose of this dissertation is to explore the use of metaevaluation to assess the quality of healthcare studies. In this dissertation, recent studies of healthcare programs designed to address chronic illness will be regarded as evaluations. When regarded as evaluations, the use of metaevaluation, which is an evaluation of an evaluation, to assess the quality of healthcare studies seems appropriate. Because a metaevaluation holds an evaluation to very high standards (Yarbrough, Shulha, Hopson, & Caruthers, 2011), it may be a useful adjunct to a grading system that encourages and holds accountable healthcare studies to the highest possible quality. If so, metaevaluation could be a key component in the promotion and sustainment of the evidence-based movement and a powerful weapon in the effort to minimize the growing social and economic consequences of chronic illnesses.
Key Terms

Study Quality. Study quality is the fit between a study’s goals and that study’s design and implementation characteristics (Valentine, 2009). Enhancing study quality is and will continue to be critical for transitioning evidence-based medicine to everyday clinical policy and practice (Gugiu & Gugiu, 2010). The criteria used to determine the merit of a study is discussed in Chapter 2 of this report. It is important at this point to distinguish between healthcare quality and healthcare study quality. As stated earlier, chronic illness in the United States is widespread, expensive, and deadly because the quality of healthcare for patients with chronic illness is poor. It is the premise of this dissertation that the quality of healthcare for patients with chronic illness is poor in large part because the quality of healthcare studies is poor. For the purposes of this dissertation, a further distinction between research study quality and evaluation study quality needs to be made. The distinction lies in the difference between research and evaluation.

Research vs. Evaluation. Research is scientific inquiry born of intellectual curiosity for the purpose of producing generalizable knowledge. This inquiry can be conducted using either quantitative or qualitative methodology (Trochim, 2006). Evaluation is the systematic assessment of the value of some object, be it a program, product, policy, or personnel. There is a difference between research and evaluation but the two are interconnected. Doing evaluation requires research. But, evaluation requires more than the collection and analysis of data. It also requires the synthesis of such data based on pre-determined criteria and standards to determine value (Mathison, 2005, p. 139). Because evaluation is designed specifically to determine value, evaluation may be a better means for judging healthcare quality, which is a value, than research alone.

A distinguishing feature of evaluation is the focus on stakeholder perspectives (Mathison, 2005, p. 397). Evaluation is different from research in that evaluators seek out value judgments from stakeholders while researchers avoid making value judgments. It is a fairly recent understanding in healthcare that context significantly influences the success of a particular intervention. Much of what constitutes the context of a program or policy is found in the perspectives of its stakeholders (Mathison,
2005, p. 398). Because evaluation seems well suited to assessing context in healthcare, this is another reason why evaluation may be a better means for judging healthcare quality than research alone.

If, indeed, evaluation is superior to research alone as a means for judging healthcare quality, then assessing evaluation study quality may be a more effective approach towards improving healthcare quality than assessing research study quality. This dissertation explores this possibility by using metaevaluation (a means for assessing evaluation study quality) to assess the quality of selected CCM studies and then comparing this approach to the HEAL grading system (a means for assessing research study quality).

**Metaevaluation.** According to Cooksy and Caracelli (2005), “Metaevaluations are systematic reviews of evaluations to determine the quality of their processes and findings (p. 2).” Michael Scriven introduced the term metaevaluation in 1969 (Cooksy & Caracelli, 2005). He used this label to refer to his evaluation of a plan for evaluating educational products. Operationally, metaevaluation is defined as the process of delineating, obtaining, and applying descriptive information and judgmental information—about the utility, feasibility, propriety, accuracy, and evaluation accountability of an evaluation (Stufflebeam, 2001; Yarbrough et al., 2011) and its systematic nature, competent conduct, integrity/honesty, respectfulness, and social responsibility (American Evaluation Association, 2009)—to guide the evaluation and/or report its strengths and weaknesses. According to Mathison (2005, p. 249), a metaevaluation evaluates evaluations; it does not merely summarize them.

To facilitate a metaevaluation, Stufflebeam (2011b) created a checklist based on the program evaluation standards mentioned above. Utility standards ensure that the evaluation serves the needs of the end users. Feasibility standards ensure that the evaluation is realistic and practical. Propriety standards ensure that the evaluation was conducted legally and ethically. Accuracy standards ensure that the evaluation provides technically adequate findings and conclusions about the program. Evaluation accountability standards ensure that evaluation quality is assured and documented (Yarbrough et al., 2011). In this dissertation, only the accuracy standards the metaevaluation checklist is used. The problem
being addressed by the dissertation is the existence of poor quality study in the healthcare field. Of the five categories of standards, only the accuracy standards directly address study quality.

Research Questions

The following research questions are addressed in this dissertation.

Research Question #1

What is the accuracy metaevaluation rating (Stufflebeam, 2011a) for each CCM study selected for this dissertation?

Research Question #2

What is the relationship between the accuracy metaevaluation rating (Stufflebeam, 2011a) and the HEAL grade (Gugiu et al., 2013) of each CCM study selected for this dissertation?

Theoretical Framework

Chronic Care Model. Chronic illness care takes place within three settings: (1) the community, with its resources and policies; (2) the health care system, including its payment structures; and (3) the provider organization, whether an integrated delivery system, a small clinic, or a loose network of physician practices (Bodenheimer, Wagner, & Grumbach, 2002a). The Chronic Care Model (CCM) consists of essential elements that encourage high-quality chronic illness care. The community-wide elements are the overall healthcare delivery system and the safety net organizations in the community. The provider-specific elements are the delivery system design of the practice, a clinical information system, and a support process for both practitioner decision-making and patient self-management (see Figure 1). These community-wide elements and provider-specific elements, in combination, may foster productive interactions between a prepared, proactive team of practitioners and informed, activated patients with chronic illness in such a way that the social and economic consequences of chronic illness may be significantly minimized (Bodenheimer, 2003).
Each of the 28 CCM programs discussed in this dissertation is based on a theoretical framework by which practitioners potentially become a prepared, proactive team of practitioners, and patients with chronic illness potentially become informed, activated patients with chronic illness. The six elements of the CCM mentioned above comprise this framework by which these selected programs effect change. A description of each CCM element is given below (Bodenheimer, Wagner, & Grumbach, 2002b).

1. **Patient self-management** support provides patient self-management training (cognitive and psychomotor), assesses a patient’s readiness to change, promotes healthy lifestyle choices, encourages collaboration to make decisions and set goals, establishes action plans, identifies barriers to healthy behaviors, and provides access to non-electronic health libraries.

2. **Delivery system design** creates multi-disciplinary clinical teams that delegate care responsibilities from the provider to other team members; assigns a care manager; facilitates interactions between the patient and multiple providers in a single visit; performs a risk stratification to guide resource utilization, follows up with patients regarding symptoms, behavioral change, or use of medication; and changes medication regimens when necessary.

3. **Decision support** improves medical decision making and patient care through evidence-based guidelines which provide medical education and increase interactions between generalists and
specialists. It also involves the use of case reviews to improve diagnosis and treatment, training providers, and develop patient assessment tools.

4. *Clinical information systems* use patient registries and electronic medical records to integrate and reinforce the first three CCM components by generating reports that facilitate communication between patients and their care team. Improved communication leads to improved adherence and access to medical information for the patient and more informed decision-making for the practitioner.

5. *Community resources* link patients to local resources by developing and maintaining a list of community resources, encouraging organization leaders to connect with other organizations to expand services and enhance continuity of care.

6. *Health care organization* provides the necessary structure for organization leaders to support and promote the CCM by developing strategy, accountability, goals, resources, evaluation, and financial and nonfinancial incentives to care providers.

Need for the Study

In 2001, the landmark report, “Crossing the Quality Chasm: A New Health System for the 21st Century,” was published (CQHCA, 2001). This report by the Committee on the Quality of Health Care in America concluded that, “Between the health care we have and the health care we could have lays not just a gap, but a chasm” (CQHCA, 2001, p. 1). This discovery-delivery chasm in healthcare is due in part to an inability and/or unwillingness to adequately address chronic illness care. Changes in public health policy are needed to help health care decision makers and health care professionals adopt the CCM and bring about the paradigm shift needed to fully meet the needs of patients with chronic illness (Bodenheimer et al., 2002b).

This paradigm shift will occur only when we begin to “understand the behavior of healthcare professionals and other stakeholders as a key variable in the sustainable uptake, adoption, and implementation of evidence-based interventions” (Madon, Hofman, Jupfer, & Glass, 2007, p. 1728).
Evaluation seems well suited to play this role in the evidence-based movement. If so, metaevaluation might be one of the best means of attaining and sustaining high-quality healthcare studies.
CHAPTER II

REVIEW OF THE LITERATURE

As stated in the previous chapter, enhancing study quality is and will continue to be critical for transitioning evidence-based medicine to everyday clinical policy and practice. The purpose of this dissertation is to explore the use of metaevaluation to assess the study quality of selected CCM studies. In this chapter, the foundations of study quality and current methods for assessing study quality will be discussed. Study quality in the context of the chronic care model (CCM) will also be presented in this chapter. Finally, implementation science as it relates to evidence-based medicine will be discussed and the role of evaluation in implementation science will be explored.

Study Quality

Research and Evaluation Methodology

Social inquiry. To more fully understand the concept of study quality, it may be helpful to review the foundation and evolution of the social inquiry discipline. At the very foundation of all systematic inquiry is epistemology. Epistemology is the theory of knowledge that defines what kind of knowledge is possible and legitimate (Feast & Melles, 2010). The major divisions of epistemology are objectivism, constructivism, and subjectivism. Methodology is the bridge that brings epistemology and social research methods together (Staller, Block, & Horner, 2008). Generally, two types of methodology are available to researchers engaged in social inquiry: quantitative methodology and qualitative methodology. Social research methods are approaches to answering research questions that may advance our understanding of the social world or some aspect of it (Staller et al., 2008). Social inquiry is the systematic study of group behavior in the social world, using a variety of social research methods (Christie & Fleischer, 2009).

History of research and evaluation methodology. Initially, social inquiries were a product of contemplation rather than empirical investigations. Not till the late 19th century did social inquiry involve the collection and analysis of empirical data (Christie & Fleischer, 2009).
The discipline of psychology introduced the experimental method to the realm of social inquiry (Mark, Donaldson, & Campbell, 2011). The impact of psychology and social experimentation are central to arguments in support of quantitative research methodology. At the basis of quantitative research methodology is the post-positivist perspective that there is a single reality that can be studied objectively, though there is no way to understand this reality in its totality. Scientists holding to this perspective believe that causation is observable and that over time predictors can be established, but also that there will always be some degree of doubt associated with the conclusion (Christie & Fleischer, 2009).

The discipline of anthropology introduced qualitative research methods to the realm of social inquiry (Mark et al., 2011). Contributions of anthropology include the use of “thick description” and observation as valid social research methods. At the basis of qualitative research methodology is the constructivist perspective that there are multiple realities which are subjective and change according to the “knower.” Scientists holding to this perspective believe that cause and effect cannot be known because relationships are bidirectional, thus everything impacts everything else (Christie & Fleischer, 2009).

Social psychology is social inquiry that considers how people's thoughts, feelings, and behaviors are influenced by the actual, imagined, or implied presence of others (Mark et al., 2011). It can involve the use of both quantitative and qualitative research methods. Kurt Lewin is considered by many to be the founder of modern social psychology. Among other contributions, Lewin is credited with initiating the study of group dynamics and promoting the integration of basic and applied research. Lewin’s contributions to the field of evaluation are not as widely recognized as his contributions to social psychology, but they are important. Perhaps the contribution to evaluation which Lewin is most widely credited is action research (Craig, 2009).

Action research is the process whereby practitioner-researchers attempt to study problems in order to guide, correct, and evaluate their decisions and actions (Craig, 2009). Action research also makes use of both quantitative and qualitative social research methods. In terms of evaluation, action research is sometimes used by those interested in improving programs. Several evaluation theories have been
influenced by Lewin through the principles of action research including empowerment evaluation, practical participatory evaluation, and developmental evaluation (Mark et al., 2011).

In the 1960s and 1970s, many social psychologists and other social scientists moved into evaluation, especially into the field of program evaluation (Mark et al., 2011). By the early 1960s, there had been massive US government expenditures (Shadish & Luellen, 2005). A multitude of relief agencies were legislated under FDR’s New Deal during the 1930s. Later, numerous social programs were initiated under the leadership of John F. Kennedy, with the goal of protecting national prosperity and providing for the well-being of American citizens. Lyndon B. Johnson followed through on these initiatives and expanded social programs to conduct what he called a War on Poverty in his quest for a Great Society (Shadish & Luellen, 2005).

Social programs in health, education, and housing grew and eventually accounted for billions of dollars in government spending (Shadish & Luellen, 2005). This created an increased demand for program evaluations which opened up new employment pathways for growing numbers of PhDs in social psychology (Mark et al., 2011). This in turn promoted the further development of social research and evaluation methodology.

In 1969, Donald T. Campbell, a social psychologist at Northwestern University, set a vision for the newly forming applied research and evaluation community which he called the Experimenting Society (Campbell, 1969). In this vision, Campbell proposed rational decision-making by politicians based on rigorous tests of bold social programs designed to improve society. This ideal society would implement reforms demonstrated to be effective by experimental and quasi-experimental research and evaluation, with the goal of improving the lives of most of the population (Donaldson, 2009). Campbell never said that all evaluation should be solely concerned with causal questions and experimental methods, but he clearly viewed other methods as subordinate to quantitatively based knowledge (Shadish & Luellen, 2013).

In the 1980s, Campbell’s vision of emphasizing experimental and quasi-experimental methodology to reduce biases in social research and program evaluations was increasingly de-
emphasized. Given the existing realities of social reforms and program management, the results of quantitative evaluation methodology were not always considered representative of the programs being evaluated (Campbell, Cook, & Shadish, 2002). For example, Cook, Shadish, and Leviton (1991) reported that reforms and programs judged to be poor by experimental standards were often considered acceptable by politicians, policy makers, and managers. It was also argued that rigorous experimental evaluations did not yield credible evidence in a timely manner (Cook et al., 1991).

These realities inspired the applied research and evaluation field to develop new tools and approaches. In particular, some social scientists turned away from quantitative methodology exemplified by the randomized controlled trial (RCT), rejecting its approach of maximum objectivity, turning instead to more qualitative and collaborative methods (Labin, 2008). This began the Paradigm Wars, a struggle between qualitative and quantitative advocates within the ranks of social scientists--program evaluators in particular (Smith, 2008).

Research and evaluation methodology today. The debate about the superiority of one methodology over the other has continued to the present, usually with good intentions to improve the profession. As stated by Smith (2008), the conflict concerning methodological superiority has created a more viable profession of evaluation. It has resulted in a shift from a reliance on narrow experimental approaches to include a wider range of quantitative designs and more sophisticated qualitative designs (Smith, 2008).

However, in 2003, the United States Department of Education issued a notice of proposed priority to establish experimental methods as the gold standard for certain educational evaluations (US DOE, 2003). Though many evaluators would agree that experimental methods are the best choice in certain cases, few would agree that they are the best method in every case (Smith, 2008). The issue is not simply whether experimental approaches are the best design, but whether experimental approaches should be a national mandate. Jennifer C. Greene (2009) argued that these mandated assumptions about methodology were not about epistemology but about political maneuverings. Tom Schwandt (2007) called this epistemological politics, where one source of knowledge is privileged over another in making
decisions about public policies. He suggested that epistemological politics were driving federal directives on methodology choice (Schwandt, 2007).

Schwandt (2009) argued that a quest for credible evidence should not center solely on methodology while ignoring what constitutes evidence and its responsible use for informing a conclusion or judgment. Instead, he proposed that a quest for credible evidence include analyses of (1) the character of evidence, (2) the ethics of evidence, (3) the contexts of the application of evidence, and (4) the nature of rationality and argumentation (Schwandt, 2009).

First, the character of evidence includes the properties of relevance, credibility, and probative force (Schwandt, 2009). Relevant evidence means it bears directly on the claim in question. Credibility is the believability of the information. Probative force is the strength of inference between the evidence and the claim. All three properties should be assessed when determining whether or not the information is useful for informing a judgment or conclusion (Schwandt, 2009). Generally, this aspect of study quality is achieved by studies of good design and execution, studies that score well using the GRADE method or the HEAL method of assessing study quality.

Second, the ethics of evidence refers to the morally correct production, interpretation, dissemination, and use of evidence (Schwandt, 2009). Edmund Pellegrino (1999, p. 34) observed that evidence “has the power to convince others and therefore has the potential to lead or mislead, and to enhance or diminish the lives of others.” Therefore, there is also an obligation to assess the individual and social implications of how evidence is created, interpreted, and used (Schwandt, 2009). While most approaches to assessing study quality, including the GRADE method or the HEAL method, do not assess the ethics of evidence in a study, a metaevaluation does. In this dissertation, though, only the accuracy standards of a metaevaluation are being employed so the ethics of evidence will not be assessed.

Third, the contexts of the application of evidence refers to the contexts within which research and evaluation take place (Schwandt, 2009). The methods used to generate evidence within these varying contexts can include quantitative methods, qualitative methods, and mixed methods. In this type of analysis, evidence is a mediation between the context of its use and the method of its production
(Upshur, 2001). First, one identifies the inquiry purposes and questions, and only then selects a methodology that fits these purposes and questions. As Greene (2009, p. 153) stated, “Method is always the servant of evidence, never the master.” When assessing study quality, it is not usual to ascertain whether or not the context of the study dictated the method of study. Because evaluation and metaevaluation place greater emphasis on context than most healthcare studies, metaevaluation seems to be a better approach to assessing the contexts of the application of evidence than the GRADE method or HEAL method.

Finally, the nature of rationality and argumentation concerns the use of evidence in research and evaluation. Such evidence may be used for persuasion, negotiation, deliberation, and/or inquiry (Schwandt, 2009). It is often assumed that all evaluative decisions are best characterized as inquiry dialogues. However, most decisions faced by practitioners are matters of practical judgment that may involve not only empirical evidence but also opinion, values, and preferences (Chelimsky, 2007). The use of evidence, therefore, is determined by the context of the situation and not by the hierarchy of research methodology (Schwandt, 2009). Again, if the assessment of study quality includes assessing how evidence in research and evaluation is used, then metaevaluation seems to be a better option than either the GRADE method or the HEAL method.

In summary, rather than allowing methodology choice to dictate research and evaluation activities, Schwandt (2009) suggests that the quest for credible evidence be dictated by a focus on evidence character, ethics, context, and use. The argument to use metaevaluation to oversee this quest for credible evidence is compelling.

Study Quality of Research and Evaluation

As stated above, study quality is the fit between a study’s goals and that study’s design and implementation characteristics (Valentine, 2009). The criteria used to determine the merit of a study depends on the type of methodology, quantitative or qualitative, used in the study. The criteria used to determine merit also depends on whether one is assessing a research study or an evaluation study. Generally, the criteria of merit for evaluation study quality is more extensive than for research study
quality since evaluation study quality must consider not only evidence character, but also evidence ethics, context, and use as well.

**Study quality of quantitative research.** Perhaps the most widely used system used to determine quantitative research quality is Campbell’s validity framework (Campbell et al., 2002). Within this framework, factors that might lead to an incorrect inference are called threats to validity. Threats to validity can be placed in four broad categories: threats to internal validity, threats to external validity, threats to construct validity, and threats to statistical conclusion validity.

Internal validity refers to the soundness of the conclusion that a certain intervention caused an observed outcome. Threats to internal validity include any factor that might plausibly cause the observed outcome even if the planned intervention had not occurred (Valentine, 2009). External validity refers to the soundness of the conclusion that a causal claim can be generalized to a population of interest. Threats to external validity include different participants, different contexts, different treatment variations, and outcomes measured in different ways (Valentine, 2009).

Construct validity refers to the soundness of the conclusion that observations or measurement tools used in a study actually represent or measure the construct being investigated. Threats to construct validity include inadequate preoperational explication of constructs, mono-operation bias, mono-method bias, and interaction of different treatments, to name just a few (Trochim, 2006). Statistical conclusion validity refers to the soundness of the statistical conclusion regarding the strength of the relationship between a presumed cause and a presumed effect. Threats to statistical conclusion include any assumption associated with a particular statistical procedure that is not met (Valentine, 2009).

Numerous grading guidelines have been developed to evaluate the quantitative research quality of healthcare studies. The Canadian Task Force on Preventative Health Care (CTFPHC) was the first organization to produce a quantitative study quality grading system in 1979. It grades healthcare studies into three basic levels: studies using randomization, studies using comparisons groups, and studies relying primarily upon expert opinion (Gugiu & Gugiu, 2010). The United States Preventative Service Task Force (USPSTF) and the Australian National Health and Medical Research Council (ANHMRC) produced similar
grading systems in 1989 and 1995 respectively. In 1999, the Oxford Centre for Evidence-Based Medicine (OCEBM) produced a grading system incorporating five levels of evidence: randomized control trials, equivalent control trials (cohort trials), non-equivalent control trials (case-control trials), uncontrolled trials (case-series trials), and expert opinion. Currently, the most popular grading system for quantitative healthcare studies is the GRADE (Grades of Recommendation, Assessment, Development, and Evaluation) method. The GRADE method also includes an assessment of risk of bias using the Cochrane risk of bias tool. The HEAL (Hierarchy of Evidence and Appraisal of Limitations) method, which is used in this dissertation to provide a comparison with metaevaluation ratings, builds on the strengths of the GRADE method by systematically identifying additional important threats to validity and more clearly distinguishing between strong and weak non-randomized controlled trials (Gugiu & Gugiu, 2010).

Gugiu et al. (2013) did an in-depth analysis of the HEAL grading system by selecting healthcare studies conducted on 28 CCM programs. In Table 1, HEAL grade for each study is given, along with the author and date of the study, the study design, and a list of design weaknesses. The grades received by the studies depended primarily upon the robustness of the research design of the study (referred to as the hierarchy of evidence) and upon the efforts of the researchers to guard against the threats to validity that limit the likelihood of attaining statistical significance (referred to as the appraisal of limitations). The criteria used in the HEAL grading system is in Appendix A.
Table 1 HEAL Scores For Selected CCM Studies

<table>
<thead>
<tr>
<th>Author/ Date</th>
<th>HEAL Grade</th>
<th>Study Design</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Bauer et al. (2006)</td>
<td>A-</td>
<td>Objective: Improve outcomes for bipolar disorder Design: Single-blind, stratified RCT Comparison: CCM versus control Duration: 3 years Intervention: PSM, DSD, DS Sample: 11 veterans affairs (VA) hospitals treating 306 veterans (approximately, 153 per group) diagnosed with bipolar disorder Participation (attrition) rate: 73.2% (7.3%)</td>
<td>An examination of patient characteristics in comparison to other studies which had large veteran samples with bipolar disorder revealed a number of differences</td>
</tr>
<tr>
<td>2. Benedetti et al (2004)</td>
<td>C-</td>
<td>Objective: Improve processes and outcomes for diabetics Design: NECT Comparison: CCM versus control Duration: 13 months Intervention: PSM, DSD, DS, CIS, HSO, CR Sample: 30 primary care providers (11 CCM and 19 control) treating 1,998 patients (698 CCM and 1,300 control) diagnosed with diabetes Participation (attrition) rate: ND (3.5%)</td>
<td>Although the authors indicated that the treatment groups had similar baseline values on HbA1c, blood pressure, and urine protein, it is not clear whether they tested for statistical equivalence, or even difference. Moreover, they would have needed to demonstrate equivalence on several other factors in order for the study to be deemed an ECT. Another problem committed in the study is that the interventions implemented a Plan-Do-Study-Act method in which the intervention constantly changed. Therefore, it is impossible to conclusively determine which aspects of the CCM were associated with improvements in outcomes</td>
</tr>
</tbody>
</table>

Continued
<table>
<thead>
<tr>
<th>Author/ Date</th>
<th>HEAL Grade</th>
<th>Study Design</th>
<th>Design Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>3. Caruso et al (2007)</td>
<td>D</td>
<td>Objective: Improve the quality of care for patients with CVD and diabetes Design: UT Comparison: Baseline versus intervention periods 1, 2, and 3 Duration: 39 months Intervention: PSM, DSD, DS, CIS Sample: 10 geriatricians treating 283 patients diagnosed with cardiovascular disease or diabetes Participation (attrition) rate: 86.0% (35.3%)</td>
<td>Caution should be exercised in the interpretation of the diabetes and cardiovascular disease subgroup analyses because significant overlap existed between these subgroups. Patients in the diabetes group also included patients diagnosed with CVD and vice-versa</td>
</tr>
<tr>
<td>4. Chin et al. (2004)</td>
<td>D</td>
<td>Objective: Improve the quality of care for diabetics Design: UT Comparison: Baseline versus post-intervention Duration: 12 months Intervention: PSM, DSD, DS, CIS, CR Sample: 19 health centers treating 969 patients diagnosed with diabetes Participation (attrition) rate: ND (38.0%)</td>
<td>Although no information was provided on the size of the population served by the 19 health centers, patients were randomly selected from the population</td>
</tr>
</tbody>
</table>
Table 1 (Continued)

<table>
<thead>
<tr>
<th>Author/Date</th>
<th>HEAL Grade</th>
<th>Study Design</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>5. Chin et al. (2007)</td>
<td>D</td>
<td>Objective: Improve processes and outcomes for diabetics Design: Clustered Cohort UT Comparison: Standard CCM versus enhanced CCM versus control Duration: 4 years Intervention: PSM, DSD, DS, CIS, HSO, CR Sample: 34 Health Centers treating 6,993 diabetic patients Participation (attrition) rate: 34.3% (35.3%)</td>
<td>Although the authors stipulated that they employed an RCT design, this was only true for the standard CCM versus enhanced CCM comparison. However, the efficacy of the CCM can only be determined by comparison to a true control group. Therefore, for this study, the efficacy of the CCM, either the standard or enhanced, could only be judged by comparing post-intervention outcomes to the corresponding baseline values (i.e., a UT design). Moreover, a cohort study design was employed because the same patients were not followed longitudinally. However, a range of demographic and clinical variables were used to control for differences between cohorts. Therefore, it is likely that the cohort samples were drawn from stable populations (i.e., there were no changes in the populations served by the health centers over time). Another problem committed in the study is that the interventions implemented a Plan-Do-Study-Act method in which the intervention constantly changed. Therefore, it is impossible to conclusively determine which aspects of the CCM were associated with improvements in outcome.</td>
</tr>
<tr>
<td>6. Chumbler et al. (2005)</td>
<td>B</td>
<td>Objective: Improve process outcomes for diabetics Design: Retrospective ECT Comparison: CCM versus control Duration: 12 months Intervention: PSM, DSD Sample: 4 clinics treating 800 veterans (400 CCM and 400 usual care) diagnosed with diabetes Participation (attrition) rate: ND (0%)</td>
<td>The treatment and control patients were not shown to be equivalent. However, propensity score matching was used to adjust for group differences on a range of variables.</td>
</tr>
<tr>
<td>Author/Date</td>
<td>HEAL Grade</td>
<td>Study Design</td>
<td>Design Weaknesses</td>
</tr>
<tr>
<td>------------</td>
<td>------------</td>
<td>---------------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>7. DiPiero et al. (2008)</td>
<td>D</td>
<td>Objective: Improve processes and outcomes for diabetics Design: Retrospective ECT Comparison: CCM versus control Duration: 6 to 24 months Intervention: PSM, DSD, CIS Sample: 565 patients (288 CCM and 277 usual care) diagnosed with diabetes Participation (attrition) rate: ND (0%)</td>
<td>The study design was compromised because lessons learned were potentially disseminated to both groups by medical residents who were exposed to the CCM and then rotated to the control group, CCM staff who may have shared methods with control staff, and/or control patients who were transferred to the CCM intervention. Thus, this could have resulted in the diffusion of treatment components to the control group</td>
</tr>
<tr>
<td>8. Dorr et al. (2005)</td>
<td>B</td>
<td>Objective: Improve processes and outcomes for diabetics Design: Retrospective ECT Comparison: CCM versus control Duration: Between 4 to 18 months Intervention: DS, DSD, CIS, CR Sample: 11 clinics treating 5,925 (1,185 CCM and 4,740 control) patients diagnosed with diabetes Participation (attrition) rate: 38.3% (0%)</td>
<td>The treatment and control patients were not shown to be equivalent. However, statistical matching was used to adjust for group differences on a range of variables. The low participation rate reflects the matching ratio set for the study (1:4 CCM to control)</td>
</tr>
<tr>
<td>Author/ Date</td>
<td>HEAL Grade</td>
<td>Objective</td>
<td>Study Design</td>
</tr>
<tr>
<td>--------------</td>
<td>------------</td>
<td>-----------</td>
<td>--------------</td>
</tr>
<tr>
<td>9. Ely et al. (2008)</td>
<td>D</td>
<td>Improve weight loss for obese patients</td>
<td>RCT</td>
</tr>
<tr>
<td>10. Glasgow et al. (2005)</td>
<td>A</td>
<td>Improve process outcomes for diabetics</td>
<td>Stratified, cluster RCT</td>
</tr>
<tr>
<td>Author/ Date</td>
<td>HEAL Grade</td>
<td>Study Design</td>
<td>Design Weaknesses</td>
</tr>
<tr>
<td>---------------------</td>
<td>------------</td>
<td>--------------</td>
<td>-----------------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
Design: Single-blind, block RCT  
Comparison: Home BP monitoring and patient website (Web CCM) versus home BP monitoring, patient website, and pharmacist care (Pharmacist CCM) versus control  
Duration: 12 months  
Intervention: PSM, DSD, DS, CIS, CR  
Sample: 10 primary care medical centers treating 778 patients (261 Pharmacist CCM, 259 Web CCM, and 258 control) diagnosed with hypertension  
Participation (attrition) rate: 64.2% (6.2%) | An analysis of differences or equivalence between participants and eligible nonparticipants was not conducted |
Design: Retrospective UT  
Duration: 3 years  
Intervention: PSM, DSD, DS, CIS, HSO, CR  
Sample: 14 practices treating 11,896 patients diagnosed with diabetes  
Participation (attrition) rate: 90.0% (0%) | The authors presented a baseline period and two follow-up periods. However, some of the medical practices began implementing the intervention during the baseline period; thus, potentially, contaminating the comparison period. Moreover, although a control group was identified for secondary analyses, no comparison group was used for the primary analysis. |
<table>
<thead>
<tr>
<th>Author/Date</th>
<th>HEAL Grade</th>
<th>Study Design</th>
<th>Design Weaknesses</th>
</tr>
</thead>
</table>
Design: Retrospective NECT  
Comparison: CCM versus control  
Duration: Between 11 to 17 months  
Intervention: PSM, DSD, DS, HSO  
Sample: 79 patients (44 CCM and 35 control) diagnosed with diabetes with high cardiovascular risk  
Participation (attrition) rate: 39.3% (0%) | The authors tested the two groups for baseline differences on age, intermediate outcomes, and medication use. However, these tests do not demonstrate that two groups are equivalent. The authors should have conducted a test for equivalence between the two groups on the baseline measures. Moreover, numerous demographic variables that are known to impact diabetes (e.g., race) were not tested. Finally, the study periods for the treatment (4/2005-8/2006) and control (1/2005-8/2006) group were not identical. Thus, given that the length of time for which controls received medical care was slightly longer than that of the intervention group, the effects of the treatment could have been masked, especially given that the study was underpowered. |
Design: Single-blind, block RCT  
Comparison: CCM versus control  
Duration: 19 + 2 months  
Intervention: PSM, DSD, DS  
Sample: 2 Veteran Affairs clinics treating 246 patients (123 CCM and 123 control) diagnosed with poorly controlled diabetes  
Participation (attrition) rate: 53.5% (15.0%) | The intensity of the intervention was weaker than anticipated by the authors. At one of the two sites, over 70% of attempted phone contacts (principle intervention modality) were unsuccessful. Likewise, case managers reported having minimal or no contact with 40% of the CCM patients. Additionally, the study was very underpowered due to inaccurate assumptions regarding the effect size and standard deviation. Finally, differences in age and mean HbA1c values were found between participants and nonparticipants. |
<table>
<thead>
<tr>
<th>Author/Date</th>
<th>HEAL Grade</th>
<th>Study Design</th>
<th>Design Weaknesses</th>
</tr>
</thead>
</table>
| 15. Mangione-Smith et al. (2005) | D          | Objective: Improve processes and outcomes for asthmatics  
Design: NECT  
Comparison: CCM versus control  
Duration: 12 months  
Intervention: PSM, DSD, DS, CIS, HSO, CR  
Sample: 13 clinics (9 CCM and 4 control) treating 501 asthmatic youth (348 CCM and 153 control)  
Participation (attrition) rate: ND (0%) | Although the authors claimed that the study used a matched design, matching only occurred for outcomes obtained on a follow-up survey. Since only outcomes abstracted from medical records could be used to investigate the efficacy of treatment, the study design was deemed to be a NECT. Moreover, differences on demographic characteristics were found to exist between the groups. Furthermore, because the interventions implemented a Plan-Do-Study-Act method in which the intervention constantly changes, it is impossible to conclusively determine which aspects of the change were associated with improvements in outcomes |
Design: Cohort UT  
Comparison: Standard CCM versus enhanced CCM versus control  
Duration: 13 months for 1 clinic, 18 months for the other 2 clinics  
Intervention: PSM, DSD, DS, CIS  
Sample: 3 clinics treating 747 women (400 pre-CCM and 347 post-CCM)  
Participation (attrition) rate: 75.2% (0%) | Although the authors stipulated that they employed a Clustered RCT design, this was only true for the standard CCM versus enhanced CCM comparison. However, the efficacy of the CCM can only be determined by comparison to a true control group. Furthermore, even if a true control group existed, the randomization process employed could not ensure that known and unknown confounding factors were evenly distributed because three clinics were randomized to two treatment groups. Examination of the demographic variables, indeed, suggested that differences existed between the samples served by the three clinics. Thus, for this study, the efficacy of the CCM, either the standard or enhanced, could only be judged by comparing post-intervention outcomes to the corresponding baseline values (i.e., a UT design). Moreover, a cohort study design was employed because patients were not followed longitudinally. Finally, although statistical tests of difference suggest that the underlying population from which the cohorts were sampled may have been the same, this can only be demonstrated by statistical tests of equivalence |

Continued
# Table 1 (Continued)

<table>
<thead>
<tr>
<th>Author/Date</th>
<th>HEAL Grade</th>
<th>Study Design</th>
<th>Design Weaknesses</th>
</tr>
</thead>
</table>
| 17. Piatt et al. (2006) | D          | Objective: Improve clinical and behavioral outcomes for diabetics  
Design: Cluster RCT  
Comparison: CCM versus provider education (PE) only versus control  
Duration: 12 months  
Intervention: PSM, DSD, DS  
Sample: 11 primary care practices (3 CCM, 3 PE, and 5 control) treating 119 patients (30 CCM, 38 PE, and 51 control) diagnosed with diabetes  
Participation (attrition) rate: 45.8% (11.8%) | Eleven clinics were randomized to three groups. As a result of the low number of clinics, the randomization process could not ensure that known and unknown confounding factors were evenly distributed between the three groups. However, because the authors utilized statistical controls in their analyses to adjust for group differences (on a few variables), their design is parallel to that of an ECT. Another problem was the authors proceeded with the study despite the fact that the initial recruitment was about half the recommended sample estimated by an a priori power analysis. Finally, differences in experience were found between participating and nonparticipating primary practices. |
Design: RCT  
Comparison: CCM versus control  
Duration: 6 months  
Intervention: PSM, DSD, CR  
Sample: 3 pediatric clinics treating 319 parents (160 CCM and 159 control) with children aged 0 to 5  
Participation (attrition) rate: 35.1% (91.5%) | The authors decided to continue with the study even though the size of the population they had access to was smaller than the sample they need to show significant differences according to their a priori power analysis. Moreover, given the brief nature of the intervention (two counseling session lasting less than 15 min), their expectation that it could reduce the percentage of noncompliant subjects, relative to the control group, by 40% seems unrealistic. Furthermore, only a third of the target population agreed to participate in the study and differences were found to exist between participants and nonparticipants, which may have been exacerbated by the high attrition rate among study participants who agreed to a home visit (i.e., data source for two primary outcomes). Also, the attrition rate between the two groups differed slightly. |
<table>
<thead>
<tr>
<th>Author/Date</th>
<th>HEAL Grade</th>
<th>Study Design</th>
<th>Design Weaknesses</th>
</tr>
</thead>
</table>
| 19. Scanlon et al. (2008) | B | Objective: Reduce medical payments and improve quality for diabetics  
Design: Retrospective ECT  
Comparison: CCM versus control  
Duration: 12 months  
Intervention: PSM, DSD, DS, CIS  
Sample: 10 CCM clinics and an unknown number of control clinics treating 386 patients (193 CCM and 193 control) diagnosed with diabetes  
Participation (attrition) rate: 18.7% (0%) | The treatment and control patients were not shown to be equivalent. However, propensity score matching was used to adjust for group differences on a range of variables. The low participation rate reflects the matching ratio set for the study (1:1 CCM to control). |
Design: Block RCT  
Comparison: Automated telephone self-management support (ATSM) versus group medical visits (GMV) versus control  
Duration: 12 months  
Intervention: PSM, DSD  
Sample: 339 patients (112 ATSM, 113 GMV, and 114 control) with poorly controlled diabetes  
Participation (attrition) rate: 82.9% (10.0%) | Some differences between participants and eligible non-participants were found with respect to language spoken, race/ethnicity, and insurance |
<table>
<thead>
<tr>
<th>Author/Date</th>
<th>HEAL Grade</th>
<th>Study Design</th>
<th>Study Design Weaknesses</th>
</tr>
</thead>
</table>
Design: Single-blind, NECT  
Comparison: CCM versus control  
Duration: 15 months  
Intervention: DS, CIS, HSO, CR  
Sample: 5 clinics (3 CCM and 2 control) treating 185 asthmatic patients (109 CCM and 76 control)  
Participation (attrition) rate: ND (0%) | Although the authors claimed that the study used a matched design, matching only occurred for outcomes obtained on a follow-up survey. Since only outcomes abstracted from medical records could be used to investigate the efficacy of treatment, the study design was deemed to be a NECT. Moreover, because the interventions implemented a Plan-Do-Study-Act method in which the intervention constantly changes, it is impossible to conclusively determine which aspects of the change were associated with improvements in outcomes |
Design: Retrospective UT  
Comparison: Baseline versus post-intervention  
Duration: Through 12 months  
Intervention: PSM, DSD, DS, CIS, HSO  
Sample: 8 hospitals and 2 primary care practices treating 705 (382 patients from the hospitals sample and 323 from the primary care practices sample) diagnosed with diabetes  
Participation (attrition) rate: 12.8% (0%) | Although a comparison group was used for one of the analyses, efficacy could not be established against this group because it also implemented the CCM. Therefore, efficacy could only be established against baseline data. The primary differences between the two groups were that the comparison group implemented the CCM in a primary care setting, as compared to a hospital setting, and began implementing the CCM about halfway through the tracking period |

Continued
<table>
<thead>
<tr>
<th>Author/Date</th>
<th>HEAL Grade</th>
<th>Study Design</th>
<th>Weaknesses</th>
</tr>
</thead>
</table>
Design: UT  
Comparison: Baseline versus post-intervention  
Duration: 12 months  
Intervention: PSM, DSD, DS, HSO  
Sample: 4 physicians treating 104 patients diagnosed with diabetes  
Participation (attrition) rate: 100% (0%) | Only 28% of patient offered diabetes self-management education participated and only 16% of patients offered the education completed the program.                                                                                          |

24. Stock et al. (2008) | D          | Objective: Improve the physical function and health-related quality of life of seniors  
Design: UT  
Comparison: CCM versus control 1 (patients of primary care physicians supported by care managers) versus control 2 (patients of primary care physicians without care managers)  
Duration: 30 months  
Intervention: PSM, DSD, DS, CIS  
Sample: 59 primary care physicians treating 1,307 geriatric patients (318 CCM, 598 control 1, and 391 control 2)  
Participation (attrition) rate: 24.9% (23.5%) | Although the authors conducted a controlled study, their analyses did not compare the treatment group to either of the two control groups. Had this comparison been made then the design would be classified as a NECT due to the numerous statistically significant differences found on baseline and demographic variables between the intervention and two control groups. Another problem committed in the study is that the interventions implemented a Plan-Do-Study-Act method in which the intervention constantly changed. Therefore, it is impossible to conclusively determine which aspects of the CCM were associated with improvements in outcomes. Finally, the reasons for dropping out did not differ significantly according to study group.                                                                                           |
<table>
<thead>
<tr>
<th>Author/Date</th>
<th>HEAL Grade</th>
<th>Study Design</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>25. Stroebel et al. (2000)</td>
<td>D</td>
<td>Objective: Improve BP control for patients with CVD</td>
<td>Although differences in age between the groups were controlled for, numerous demographic variables related to BP were not considered. Thus, the equivalence of the groups was not adequately demonstrated. Furthermore, some of the CCM components may have diffused to the control physicians because the clinical teams that treated the intervention and control patients were located in the same clinic. Likewise, nine months into the intervention period, the intervention was expanded by the addition of 10 clinical teams originally assigned to the control group. Thus, a significant portion of the treatment group was exposed to a lower “dosage” of the intervention than their counterparts. There is also no way of determining the nature of the selection bias that occurred (e.g., did clinicians with healthy patients leave the control group or did clinicians with unhealthy patients leave the control group?) or whether one even occurred.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Design: Retrospective NECT</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comparison: CCM versus control</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Duration: 27 months</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intervention: PSM, DSD</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sample: 30 clinical teams (15 CCM and 15 control) treating 867 patients (152 CCM and 715 control) diagnosed with hypertension</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Participation (attrition) rate: 18.9% (0%)</td>
<td></td>
</tr>
<tr>
<td>26. Stroebel et al. (2005)</td>
<td>D</td>
<td>Objective: Improve the quality of chronic illness care to medically uninsured patients</td>
<td>Nearly 27% of the intent to treat sample (N =149) left the study either because they acquired insurance (11.4%), moved from the area (8.7%), or were unaccounted for (6.7%). Also, the length of the intervention varied greatly among the patients. It is worth noting that because the population sampled was uninsured, baseline clinical measures are likely to reflect unmanaged disease states, rather than states attributed to patients receiving treatment as usual.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Design: UT</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comparison: Baseline versus post-intervention versus national standards</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Duration: Up to 22 months with average follow-up of 12 months</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intervention: PSM, DSD, DS, CIS, HSO</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sample: 149 uninsured patients diagnosed with diabetes, hypertension, or hyperlipidemia</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Participation (attrition) rate: ND (26.8%)</td>
<td></td>
</tr>
</tbody>
</table>

Continued
<table>
<thead>
<tr>
<th>Author/Date</th>
<th>HEAL Grade</th>
<th>Objective: Reduce CVD risk for diabetics</th>
<th>Study Design</th>
<th>Design Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>27. Vargas et al. (2007)</td>
<td>B-</td>
<td>Design: ECT</td>
<td>Comparison: CCM versus control</td>
<td>Duration: 11 months</td>
</tr>
</tbody>
</table>

| 28. Warm et al (2008) | D | Objective: Improve the ambulatory experience for residents and patients | Design: UT | Comparison: Baseline versus post-intervention | Duration: 12 months | Intervention: PSM, DSD, DS, CIS | Sample: 108 residents treating 4,947 patients | Participation (attrition) rate: ND (ND) | Many aspects of the intervention were changed simultaneously due to numerous improvement projects, so it is impossible to conclusively determine which aspects of the change were associated with improvements in outcome. |

Note. ADA = American Diabetes Association; BP = blood pressure; CCM = intervention implemented by study; CDE = certified diabetes educator; CIS = clinical information system; CR = community resources; CVD = cardiovascular disease; DS = decision support; DSD = delivery system design; EMR = electronic medical record; HbA1c = hemoglobin A1c; HSO = health system organization; LDL = low-density lipoprotein; ND = no data; PSM = patient self-management; HEAL = Hierarchy of Evidence and Appraisal of Limitations.
Regarding the hierarchy of evidence in the HEAL grading system, ECTs (equivalent controlled trials) are not as robust in design as RCTs (randomized controlled trials) because they do not employ randomized assignment to assure equivalency between groups, but they do employ statistical techniques or study design elements to reduce baseline differences between groups. NECTs (nonequivalent controlled trials) and UTs (uncontrolled trials) are less robust in design than either RCTs or ECTs because they make minimal efforts (NECTs) or no effort (UTs) to reduce baseline differences between groups (Gugiu et al., 2013).

Of the 28 studies in Table 1, 10 were reported as RCTs, 5 as ECTs, 6 as NECTs, and 7 as UTs. In actuality, three of the studies (two reported as RCTs and one reported as an NECT) did not use true control groups but rather compared a standard CCM treatment against an enhanced CCM treatment. Those studies were downgraded to UTs, which changed the hierarchy of evidence scoring to 8 RCTs, 5 ECTs, 5 NECTs, and 10 UTs (Gugiu et al, 2013).

In a perfect world, the hierarchy of evidence based on research design would be sufficient to distinguish between higher study quality and lower study quality. But upon examination of the selected CCM studies, this is hardly a perfect world. Regarding the appraisal of limitations, few of the studies in Table 1 (6 of 28) were executed flawlessly. Gugiu et al. (2013) assessed the limitations (i.e. flaws) of these studies using three categories of limitations: serious limitations, design-specific limitations, and minor limitation (see Appendix A).

The top four serious limitations found in the studies in Table 1 were low dosage, low statistical power, violation of the intent-to-treat principle, and the absence of a “true” control group (Gugiu et al., 2013). The first two limitations increase the likelihood of a Type II error (failure to reject a false null hypothesis), whereas the latter two limitations increase the likelihood of a Type I error (rejection of a true null hypothesis). Of the 28 studies, 13 had at least one serious limitation (Gugiu et al., 2013).

Of the 28 studies in Table 1, 11 had design-specific limitations. Examples of common design-specific limitations include flawed randomization in RCTs or no tests for statistical equivalence in ECTs. Of the 28 studies, 20 had one or more minor limitations, most commonly the limiting effects of constantly
changing interventions. CCM fidelity is difficult to maintain. This may explain why so many of the selected CCM studies were plagued by this particular limitation (Gugiu et al., 2013).

One type of changing intervention is the number of CCM components implemented in each study. Of the 28 studies in Table 1, 4 studies implemented all six components of the CCM, 4 studies implemented five components of the CCM, 11 studies implemented four components, 5 studies implemented two components of the CCM, and 4 studies implemented one component of the CCM. These varying numbers of implemented CCM components makes it difficult to compare the results of the studies (Gugiu et al., 2013). Another type of changing intervention found in the studies in Table 1 is the likelihood of variation within each applied CCM component. For example, if a care coordinator is used in the delivery system design of one study but not in the delivery system design of another study, it is difficult to compare results between the two studies (Gugiu et al., 2013).

Although the reliability and validity of the HEAL grading system has yet to be assessed (Gugiu et al., 2013), it is arguably one of the best grading system available for assessing study quality. However, the utility of the HEAL grading system is limited to assessing quantitative research study quality. Neither this grading system, nor any of the other available grading systems including the GRADE system, is suitable for assessing qualitative research or evaluation study quality.

Study quality of qualitative research. Qualitative researchers strive for a deeper understanding of a phenomenon that requires visiting personally with participants, spending extensive time in the field, and probing to obtain meaning. During and after a study, the qualitative researcher must ask if the account is valid. In the assessment of qualitative research quality, the concept of validation varies depending on procedural, interpretive, emancipatory, and postmodern perspectives (Creswell, 2007).

LeCompte and Schensul (1999) described the use of parallel, qualitative equivalents to internal validity, external validity, reliability, and objectivity. Lincoln and Guba (1985) provided alternative concepts that apply more to naturalistic axioms, including credibility, transferability, dependability, and confirmability. Rather than addressing validation, Eisner (1991) sought to provide reasonable standards for judging credibility of qualitative research by introducing the concepts structural corroboration,
consensual validation, referential adequacy, and ironic validity. Working from postmodern and interpretive perspectives, Lather (1993), Wolcott (1994), Angen (2000), Wittemore, Chase, and Mandle (2001), and Richardson and St. Pierre (2005) all either re-conceptualized or abandoned the notion of validity altogether in pursuit of a deeper and more dynamic understanding of research quality.

Creswell and Miller (2000) described eight strategies that are most commonly used to assess the validity of a qualitative study. These strategies include (1) prolonged engagement and persistent observation in the field, (2) triangulation among different data sources, (3) peer review or debriefing, (4) negative case analysis, (5) clarifying research bias, (6) taking the written narrative back to participants in member checking, (7) writing with detailed and thick description, and (8) external audits. Creswell (2007) recommends that qualitative researchers engage in at least two of these strategies to ensure adequate qualitative research quality.

**Study quality of evaluation.** The primary purpose of evaluation is to not only to contribute to the understanding of how an evaluand works, but also to assign merit and worth to that evaluand. Therefore, the study quality of evaluation is judged not only by accuracy, but also by utility, feasibility, propriety, and evaluation accountability. The assignment of merit and worth also requires a universal focus on stakeholder perspectives. The study quality of evaluations is therefore judged on whether or not these perspectives are considered (Yarbrough et al., 2011).

Metaevaluations are systematic reviews of evaluations to determine their study quality (Cooksy & Caracelli, 2005). Metaevaluation can be used to assess the quality of a single study or a set of studies. Metaevaluation of a single study can be conducted for a formative purpose to improve an on-going evaluation or for a summative purpose to approve a completed evaluation (Cooksy & Mark, 2011).

Most metaevaluation approaches rely on an appraisal checklist (Pawson, 2006). Manning (2012) notes that though there is a proliferation of checklists, little work has been done to evaluate them. In mixed methods synthesis, one challenge with the use of checklists is how to operationalize quality criteria from different methodologies (Manning, 2012). Saini and Shlonsky (2012) developed a checklist intended to appraise studies in terms of epistemological frameworks, ethical issues, and judgments about fairness.
and justice. As the evaluation field develops the area of mixed methods synthesis, the metaevaluation aspect of the review becomes increasingly complex (Cooksy & Mark, 2011).

Reviews of sets of evaluation have two major purposes. The first major purpose for reviewing a set of evaluations is to decide which evaluations to include in an evaluation synthesis (Cook et al., 1992). Similar to research synthesis, an evaluation synthesis combines information from several evaluations in order to come to general statements about an intervention. Evaluation syntheses are used to assess overall impact but they are only as good as the evaluation findings they synthesize. Metaevaluation screens out those studies that are so weak methodologically that the information they include is not defensible (Cooksy & Caracelli, 2005).

The second major purpose for reviewing a set of evaluations is to identify strengths and weaknesses in evaluation practice in order to develop evaluation capacity (Bickman, 1997). For example, several evaluations conducted for a particular program can be evaluated to determine what kind of evaluation training is needed to produce higher quality evaluations for that same program.

There are many approaches to metaevaluation. Cook and Gruder (1978) list seven suggested approaches, including (1) review of an evaluation report, (2) review of literature about a specific program, (3) re-evaluation of a program, (4) evaluation of multiple data sets from a program, (5) consultant metaevaluation, (6) simultaneous secondary analysis of data sets from a program, or (7) multiple independent evaluations of a program. Scriven (2007) suggests three approaches to metaevaluation, including (1) re-evaluation of a program using the same methodology, (2) re-evaluation of a program using different methodology, or (3) assessing the quality of the evaluation using a set of standards such as the Program Evaluation Standards (Yarbrough et al., 2011) or the Key Evaluation Checklist (Scriven, 2007).

According to Wingate (2009), the Program Evaluation Standards are the most widely used criteria for evaluating educational evaluations in the United States. The typical metaevaluation involves one rater who relies mostly on evaluation reports to inform the metaevaluation and the Program Evaluation Standards as criteria to determine evaluation study quality (Wingate, 2009).
Program Evaluations Metaevaluation Checklist. The Program Evaluations Metaevaluation Checklist was created by Dan Stufflebeam (2011) in conjunction with the creation of evaluation standards for programs and personnel (Stufflebeam & Coryn, 2012). The purpose of the Checklist is to assess program evaluations against standards for high quality evaluations. The checklist is used to evaluate evaluations based on the Program Evaluation Standards (PES). The five categories of standards of the PES are utility, feasibility, propriety, accuracy, and evaluation accountability (Stufflebeam, 2011). The checklist is not designed to be completed using only the evaluation report. According to Stufflebeam (2001), the metaevaluator often must collect additional needed information other than the information gathered from the evaluation report. This includes phone and on-site interviews, observations, and the study of portfolios. To reach valid conclusions, metaevaluators need to be able to access all available information and collect any further information if needed (Stufflebeam, 2001).

Reliability is a serious concern when applying the Program Evaluations Metaevaluation Checklist (Wingate, 2009). Reliability is the degree to which an assessment tool produces stable and consistent results (Allen & Yen, 1979). Validity refers to how well a test measures what it is purported to measure. The reliability of a test affects the validity of a test since a test cannot correlate more highly with any other score (the indicator of validity) than it correlates with its own true score (the indicator of reliability) (Allen & Yen, 1979). If, indeed, the checklist is invalid, it cannot be relied on to assess evaluation study quality.

By doing an in-depth analysis of the Program Evaluations Metaevaluation Checklist using multiple raters from diverse perspectives, Wingate (2009) found that the inter-rater reliability between raters using the checklist was uniformly low. On individual standards, ICC values ranged from .01 to .57. There was somewhat greater reliability when rating according to the accuracy standards (Wingate, 2009). This poor showing of the metaevaluation checklist regarding its reliability should raise concern and caution for evaluation stakeholders and should also encourage evaluators to take necessary steps to increase the reliability, and thereby the validity, of metaevaluation checklist.
There are steps that can be taken to increase the reliability of the metaevaluation checklist including (1) increasing documentation and transparency in the process, (2) creating and agreeing on decision rules for determining the extent to which a standard was met, (3) using more than one rater, and (4) calibrating the raters to increase inter-rater reliability (Wingate, 2009). One rater may be used for purposes of reflection or improvement, but not for metaevaluation. Finally, the checklist itself may need to be modified, such as rewriting items on the checklist that assess multiple criteria, and the scoring may need to be revised, such as changing how N/A is scored (Wingate, 2009).

Chronic Care Model

Development of the Chronic Care Model

Healthcare quality. In 1910, Abraham Flexner wrote the Report to the Carnegie Foundation which documented the poor quality of medical practice and education being provided by the nation’s medical schools and major hospitals (Flexner, 1910). In response to this report, the American College of Surgeons created the Hospital Standardization Program in 1917 which established minimum quality standards for the care of patients within the hospital setting (Luce, Bindman, & Lee, 1994). Additional standards were created over the next few decades addressing other hospital concerns including physical plant issues, equipment, and administrative structure (Roberts, Coale, & Redman, 1987).

In 1952, the American College of Surgeons was joined by the American College of Physicians, the American Medical Association, the American Hospital Association, and the Canadian Medical Association to create the Joint Commission on Accreditation of Hospitals (Roberts et al., 1987). Although the Joint Commission initially used minimum quality standards for its accreditation process, it abandoned this approach in 1966 for the use of optimal achievable standards. These new standards were adopted because most American hospitals were already meeting the minimum standards, the newly established Medicare law set more rigorous standards that hospitals had to meet, and the research and evaluation methodology used to assess healthcare quality had grown more sophisticated (Luce et al., 1994).
In 1966, Donabedian attempted to describe and evaluate current methods for assessing the quality of medical care. He recommended that healthcare quality be evaluated in three areas: structure (healthcare facilities and staffing), process (delivery of healthcare), and outcome (outcome of healthcare delivery) (Donabedian, 1966). The Joint Commission ultimately embraced the structure-process-outcome model and still uses it today (Roberts et al., 1987).

In 1988, the Joint Commission adapted a technique called continuous quality improvement (CQI) that had been developed earlier for the manufacturing industry. This method evaluates the performance of groups rather than the performance of individuals (Berwick, 1989). Representatives from each of the disciplines in the group form a task force that continually analyzes and improves the quality of healthcare. The Joint Commission had grown to include long-term care, community mental health, ambulatory care, and hospices. In recognition of its expanded role, its name was changed in 1987 to the Joint Commission on Accreditation of Healthcare Organizations (Roberts et al., 1987).

In 1996, the Institute of Medicine (IOM), a component of the National Academy of Sciences, Washington, DC, convened the National Roundtable on Health Care Quality, which met six times between February 1996 and January 1998. During this time, the IOM roundtable reached a consensus on the precise definition of healthcare quality and how best to measure healthcare quality (Chassin & Galvin, 1998).

The roundtable agreed that the IOM’s definition of quality, developed in 1990, was still robust: “Quality of care is the degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge (Lohr, 1990, p. 10).” The term “health services” refers to all services that intend to improve health, including services for both physical and mental health. The phrase, “individuals and populations,” draws attention to the two primary perspectives on healthcare quality that must be addressed. The phrase, “health outcomes,” refers to health outcomes that patients may desire. The phrase, “increases the likelihood,” indicates that quality is not identical to positive outcomes because disease often prevails. Finally, the
phrase, “current professional knowledge,” emphasizes that practitioners must stay abreast of the knowledge base in each profession and use that knowledge appropriately (Chassin & Galvin, 1998).

The roundtable agreed on the need to broaden the domain of measures to include outcomes as well as processes of care and to speak to the concerns of consumers by developing outcome measures that go beyond immediate morbidity and mortality to include various kinds of functional status. It concluded that either processes or outcomes may be valid measures of quality (Chassin & Galvin, 1998). For an outcome to be a valid measure, it must be closely related to processes of care that can be modified to affect the outcome. For a process to be a valid measure, it must be closely related to an outcome that we care about. It was also concluded that quality of care for a great variety of specific clinical conditions and procedures can be measured with sufficient precision to make judgments and take needed actions to bring about improvement (Chassin & Galvin, 1998).

Around the same time that the IOM roundtable released its report, President Clinton’s Advisory Commission on Consumer Protection and Quality in the Health Care Industry released its final report on how to define, measure, and promote quality of health care. To produce the report, the Advisory Commission conducted a review of academic literature on the quality of healthcare in the United States. In the report, the Advisory Commission noted that there was a surprisingly small amount of systematic knowledge available on the quality of health care delivered in the United States (Schuster, McGlynn, & Brook, 1998).

The Advisory Commission determined that good healthcare quality means providing patients with appropriate services in a technically competent manner, with good communication, shared decision making, and cultural sensitivity. The Commission also determined that poor healthcare quality can mean too much care (e.g., providing unnecessary tests, medications, and procedures, with associated risks and side effects), too little care (e.g., not providing an indicated diagnostic test or a lifesaving surgical procedure), or the wrong care (e.g., prescribing medicines that should not be given together or using poor surgical technique) (Schuster et al., 1998).
The Advisory Commission highlighted the National Committee for Quality Assurance’s (NCQA) Health Plan Employer Data and Information Set (HEDIS). HEDIS is a performance measurement tool designed to assist purchasers and consumers in evaluating managed care plans and to hold plans accountable for the quality of their services. It is a useful system that could eventually lead to the development of a more comprehensive, national healthcare quality assessment system. The Advisory Commission recommended such a system to provide a comprehensive assessment of quality of care for the nation—including how quality varies by population subgroups (e.g., gender, age, race/ethnicity, income, region of country, size of community) and how quality is changing over time (Schuster et al., 1998).

As a result of these two major reports in 1998 on healthcare quality, the National Quality Forum (NQF) was established (Kizer, 2000). The Forum is a private, non-profit, membership organization proposed by the Advisory Commission on Consumer Protections and Quality in the Health Care Industry. The primary goal of the NQF is to improve the quality of healthcare. The primary strategy the Forum uses to reach this goal is to improve quality measurement and reporting mechanisms (Kizer, 2000).

This nationwide conversation on healthcare quality continued in 2001 when the landmark report, “Crossing the Quality Chasm: A New Health System for the 21st Century,” was published (CQHCA, 2001). This report by the Committee on the Quality of Health Care in America addressed quality-related issues more broadly, providing strategic direction for redesigning the health care delivery system of the 21st century. The overall conclusion of the report was that quality problems are everywhere in the country, affecting many patients. According to the report, “Between the health care we have and the health care we could have lays not just a gap, but a chasm” (CQHCA, 2001, p. 1).

In this CQHCA report, four underlying reasons were cited for the healthcare quality chasm in the United States: (1) the growing complexity of science and technology, (2) the increase in chronic conditions, (3) a poorly organized healthcare delivery system, and (4) constraints on exploiting the revolution in information technology (CQHCA, 2001). Two of these four factors, the increase in chronic conditions and a poorly organized healthcare delivery system, are relevant to this dissertation.
Burden of chronic illness. The average American born today can expect to live more than 76 years (CDC, 2007). Roughly 1 additional year has been added to life expectancy every 5 years since 1965. One consequence of growing older is the increase in chronic conditions. Chronic conditions, defined as illnesses that last longer than 3 months and are not self-limiting, are now the leading cause of morbidity and mortality in this country, and affect almost half of the U.S. population. The majority of health care resources are now devoted to the treatment of chronic disease. In 2005, the direct medical costs for persons with chronic conditions were $425 billion, nearly 70 percent of all personal health care expenditures (DeVol & Bedroussian, 2007).

Unlike acute care, chronic care is a collaborative process which both the practitioner and the patient must understand (CQHCA, 2001). There is a need for collaboration in the development and execution of a care plan. The patient (and perhaps the patient’s family) requires training in self-management and the accessing of support services. Often, there must be sustained follow-up visits, telephone calls, and other forms of continuous communication. The collaboration involved in much of the care provided to the chronically ill creates another layer of complexity to the delivery of health care to this growing segment of the population (CQHCA, 2001).

The current healthcare delivery system is not centralized. Of all the physician offices in the nations, 40% are still one-physician offices, and 80% of all physicians practice in offices with fewer than ten physicians (American Medical Association, 1998). Consequently, patients are passed from practitioner to practitioner in a confusing and often impersonal manner. This process is wasteful, leads to loss of information, and does not allow the formation of vital therapeutic relationships between patient and practitioner. In a population increasingly afflicted by chronic conditions, the health care delivery system is poorly organized to provide care to the chronically ill (CQHCA, 2001).

McGlynn, Goss, Elmore, and Lessler (2003) gave further testimony to the healthcare quality gap in America. They found that, on average, Americans receive about half of recommended medical care processes. This gap persists despite the initiatives by both the federal government and private health care delivery systems to improve care. It was recommended as part of the solution to this quality gap that
a national baseline for performance be established to assess the effect of policy changes and to evaluate large-scale efforts to improve quality (McGlynn et al., 2003).

A report from the Milkin Institute states that more than half of Americans suffer from one or more chronic illnesses. This increase in incidence and prevalence of chronic illnesses in America is a major contributor to the growth in medical expenditures. The report states that much of this cost is avoidable (DeVol & Bedroussian, 2007).

Seven of the most common chronic illnesses in America—cancer, diabetes, hypertension, stroke, heart disease, pulmonary conditions, and mental disorders—were analyzed for economic costs that could be avoided through more effective prevention and treatment. More than 109 million Americans report having at least one of these seven diseases, for a total of 162 million cases. The total impact of these diseases on the economy is $1.3 trillion annually (DeVol & Bedroussian, 2007).

The researchers projected rates of disease and associated costs assuming modest improvements in preventing and treating disease. They found that 40 million cases of chronic illness could be prevented and the economic impact of disease could be reduced by $1.1 trillion. They concluded that incentives should be included in the healthcare system that promote prevention and early intervention of chronic illnesses (DeVol & Bedroussian, 2007).

Even with incentives in place for improving prevention and treatment of chronic illnesses, the United States healthcare workforce is not prepared for the growing chronic disease burden (Bodenheimer et al., 2009). Increased numbers of primary care clinicians (i.e. primary care physicians, nurse practitioners, and physician assistants) will be needed, as well as a reimbursement policy that favors the use of multidisciplinary teams that work specifically with patients with chronic illnesses. Currently, however, the ratio of primary care clinicians to population is projected to fall by 9% by 2020 while the ratio of specialist physicians to population will rise 14%. Also, the vast majority of ambulatory care visits are paid for on a per visit, fee-for-service basis which discourages the use of multidisciplinary teams (Bodenheimer et al., 2009).
Pearson, Bhat-Schelbert, and Probst (2012) analyzed cross-sectional data from the 2009 Behavioral Risk Factor Surveillance System to determine the life burden of patients with one or more of nine chronic health conditions. They concluded that as the population of patients with chronic illness grows, the workloads of primary care clinicians will increase, leading to inadequate payment for the required level of management. They suggested that an emphasis on practice redesign to reward clinics that use the chronic care model within the patient centered medical home (PCMH) may help reduce costs by rewarding integration and collaborative care required for patients with multiple chronic diseases (Pearson et al., 2012).

The Chronic Care Model. The effective management of chronic illness requires that patients and their families appropriately cope with the illness and its therapies. Clark et al. (1991) described the self-management tasks that patients need to learn to cope with their illness. These tasks included (1) engaging in activities that promote health such as exercise, a healthy diet, social activation, and good sleep hygiene, (2) engaging with the medical community and adhering to their recommended treatment protocols, and (3) monitoring their physical and emotional status and making appropriate management decisions based on symptoms and signs (Clark et al., 1991).

However, the medical community often failed to include the intervention components that contribute to more effective self-management. This is because medical practices are generally organized to respond to the acute and urgent needs of their patients. Because primary care practices are so oriented to acute illness, they may not differentiate their clinical approaches to patients with acute and chronic illness, relying instead on patient-initiated visits and relief of symptoms (Wagner, Austin, & Von Korff, 1996).

Wagner et al. (1996) evaluated special chronic disease clinics in the United States and in Europe that showed improved outcomes and found that there were significant similarities between successful clinics. There were five common elements: (1) the use of care plans, (2) the reorganization of the practice to meet the complex needs of the chronically ill patient, (3) a systematic attention to the behavioral
change needs of patients, (4) ready access to necessary expertise, and (5) a supportive information system (Wagner et al., 1996).

In 2001, a Cochrane Collaborative review examined 41 studies of interventions designed to improve primary care for diabetes (Renders et al., 2001). The successful approaches identified by review included one or more of the following: (1) provider-oriented components such as continuing education or physician feedback, (2) organizational changes in personnel or the management of visits and follow-up, information systems changes, and (3) patient-oriented interventions of an educational or supportive nature. The Cochrane reviewers concluded that the more comprehensive the intervention, the more likely it was to be successful. They also concluded that interventions must not only target providers' behavior but also patients' behavior, including knowledge, skills, and confidence to manage their condition (Renders et al., 2001).

The Group Health Cooperative independently arrived at the same conclusions and, with support from the Robert Wood Johnson Foundation, developed a guide to chronic care improvement called the Chronic Care Model (CCM) (Wagner et al., 2001). The CCM is not intended to be an explanatory theory but rather a synthesis of the best available evidence. It is intended to be flexible and subject to change when new evidence emerges. However, effectiveness studies show that guidelines have only a minimal impact on quality of healthcare unless those guidelines are fully integrated into a medical practice through education and decision-support interventions (Woolf, 1999).

Application of the Chronic Care Model

**CCM systematic reviews.** This model was extended in a national program, Improving Chronic Illness Care, which used the CCM in collaborative chronic disease improvement programs involving large numbers of diverse health care delivery organizations. By 2001, a total of 104 practices had undergone a yearlong quality improvement program where the CCM was used to improve the care of a single chronic condition such as diabetes or asthma (Wagner et al., 2001).
One lesson learned from this national quality improvement program is that effective chronic illness management requires comprehensive system changes that involve more than simply adding features to a system designed to focus on acute care. The program also showed that smaller systems benefit from the CCM as much as larger systems, though the fee-for-service payment system of smaller practices created disincentives to test different ways of using existing personnel or organizing patient visits (Wagner et al., 2001).

Barriers to the integration of the CCM in this national program included reimbursement policies and the behavior of health plans and insurers. Reimbursement policies do not provide reimbursement for patient-provider interactions that are not face-to-face (e.g. telephone follow ups), care for patients in group settings, or self-management education. Health plans and insurers often provided their own disease management interventions, creating confusion and extra work in the practices. However, smaller practices were helped by the information system support provided by some of the health plans (Wagner et al., 2001).

Bodenheimer et al (2002b) reviewed 39 studies on the implementation of the CCM to treat diabetes to determine the extent to which the CCM can improve the management of chronic conditions and reduce health care costs. Thirty-two of the studies showed that interventions based on the CCM improved at least one process or outcome measure for diabetic patients (Bodenheimer et al., 2002b).

Studies seem to indicate that, in order to successfully implement the CCM, conditions must be favorable both externally (i.e. the overall healthcare system and in the community) and internally (i.e. the physician practice) (Bodenheimer et al., 2002b). Externally, visionary leaders are needed to make systemic changes that favor chronic illness care in the community and the healthcare system. Internally, a favorable financial system is necessary to provide incentives to allow the physician practice to make systemic changes internally that favor chronic illness care (Bodenheimer et al., 2002b).

Tsai, Morton, Mangione, and Keeler (2005) conducted a meta-analysis on randomized and non-randomized controlled trials of interventions using the CCM to address asthma, congestive heart failure (CHF), depression, and diabetes. The effect sizes for processes of care and outcomes of care showed a
beneficial effect of employing one or more elements of the CCM. The effects on quality of life were mixed with only CHF and depression studies showing benefit (Tsai et al., 2005).

**CCM collaborative evaluations.** As of 2009, various components of the CCM have been applied to over 1500 physician practices in the United States and internationally (Coleman et al., 2009). The CCM is also an integral part of the patient-centered medical homes (PCMH) model that is currently being tested across the country (Berenson et al., 2008). Three major national collaboratives, Breakthrough Series (BTS), the Health Resources and Services Administration’s (HRSA) Health Disparities Collaboratives (HDCs), and the Improving Chronic Illness Care (ICIC) Collaboratives, have been the dominant method used to help practices implement the CCM (Coleman et al., 2009).

A RAND study surveyed 52 practices that were participating in one of the above collaboratives to determine if they were able to implement and sustain elements of the CCM. All of them were able implement elements of the CCM, while 39 of the 52 practices were able to sustain those changes at least one year (Pearson et al., 2005).

When assessed for evidence of improving care for specific diseases, the CCM guidelines appear to be improving outcomes. A study of practices participating in a congestive heart failure (CHF) collaborative visited the emergency department (ED) less often and experienced 35% fewer days in the hospital (Asch, 2005). When practices redesigned asthma care according to the CCM, patients with asthma were more likely to have an asthma action plan and show an improvement on quality of life questionnaires (Mangione-Smith et al., 2005). Finally, another study showed that patients with diabetes seen at practices participating in CCM collaboratives experienced reduced risk of cardiovascular disease (Vargas et al., 2007).

There have been three large evaluations of HRSA’s Health Disparities Collaboratives. Chin (2004) studied 19 community health centers (CHCs) that were involved in diabetes collaboratives. They found significant improvements in the processes of care, but not in outcomes of care. Landon et al. (2007) studied another sample of CHCs that were involved with diabetes and asthma collaboratives. They also found improvements in the processes of care but not the outcomes of care. Suspecting that outcomes of
care were not improving because the follow-up period was too short (one year), Chin (2007) studied another sample of CHCs with a one-year follow up and a three-year follow up. The one-year follow up showed improvements in the processes of care while the three-year follow-up showed improvements in the outcomes of care (such as decreases in HbA1c and low-density lipoprotein levels).

In general, evaluations of CCM programs have shown that high-performing practices make changes across multiple elements of the CCM rather than implementing just one element of the CCM. Also, most of the studies involve highly motivated practices focusing on patients with a single chronic condition. Evidence to date gives no assurance that practices changes spread to the care of other illnesses or to other less motivated practices in an organization (Coleman et al., 2009).

Other applications of the CCM. In 2004, the World Health Organization (WHO) joined with the MacColl Institute for Health Innovation to adapt the CCM from a global perspective resulting in the Innovative Care for Chronic Conditions (ICCC) framework (Epping-Jordan, 2004). (See Figure 2.) The framework includes micro components (patient and family), meso components (healthcare organization and community), and a macro component (policy). The ICCC framework works similar to the CCM with the added feature of emphasizing close community networks and international policy aspects of improving chronic illness care (Epping-Jordan, 2004).
Kaiser Permanente in Washington DC introduced the Guided Care model which is based on the CCM (Boult, Karm, and Groves, 2008). Guide Care is facilitated by a nurse in a primary care office who assists 3-4 physicians in providing care for patients with chronic illnesses. For each patient, the Guided Care nurse provides assessment, a care plan, on-going monitoring, coaching, self-management education, supporting caregiver education, coordination of transitions between sites of care, and access to community services. Early positive outcomes of a pilot study indicate that this model may be an effective way to integrate the CCM into physician practices (Boult et al., 2008).

**CCM vs. disease management.** Two models currently dominate the discussion about how to improve care for patients with chronic illness: the physician-based chronic care model and the health insurance-based disease management model (Casalino, 2005). The components of each model are similar to each other but their primary strategies differ significantly. The strategy of the CCM model is to reorganize physician practices. The strategy of the disease management model is to bypass physician
practices and employ health (insurance) plans which work directly with patients via telephone, internet, or other distance media (Leeman & Mark, 2006).

Both approaches to chronic illness care have strengths and weaknesses. Disease management programs usually have a more developed clinical information system which provides extensive patient data. They also have the ability to dedicate staff to specific tasks which increases competence and consistency. A restructured physician practice using the CCM has the advantage of integrating chronic illness care with other care provided by the provider team which can enhance provider-patient relationships. The provider team is also more likely to be familiar with the community and can link patients to community resources more effectively (Leeman & Mark, 2006).

Health plans could choose to employ the CCM. However, health plans are more often choosing to invest in disease management programs rather than trying to change care in a physician practice using the CCM. Leeman and Mark (2006) propose that increases in transaction costs are driving health plans to adopt the disease management model rather than the CCM model. To implement the CCM, both the health plan and the physician practice must invest in the process. These investments include the adoption of evidence-based protocols, the implementation of clinical information systems, training of staff as care managers or self-management educators, and changes in the contractual relationship to ensure compliance with the CCM approach (Leeman & Mark, 2006).

One factor that has increased interest in the disease management model in the last 15 years has been aggressive marketing by a growing disease management industry seeking profits in a new market. The pharmaceutical industry has led this development since the early 1990s (Geyman, 2007). Although these programs are marketed as strategies to contain costs and improve the quality of care of patients with chronic diseases (such as diabetes or asthma), pharmaceutical manufacturers stand to profit since disease management programs promote their own product lines. Increased sales can be leveraged by pharmacy benefit management companies contracting with employers and managed care organizations, and patients not yet taking medications can be identified (Bodenheimer, 2003).
Two thirds of employers with 200 or more employees in 2005 participated in programs based on the disease management model. More than one half of all workers with employment-based insurance are enrolled in disease management programs. In that same year on the public sector side, more than 20 states contracted for disease management programs for their Medicaid enrollees (Wheatley, 2002), and Medicare had signed up 100,000 beneficiaries for disease management programs (Landro, 2006).

The movement towards disease management programs and away from the CCM is contrary to the healthcare delivery redesign envisioned by the Institute of Medicine (CQHCA, 2001). Yet, the expanding disease management industry has emerged to address the poor quality of chronic disease care in many primary care settings across the country. Many factors account for this problem, including inadequate design of office practice for team management of chronic illness, under-reimbursement of chronic disease care, lack of time, and lack of enabling information technology (Casalino, 2005).

The evidence for the effectiveness of disease management programs is limited. They may actually further fragment healthcare since they bypass the physician and since they typically target only a subset of patients (Leeman & Mark, 2006). And there is no solid evidence yet that commercial for-profit disease management vendors will save money and improve care of chronic illness on a long-term basis. But, evaluations of disease management programs are methodologically challenging since it is difficult to create a control group or collect data on longer-term outcomes. To date, there are no studies that directly compare the outcomes of CCM programs integrated in primary care settings with the outcomes of commercial disease management programs (Casalino, 2005).

To encourage third-party payers and primary care providers to choose the CCM over the disease management model, the Group Health Cooperative has taken its expertise and lessons on the road in an effort to train primary care teams in other parts of the country. After 13 months of training and collaboration, 82% of the pilot sites reported decreases in patients’ average HbA1c levels in diabetes from 8.4% to 7.6%, and combined cardiac risk reduction scores showed an absolute risk reduction of 2.4% (McCulloch, Davis, Austin, and Wagner, 2004). These results require commitment, expertise, and adequate funding, together with an infrastructure that includes electronic medical records, disease
registries, decision-support systems, patient reminders, and self-management materials. Unfortunately, those needs are not available in most primary care practices across the country, which are already under-reimbursed by private and public payers for the challenge of improving quality of chronic illness care (Geyman, 2007).

According to the CCM, optimal management of chronic conditions is best done by well-trained primary care physicians working closely with other health professionals on a team basis. Management decisions are often difficult and must be individualized to each patient and family in a continuity of care relationship. With a strong link to primary care, quality of care can be improved, but cost savings cannot be assumed. Costs may actually increase as better quality of care is provided to patients previously undertreated for chronic conditions (Geyman, 2007).

As long as costs drive the health care bus, it is unlikely that the CCM will be adopted over disease management models. Changes in public health policy are needed to help health care decision makers and health care professionals adopt the CCM and bring about the paradigm shift needed to fully meet the needs of patients with chronic illness. One example of such a public policy change are the following proposals by Ham (2010) to create a high-performing chronic care system (See Table 2.)

<table>
<thead>
<tr>
<th>Table 2. Proposals Creating a High-Performing Chronic Care System</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Universal coverage is necessary for without universal coverage it is difficult to act consistently on the other characteristics</td>
</tr>
<tr>
<td>2. Care should be free at the point of use or at least provided at a cost that does not act as a major deterrent to sick patients seeking medical help</td>
</tr>
<tr>
<td>3. Care should be focused on the prevention of ill health and not just the treatment of sickness</td>
</tr>
<tr>
<td>4. Priority should be given to patients to self-manage their conditions with support from caregivers and families</td>
</tr>
<tr>
<td>5. Priority should be given to primary health care since there is evidence of the value of multidisciplinary team work that is most likely in a primary care setting</td>
</tr>
<tr>
<td>6. Population management should be emphasized, stratifying people with chronic diseases according to their risk and offering support commensurate with this risk</td>
</tr>
<tr>
<td>7. Primary health care teams should be able to access specialist advice and support when needed</td>
</tr>
<tr>
<td>8. Maximize the benefit of information technology by enabling primary health care teams to develop disease registers and stratify the population according to risk</td>
</tr>
<tr>
<td>9. Coordination of care from site to site is necessary for people with multiple conditions who are much more likely to be in transition between chronic care and acute care</td>
</tr>
<tr>
<td>10. Apply all nine characteristics as a whole since it is the cumulative effect that explains the impact of the CCM rather than any of its individual elements</td>
</tr>
</tbody>
</table>
Implementation Science

As defined by the Annual NIH Conference on Implementation and Dissemination (Madon et al., 2007, p. 1728), implementation is “the use of strategies to adopt and integrate evidence-based health interventions and change practice patterns within specific settings.” Implementation science seeks to “understand the behavior of healthcare professionals and other stakeholders as a key variable in the sustainable uptake, adoption, and implementation of evidence-based interventions” (Madon et al., 2007, p. 1728).

Malveaux and Butterfoss (2011) state that the ill-defined lexicon that describes implementation science is a significant issue. In a recent literature review of articles that were written in 12 different health care journals in 2006, one hundred different terms were identified as being equivalent or closely related to implementation science (Malveaux & Butterfoss, 2011).

For the sake of clarity, the terminology set forth by (Fleisher, Goldstein, & Schillinger, 2010) will be used in this discussion about implementation science (IS), which lies in the pathway of harnessing science to promote health. (See Figure 3.) Scientific development for human health has been conceived as occurring through five steps: (1) basic research, (2) treatment development, (3) efficacy studies, (4) effectiveness studies, and (5) adaptation to the real world. Basic research, treatment development, and efficacy studies take place within the T1 loop in pathway in Figure 3. Effectiveness studies and guideline development occur within the T2 loop, and studies that facilitate adaptation to the real world occur within the T3 loop. Fleisher et al. (2010) define efficacy as the intervention’s ability to do more good than harm among the target population in an ideal setting, and effectiveness as the intervention’s ability to do more good than harm for the target population in a real world setting. The T3 loop is that juncture in the healthcare research pathway in Figure 3 where the discover-delivery gap occurs and where implementation science (i.e. implementation research) is employed.
Grimshaw and Eccles (2004) did a systematic review of 235 evaluations of different guideline dissemination and implementation strategies published up to 1998. The results of the review revealed that there was a median 10% improvement across studies, suggesting that it is possible to improve healthcare provider behavior and the quality of care. The review also showed that most dissemination and implementation strategies resulted in small to moderate improvements in care (Grimshaw & Eccles, 2004). The authors of the study concluded that establishing an empirically tested theoretical base for healthcare professional and organizational behavior would facilitate more effective interventions (Grimshaw & Eccles, 2004).

The problem of translating healthcare research into practice became evident in the mid-1990s as evidence-based practice guidelines were being developed with the aim of improving patient-oriented outcomes (Green & Seifert, 2005). It became evident that physicians and health care systems were not putting new knowledge into practice nearly quickly enough. As implementation science developed, the translation of research into clinical practice was conceptualized as proceeding from awareness to acceptance to adoption. Green and Seifert (2005) proposed that improving the translation of healthcare
research into practice requires a detailed understanding of how adoption (i.e. the T3 loop in Figure 2.2) takes place.

On average, just over half of recommended health care practices are implemented, and the situation may be even worse for health behavior change interventions (Glasgow & Emmons, 2007). Minority populations are at even greater disadvantage; receiving evidence-based care only 35% or less of the time. Multiple reasons can be given for the failure of health research findings to translate into practice, including historical, political, social, economic, scientific, cultural, and organization factors that slow or impede transfer of research into practice (Glasgow & Emmons, 2007).

Implementation science is inherently complex (Ohadike, Malveaux, Lesch, 2011). It requires a systematic assessment of clinical process and outcomes and a systematic assessment of the effect of contextual factors (political and professional, economic, social, organizational, attitudes and behavior of local stakeholders) (Madon et al., 2007). Understanding this complexity may require more methodology than what research methodology has to offer.

Much research fails to translate into practice because the programs being implemented fail to address these contextual factors (Glasgow & Emmons, 2007). When contextual factors are considered, it is from a limited and often researcher-centric perspective. Multiple stakeholder perspectives are not taken into consideration. Evaluation methods are needed to integrate and synthesize contextual and external validity evidences from different perspectives that can aid both local decision makers and policy-making bodies (Glasgow & Emmons, 2007). The decisions these two groups need to make about program adoption are different. Local decision makers are interested in whether evidence is relevant and will fit their situation. Policy makers are concerned with the generalizability of evidence—that is, the breadth of conditions across which this evidence will apply. If health researchers can develop and evaluate programs with greater attention to context and external validity and in partnership with relevant decision makers and target audiences, it will be much easier for both local practitioners and policy makers to judge program relevance (Glasgow & Emmons, 2007).
Beginning with diffusion theory (Dearing, 2008), Green, Ottoson, Garcia, and Hiatt (2009) reviewed the concepts of dissemination and implementation as applied in the realm of public health. As a result of this review, the authors suggested ways to provide a more decentralized approach to dissemination and implementation in public health (Green et al. 2009).

- Give greater attention to external validity in the systematic reviews of evidence for adoption by policy makers, practitioners, or the public
- Research beyond efficacy trials needs to be more practice based, outside highly controlled academic circumstances, to be more relevant, believable, and actionable to practitioners
- Engage practitioners or policy makers or community-based residents or patients to help shape the research questions and methods for sampling, design, analysis, and interpretation
- Conduct evaluations of actual public health programs in real-time, typical situations, with typical personnel conducting the interventions (and participating in the evaluations)
- Utilize enhanced surveillance systems to track comparable data over time and between jurisdictions to make public health program evaluation more feasible and useful beyond the particular program being evaluated

Green and Glasgow (2006) state that much of the research upon which clinical practice guidelines have been based is strong on internal validity, thanks to the emphasis that has been given to experimental control. However, they contend that this research is weak on external validity because the varied circumstances of practice have not been taken into account. The authors provided two conclusions based on their observations of guidelines implementation: that more emphasis needs to be placed on developing criteria and measures of external validity, and that the healthcare profession needs to find ways to generate more practice-based evidence that addresses external validity and local realities, stating “if we want more evidence-based practice, we need more practice-based evidence” (Green & Glasgow, 2006).

The RE-AIM (reach, effectiveness, adoption, implementation, and maintenance) model, developed by Glasgow, Vogt, and Boles (1999), was adapted from diffusion theory (Rogers, 2003) and
Program developers are often primarily concerned with the fidelity with which their intervention protocols are translated into practice (Green & Glasgow, 2006). However, it is well established that program adopters seldom adopt or implement a program exactly as it was originally tested. This is not only pervasive but also generally desirable. The balance between these two opposing criteria of complete fidelity and complete adaptation to local settings, clientele, resources, and priorities is unknown. This balance is more difficult to achieve when the intervention is a complex program made up of many discrete interventions, such as the full range of interventions required to predispose, enable, and
reinforce a set of behavioral and environmental determinants of a specific health outcome as in the chronic care model (Green & Glasgow, 2006).

A measurement approach capable of providing critical information about implementation is formative evaluation (FE). Formative evaluation is an assessment process designed to identify influences on the progress of implementation efforts (Stetler et al., 2006). Formative evaluation can be used to study the complexity of implementation projects and suggests ways to answer questions about context, adaptations, and response to change. The Department of Veterans Affairs (VA) Quality Enhancement Research Initiative (QUERI) has integrated FE into its implementation program (Stetler et al., 2006). QUERI, begun in 1998, is an improvement initiative that focuses on implementation of empirically based practices and on the evaluation and refinement of these implementation efforts.

Research methods such as experimental and quasi-experimental approaches provide summative data on the effectiveness of a program. Summative data are essential but insufficient to meet the needs of implementation science (Stetler et al., 2006). Evaluative information is needed beyond whether a chosen program worked. Implementation studies seek to answer critical questions about the (1) feasibility of implementation strategies, (2) degree of real-time implementation, (3) status and potential influence of contextual factors, (4) response of project participants, and (5) any adaptations necessary to achieve optimal change. Formative evaluation provides techniques for obtaining such information and for overcoming limitations identified in early implementation. As Stead, Hastings, and Eadie (2002) suggest, traditional intervention research can fail to “capture the detail and complexity of intervention inputs and tactics,” thereby missing the true nature of interventions as well as significant organizational factors important for replication. However, the complementary use of FE within an experimental study can create a useful hybrid style approach for implementation research (Stetler et al., 2002).

In 2004, the Royal College of Nursing (RCN) Institute in the United Kingdom published a PARiHS framework (Promoting Action on Research Implementation in Health Services) to facilitate implementation science (Kitson et al., 2008). This framework uses a two-stage process whereby evaluation is used to measure activities and outcomes and context, and then the results of the evaluation is used to determine the most appropriate facilitation method for the program. The exact nature of the
intervention is thus determined by the specific stakeholders in the specific context at a specific time and place. PARiHS thus has the potential to be used as a practical and pragmatic tool by practitioners and researchers at the local level (Kitson et al., 2008). The purpose of the PARiHS framework is to provide a map to enable others to make sense of the complexity of implementation, and the elements that require attention if implementation is more likely to be successful. The PARiHS conceptual framework may also lend itself to guiding the development of evaluative approaches and instruments, which could be used by both researchers and practitioners (Kitson et al., 2008).

In 2009, a research group from the University of Michigan proposed the Consolidated Framework for Implementation Research (CFIR) to promote implementation theory development and verification about what works where and why across multiple contexts (Damschroder et al., 2009). The group began with a research synthesis by Greenhalgh, Robert, Macfarlane, Bate, and Kyriakidou (2004) of nearly 500 published sources across 13 fields of research which considered the determinants of diffusion, dissemination, and implementation of innovations in health service delivery. They included theories related to dissemination, innovation, organizational change, implementation, knowledge translation, and research uptake that have been published in peer reviewed journals, including the Quality Enhancement Research Initiative (QUERI) (Stetler et al., 2006) and the Promoting Action on Research Implementation in Health Services (PARiHS) framework (Kitson et al., 2008).

The CFIR comprises five major domains. The first major domain is the intervention being implemented into a particular organization. The next two domains are inner and outer setting within which the intervention takes place. The outer setting includes the economic, political, and social context within which an organization resides, and the inner setting includes structural, political, and cultural contexts through which the implementation process takes place. The fourth domain is the individuals involved with the intervention and/or implementation process. The fifth domain is the actual implementation process. These five domains of the CFIR’s overarching structure support the exploration of essential factors that may be encountered during implementation through formative evaluations (Damschroder et al., 2009).
Other hybrid designs simultaneously evaluate the effectiveness and implementation of a program (Bernet, Willens, & Bauer, 2013). Conducting both evaluations simultaneously enhances the ability to identify important intervention-implementation interactions, which inform decisions about optimal deployment and generalized impact. Hybrid evaluation designs may serve summative or formative purposes. Formative evaluation makes use of data throughout the intervention trial to modify the intervention procedures or implementation process during the study. Formative evaluation can be used to refine and improve the clinical intervention and implementation process while under study. Summative evaluation provides information about the impact of the intervention, similar to classical clinical trials. The summative evaluation outcomes may include patient level health outcomes for a clinical intervention, population-level health status, or an index of system function for an organizational-level intervention (Bernet et al., 2013).

Finally, Glasgow (2013) states that there is an increased need for healthcare studies to be pragmatic. At the heart of pragmatic approaches to health research is keeping the focus on issues and data relevant for making decisions and taking action. The overall goal of pragmatic approaches is to produce results that are both rigorous and relevant to stakeholders. Pragmatic studies are different than efficacy studies in that: (1) the questions and outcomes studied are important to stakeholders such as policy makers, practitioners, and patients; (2) studies are conducted in settings similar to those in practice; and (3) comparison conditions are real-world alternatives; for example, a current standard of care rather than no treatment or placebo controls (Glasgow, 2013).
CHAPTER III

METHODOLOGY

Research Design

The overall design of this dissertation is a non-experimental cross-sectional design. The study procedure consists of (1) the selection of CCM studies using specific exclusion criteria, (2) the metaevaluation of each study using the accuracy standards of the Program Evaluations Metaevaluation Checklist (Stufflebeam, 2011), and (3) the processing and analysis of data to answer the research questions above.

Selection of Studies

Since this study is building on the work of Gugiu et al. (2013), the articles used in their study will also be used in this study. This consistency is necessary to answer the second research question above which is a comparison of the HEAL grading approach to assessing study quality to the metaevaluation approach to assessing study quality.

There are 28 CCM studies included in this study. These studies were identified through systematic searches of MEDLINE and several social sciences databases for studies published on or after 1996 that included the search terms chronic care model or CCM. All articles found were then screened based on a set of exclusion criteria (see Table 4). The studies that remained following the screening process cover a variety of chronic illnesses, employ multiple CCM elements, and measure numerous clinical outcomes. In Table 5, a breakdown of the 28 CCM studies by research design is given. A bibliography of the 28 CCM studies is provided at the end of this report.
Table 4. Exclusion Criteria for Relevant CCM Studies

- Described or referenced search term only
- Provided descriptive results only
- Measured implementation of CCM only
- Statistically modeled CCM but did not test for effectiveness
- Provided only a literature review

Table 5. Breakdown of CCM Studies by Research Design

- Randomized controlled trials: 8
- Quasi-experimental with matched control: 7
- Quasi-experimental with nonequivalent control: 4
- Pretest-posttest: 9

Instrumentation

The metaevaluations of the selected 28 CCM studies have been conducted using the accuracy standards portion of the Program Evaluations Metaevaluation Checklist (Stufflebeam, 2011b). There are eight accuracy standards that have been created by the Joint Committee on Standards for Educational Evaluation (Yarbrough et al, 2011). Each accuracy standard is broken down into six declarative statements (checkpoints) in the Checklist. Each checkpoint refers to a desired element of the accuracy standard. If an element was present in a CCM study, the rater marked “Yes” by the checkpoint. If the element was not present in the evaluation, the rater marked “No” by the checkpoint. If there was insufficient information to make a judgment, the rater marked “???” by the checkpoint. This process simplifies metaevaluation by reducing rating to a series of dichotomous scale measures. See Figure 4 for an example of one accuracy standard represented in the Checklist.
If all checkpoints within an accuracy standard are marked “Yes,” that standard is assigned a rating of “Excellent”. The rating of the accuracy standard drops from “Excellent” to “Very Good” to “Good” to “Fair” to “Poor” as fewer checkpoints are marked “Yes.” The rating for each accuracy standard is then weighted, with the number of “Excellent” ratings multiplied by 4, the number of “Very Good” ratings multiplied by 3, the number of “Good” ratings multiplied by 2, and the number of “Fair” ratings multiplied by 1. The weighted scores are then summed to obtain a total score for all eight accuracy standards in the Checklist. The total accuracy score is divided by the maximum number of points possible and then multiplied by 100 to arrive at a percentage score. See Figure 5 for an example of this analysis.
Data Processing and Analysis

Research Question #1

What is the accuracy metaevaluation rating (Stufflebeam, 2011b) for each CCM study selected for this dissertation?

To answer research question #1, each of the CCM studies was evaluated using the accuracy standards portion of the Program Evaluations Metaevaluation Checklist. Two raters were used to conduct the metaevaluations of the selected CCM studies. After a trial metaevaluation of three studies and a preliminary assessment of inter-rater reliability, decision rules for subsequent metaevaluations were agreed upon, including (1) all CCM studies are to be metaevaluated against one accuracy standard at a time before proceeding to the next accuracy standard, (2) mark “Yes” for a checkpoint if one or more accuracy criterion in the checkpoint is present in the CCM study report, (3) mark “No” for a checkpoint if it is clear from the CCM study report that the accuracy criterion was not addressed, and (4) mark “???” for a checkpoint if it is not clear from the CCM study report if the accuracy criterion was addressed. Using these decision rules, the inter-rater agreement between the two raters was increased to an inter-rater reliability coefficient of .93.

Research Question #2

What is the relationship between the accuracy metaevaluation rating (Stufflebeam, 2011b) and the HEAL grade (Gugiu et al., 2013) of each CCM study selected for this dissertation?

To answer research question #2, the accuracy metaevaluation ratings for each of the 28 CCM studies was compared to the HEAL grade for each of the 28 CCM studies. The HEAL grading system is discussed in detail in Chapter 2 of this dissertation. The HEAL grades for each of the 28 CCM studies can be found in Table 2.1 in Chapter 2. The criteria used by the Hierarchy of Evidence and Study Limitations (HEAL) grading system to assess study quality are provided in Appendix A.

A Spearman’s rank-order correlation was calculated to assess the strength of association between accuracy metaevaluation ratings and HEAL grades for each of the 28 CCM studies. To assure
statistical significance at the .05 level of significance given $n = 28$, the coefficient above which a meaningful relationship can be assumed is $\rho = .38$. 
CHAPTER IV

RESULTS

As stated in Chapter 1, the purpose of this dissertation is to explore the use of metaevaluation to assess the quality of healthcare studies. The study design is a non-experimental cross-sectional design to answer two research questions related to the stated purpose of the dissertation. The results pertaining to the two research questions are presented below.

Results Pertaining to Research Question #1

What is the accuracy metaevaluation rating (Stufflebeam, 2011b) for each CCM study selected for this dissertation?

Table 6 contains the author, date, and study title of the selected CCM studies as well as the metaevaluation rating for overall accuracy for each CCM study. Table 6 also contains the metaevaluation ratings for each of the eight accuracy standards for the selected CCM studies. A study was rated Excellent for a particular accuracy standard if all six items pertaining to the standard were marked “Yes.” A study was rated Very Good per standard if five items were affirmative, rated Good if four items were affirmative, rated Fair if two or three items were affirmative, and rated Poor if only one or no items were affirmative.
<table>
<thead>
<tr>
<th>Study</th>
<th>Title</th>
<th>Accuracy Rating</th>
<th>(continued)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Very Good</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fair</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Excellent</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Poor</td>
</tr>
</tbody>
</table>

(continued)
<table>
<thead>
<tr>
<th>Author/Date</th>
<th>CCM Study Title</th>
<th>Metaevaluation Accuracy Standard Ratings</th>
</tr>
</thead>
</table>
| 4. Chin et al. (2004) | Improving diabetes care in Midwest community health centers with the health disparities collaborative | Overall Accuracy Rating  
 |                     |                                                                                 | A1 Justified Conclusions and Decisions  
 |                     |                                                                                 | A2 Valid Information  
 |                     |                                                                                 | A3 Reliable Information  
 |                     |                                                                                 | A4 Explicit Program and Context Descriptions  
 |                     |                                                                                 | A5 Information Management  
 |                     |                                                                                 | A6 Sound Designs and Analyses  
 |                     |                                                                                 | A7 Explicit Evaluation Reasoning  
 |                     |                                                                                 | A8 Communicating and Reporting  
 |                     |                                                                                 | Fair  
 |                     |                                                                                 | Very Good  

| 5. Chin et al. (2013) | Diabetes care in community and sustaining improving with the health disparities collaboratives health centers | Overall Accuracy Rating  
 |                     |                                                                                 | A1 Justified Conclusions and Decisions  
 |                     |                                                                                 | A2 Valid Information  
 |                     |                                                                                 | A3 Reliable Information  
 |                     |                                                                                 | A4 Explicit Program and Context Descriptions  
 |                     |                                                                                 | A5 Information Management  
 |                     |                                                                                 | A6 Sound Designs and Analyses  
 |                     |                                                                                 | A7 Explicit Evaluation Reasoning  
 |                     |                                                                                 | A8 Communicating and Reporting  
 |                     |                                                                                 | Fair  
 |                     |                                                                                 | Very Good  

 |                     |                                                                                 | A1 Justified Conclusions and Decisions  
 |                     |                                                                                 | A2 Valid Information  
 |                     |                                                                                 | A3 Reliable Information  
 |                     |                                                                                 | A4 Explicit Program and Context Descriptions  
 |                     |                                                                                 | A5 Information Management  
 |                     |                                                                                 | A6 Sound Designs and Analyses  
 |                     |                                                                                 | A7 Explicit Evaluation Reasoning  
 |                     |                                                                                 | A8 Communicating and Reporting  
 |                     |                                                                                 | Fair  
 |                     |                                                                                 | Good  

(continued)
<table>
<thead>
<tr>
<th>Author/Date</th>
<th>CCM Study Title</th>
<th>Metaevaluation Accuracy Standard Ratings</th>
</tr>
</thead>
</table>

(continued)
<table>
<thead>
<tr>
<th>Author/Date</th>
<th>CCM Study Title</th>
<th>Metaevaluation</th>
<th>Accuracy Standard Ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>10. Glasgow et al. (2005)</td>
<td>Randomized effectiveness trial of a computer-assisted intervention to improve diabetes care</td>
<td>Overall Accuracy Rating</td>
<td>Good</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A1 Justified Conclusions and Decisions</td>
<td>Excellent</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A2 Valid Information</td>
<td>Good</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A3 Reliable Information</td>
<td>Fair</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A4 Explicit Program and Context Descriptions</td>
<td>Excellent</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A5 Information Management</td>
<td>Poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A6 Sound Designs and Analyses</td>
<td>Very Good</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A7 Explicit Evaluation Reasoning</td>
<td>Good</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A8 Communicating and Reporting</td>
<td>Poor</td>
</tr>
<tr>
<td>11. Green et al. (2008)</td>
<td>Effectiveness of home blood pressure monitoring, web communication, and pharmacist care on hypertension control</td>
<td>Overall Accuracy Rating</td>
<td>Fair</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A1 Justified Conclusions and Decisions</td>
<td>Very Good</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A2 Valid Information</td>
<td>Good</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A3 Reliable Information</td>
<td>Poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A4 Explicit Program and Context Descriptions</td>
<td>Excellent</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A5 Information Management</td>
<td>Poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A6 Sound Designs and Analyses</td>
<td>Very Good</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A7 Explicit Evaluation Reasoning</td>
<td>Fair</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A8 Communicating and Reporting</td>
<td>Poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A1 Justified Conclusions and Decisions</td>
<td>Good</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A2 Valid Information</td>
<td>Fair</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A3 Reliable Information</td>
<td>Poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A4 Explicit Program and Context Descriptions</td>
<td>Very Good</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A5 Information Management</td>
<td>Poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A6 Sound Designs and Analyses</td>
<td>Fair</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A7 Explicit Evaluation Reasoning</td>
<td>Poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A8 Communicating and Reporting</td>
<td>(continued)</td>
</tr>
<tr>
<td>Author/ Date</td>
<td>CCM Study Title</td>
<td>Metaevaluation Accuracy Standard Ratings</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>-----------------</td>
<td>------------------------------------------</td>
<td></td>
</tr>
</tbody>
</table>
| 13. Kirsh et al. (2007) | Shared medical appointments based on the chronic care model: a quality improvement project to address the challenges of patients with diabetes with high cardiovascular risk | Overall Accuracy Rating  
A1 Justified Conclusions and Decisions  
A2 Valid Information  
A3 Reliable Information  
A4 Explicit Program and Context Descriptions  
A5 Information Management  
A6 Sound Designs and Analyses  
A7 Explicit Evaluation Reasoning  
A8 Communicating and Reporting |
A1 Justified Conclusions and Decisions  
A2 Valid Information  
A3 Reliable Information  
A4 Explicit Program and Context Descriptions  
A5 Information Management  
A6 Sound Designs and Analyses  
A7 Explicit Evaluation Reasoning  
A8 Communicating and Reporting |
A1 Justified Conclusions and Decisions  
A2 Valid Information  
A3 Reliable Information  
A4 Explicit Program and Context Descriptions  
A5 Information Management  
A6 Sound Designs and Analyses  
A7 Explicit Evaluation Reasoning  
A8 Communicating and Reporting |
<table>
<thead>
<tr>
<th>Author/ Date</th>
<th>CCM Study Title</th>
<th>Metaevaluation Accuracy Standard Ratings</th>
</tr>
</thead>
</table>
A1 Justified Conclusions and Decisions: Good  
A2 Valid Information: Fair  
A3 Reliable Information: Fair  
A4 Explicit Program and Context Descriptions: Very Good  
A5 Information Management: Poor  
A6 Sound Designs and Analyses: Good  
A7 Explicit Evaluation Reasoning: Poor  
A8 Communicating and Reporting: Poor |
| 17. Piatt et al. (2006) | Translating the chronic care model into the community | Overall Accuracy Rating: Fair  
A1 Justified Conclusions and Decisions: Good  
A2 Valid Information: Fair  
A3 Reliable Information: Poor  
A4 Explicit Program and Context Descriptions: Very Good  
A5 Information Management: Poor  
A6 Sound Designs and Analyses: Very Good  
A7 Explicit Evaluation Reasoning: Fair  
A8 Communicating and Reporting: Poor |
A1 Justified Conclusions and Decisions: Good  
A2 Valid Information: Fair  
A3 Reliable Information: Poor  
A4 Explicit Program and Context Descriptions: Excellent  
A5 Information Management: Poor  
A6 Sound Designs and Analyses: Good  
A7 Explicit Evaluation Reasoning: Poor  
A8 Communicating and Reporting: Poor |

(continued)
<table>
<thead>
<tr>
<th>Author/Date</th>
<th>CCM Study Title</th>
<th>Metaevaluation Accuracy Standard Ratings</th>
</tr>
</thead>
</table>
A1 Justified Conclusions and Decisions: Fair  
A2 Valid Information: Good  
A3 Reliable Information: Fair  
A4 Explicit Program and Context Descriptions: Poor  
A5 Information Management: Very Good  
A6 Sound Designs and Analyses: Fair  
A7 Explicit Evaluation Reasoning: Very Good  
A8 Communicating and Reporting: Poor |
A1 Justified Conclusions and Decisions: Fair  
A2 Valid Information: Good  
A3 Reliable Information: Poor  
A4 Explicit Program and Context Descriptions: Very Good  
A5 Information Management: Poor  
A6 Sound Designs and Analyses: Very Good  
A7 Explicit Evaluation Reasoning: Good  
A8 Communicating and Reporting: Poor |
A1 Justified Conclusions and Decisions: Very Good  
A2 Valid Information: Good  
A3 Reliable Information: Fair  
A4 Explicit Program and Context Descriptions: Very Good  
A5 Information Management: Fair  
A6 Sound Designs and Analyses: Very Good  
A7 Explicit Evaluation Reasoning: Very Good  
A8 Communicating and Reporting: Poor |
<table>
<thead>
<tr>
<th>Author/Date</th>
<th>CCM Study</th>
<th>Metaevaluation Accuracy Standard Ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Overall Accuracy Rating</td>
</tr>
<tr>
<td>22. Siminerio et al. (2006)</td>
<td>Deploying the chronic care model to implement and sustain diabetes self-management training programs</td>
<td>Fair</td>
</tr>
<tr>
<td>23. Siminerio et al. (2005)</td>
<td>Implementing the chronic care model for improvements in diabetes care and education in a rural primary care practice</td>
<td>Fair</td>
</tr>
<tr>
<td>24. Stock et al. (2008)</td>
<td>Developing a senior healthcare practice using the chronic care model: effect on physical function and health-related quality of life</td>
<td>Good</td>
</tr>
<tr>
<td>Author/Date</td>
<td>CCM Study Title</td>
<td>Metaevaluation Accuracy Standard Ratings</td>
</tr>
<tr>
<td>-------------</td>
<td>-----------------</td>
<td>------------------------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A1 Justified Conclusions and Decisions: Good</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A2 Valid Information: Fair</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A3 Reliable Information: Poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A4 Explicit Program and Context Descriptions: Very Good</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A5 Information Management: Fair</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A6 Sound Designs and Analyses: Very Good</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A7 Explicit Evaluation Reasoning: Fair</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A8 Communicating and Reporting: Poor</td>
</tr>
<tr>
<td>26. Stroebel et al. (2005)</td>
<td>Adapting the chronic care model to treat chronic illness at a free medical clinic</td>
<td>Overall Accuracy Rating: Poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A1 Justified Conclusions and Decisions: Poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A2 Valid Information: Poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A3 Reliable Information: Poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A4 Explicit Program and Context Descriptions: Good</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A5 Information Management: Poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A6 Sound Designs and Analyses: Good</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A7 Explicit Evaluation Reasoning: Poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A8 Communicating and Reporting: Poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A1 Justified Conclusions and Decisions: Good</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A2 Valid Information: Fair</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A3 Reliable Information: Poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A4 Explicit Program and Context Descriptions: Very Good</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A5 Information Management: Fair</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A6 Sound Designs and Analyses: Very Good</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A7 Explicit Evaluation Reasoning: Good</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A8 Communicating and Reporting: Poor</td>
</tr>
</tbody>
</table>

(continued)
<table>
<thead>
<tr>
<th>Author/Date</th>
<th>CCM Study Title</th>
<th>Metaevaluation Accuracy Standard Ratings</th>
</tr>
</thead>
</table>
Results Pertaining to Research Question #2

What is the relationship between the accuracy metaevaluation rating (Stufflebeam, 2011b) and the HEAL grade (Gugiu et al., 2013) of each CCM study selected for this dissertation?

Table 7 contains the correlations between the accuracy metaevaluation ratings in Table 6 and the HEAL grades in Table 1 (see Chapter 2). As stated in Chapter 3, the coefficient above which a meaningful relationship can be assumed between accuracy metaevaluation ratings and HEAL grades is \( p = .38 \). All of the correlations listed in Table 7 are below this benchmark. Thus, none of the correlations between accuracy metaevaluation ratings and HEAL grades are considered to be statistically significant.
## Table 7: Correlations between Metaevaluation Ratings and HEAL Grades for Selected CCM Studies

<table>
<thead>
<tr>
<th>Accuracy Standards</th>
<th>Accuracy Standard Description</th>
<th>MEV Rating to HEAL Grade Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Overall</strong></td>
<td>Evaluation Accuracy Standards</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>The accuracy standards are intended to increase the dependability and truthfulness of evaluation representations, propositions, and findings, especially those that support interpretations and judgments about quality</td>
<td></td>
</tr>
<tr>
<td>A1</td>
<td>Justified Conclusions and Decisions</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>Evaluation conclusions and decisions should be explicitly justified in the cultures and contexts where they have consequences</td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>Valid Information</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>Evaluation information should serve the intended purposes and support valid interpretations</td>
<td></td>
</tr>
<tr>
<td>A3</td>
<td>Reliable Information</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>Evaluation procedures should yield sufficiently dependable and consistent information for the intended uses</td>
<td></td>
</tr>
<tr>
<td>A4</td>
<td>Explicit Program and Context Descriptions</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>Evaluations should document programs and their contexts with appropriate detail and scope for the evaluation purposes</td>
<td></td>
</tr>
<tr>
<td>A5</td>
<td>Information Management</td>
<td>-0.05</td>
</tr>
<tr>
<td></td>
<td>Evaluations should employ systematic information collection, review, verification, and storage methods</td>
<td></td>
</tr>
<tr>
<td>A6</td>
<td>Sound Designs and Analyses</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>Evaluations should employ technically adequate designs and analyses that are appropriate for the evaluation purposes</td>
<td></td>
</tr>
<tr>
<td>A7</td>
<td>Explicit Evaluation Reasoning</td>
<td>-0.01</td>
</tr>
<tr>
<td></td>
<td>Evaluation reasoning leading from information and analyses to findings, interpretations, conclusions, and judgments should be clearly and completely documented</td>
<td></td>
</tr>
<tr>
<td>A8</td>
<td>Communication and Reporting</td>
<td>N/A*</td>
</tr>
<tr>
<td></td>
<td>Evaluation communications should have adequate scope and guard against misconceptions, biases, distortions, and errors</td>
<td></td>
</tr>
</tbody>
</table>

HEAL = Hierarchy of Evidence and Appraisal of Limitations; MEV = Metaevaluation
*The correlation between HEAL scores and A8 accuracy standard ratings for the CCM studies could not be calculated since the A8 standard ratings were all Poor and assigned a score of zero (0).
Discussion of Results

An in-depth discussion follows centered on tables containing the checkpoints of each of the eight accuracy standards along with how the CCM studies fared against each checkpoint, plus a breakdown of CCM studies based on their metaevaluation ratings for each accuracy standard. Also, for each accuracy standard, criteria that is assessed with metaevaluation but not with grading systems like the HEAL grading system will be discussed.

**A1 Justified conclusions and decisions.** In Table 8, the CCM studies fared well when rated on the A1 accuracy standard. Of the 28 studies, 24 studies received a rating of Good or better. However, less than half of the studies were rated affirmatively according to items 4 and 6 of the A1 checklist. Regarding item 4, information about the contributions of other people to the conclusions of a research or evaluation study is not normally included in a written report. Regarding item 6, exploring and addressing plausible alternative explanations is more of a concern of an evaluator than a concern of a researcher.

Table 8  Summary of A1 Checkpoints for Selected CCM Studies

<table>
<thead>
<tr>
<th>A1 Checkpoints</th>
<th>Percent of CCM Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Address each contracted evaluation question based on information that is sufficiently broad, deep, reliable, contextually relevant, culturally sensitive, and valid</td>
<td>100%</td>
</tr>
<tr>
<td>2. Derive defensible conclusions that respond to the evaluation’s stated purposes, e.g., to identify and assess the program’s strengths and weaknesses, main effects and side effects, and worth and merit</td>
<td>93%</td>
</tr>
<tr>
<td>3. Limit conclusions to the applicable time periods, contexts, purposes, and activities</td>
<td>96%</td>
</tr>
<tr>
<td>4. Identify the persons who determined the evaluation’s conclusions, e.g., the evaluator using the obtained information plus inputs from a broad range of stakeholders</td>
<td>39%</td>
</tr>
<tr>
<td>5. Identify and report all important assumptions, the interpretive frameworks and values employed to derive the conclusions, and any appropriate caveats</td>
<td>82%</td>
</tr>
<tr>
<td>6. Report plausible alternative explanations of the findings and explain why rival explanations were rejected</td>
<td>14%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A1 Ratings</th>
<th>Breakdown of CCM Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>2</td>
</tr>
<tr>
<td>Very Good</td>
<td>10</td>
</tr>
<tr>
<td>Good</td>
<td>11</td>
</tr>
<tr>
<td>Fair</td>
<td>4</td>
</tr>
<tr>
<td>Poor</td>
<td>1</td>
</tr>
</tbody>
</table>
Considering just the checkpoints for the A1 accuracy standard, the following criteria would be overlooked if only a grading system such as the HEAL grading system would be used to assess study quality.

- Include information that is contextually relevant and culturally sensitive
- Derive conclusions that respond to a program’s worth and merit
- Identify and report all values employed to derive the conclusions
- Report plausible alternative explanations and explain why rival explanations were rejected

A2 Valid Information. In Table 9, the ratings for the CCM studies on the A2 accuracy standard are quite low. Only 6 CCM studies had a good rating on the A2 standard while none of the studies had a Very Good or Excellent rating. Less than 50% of the CCM studies rated affirmatively in four of the six items on the A2 checklist. Regarding item 1, communication with a full range of stakeholders is not possible to assess in a written report, plus research studies are generally not concerned with getting input from a full range of stakeholders. Regarding item 2, complete information about the training of data collectors is often not provided in written reports. Regarding item 3, data storage is usually not a concern in research studies. Regarding item 4, most research studies do not involve as many stakeholders as evaluation studies.
### Table 9: Summary of A2 Checkpoints for Selected CCM Studies

<table>
<thead>
<tr>
<th>A2 Checkpoints</th>
<th>Percent of CCM Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Through communication with the full range of stakeholders develop a coherent, widely understood set of concepts and terms needed to assess and judge the program within its cultural context</td>
<td>0%</td>
</tr>
<tr>
<td>2. Assure—through such means as systematic protocols, training, and calibration—that data collectors competently obtain the needed data</td>
<td>46%</td>
</tr>
<tr>
<td>3. Document the methodological steps taken to protect validity during data selection, collection, storage, and analysis</td>
<td>14%</td>
</tr>
<tr>
<td>4. Involve clients, sponsors, and other stakeholders sufficiently to ensure that the scope and depth of interpretations are aligned with their needs and widely understood</td>
<td>32%</td>
</tr>
<tr>
<td>5. Investigate and report threats to validity, e.g., by examining and reporting on the merits of alternative explanations</td>
<td>68%</td>
</tr>
<tr>
<td>6. Assess and report the comprehensiveness, quality, and clarity of the information provided by the procedures as a set in relation to the information needed to address the evaluation’s purposes and questions</td>
<td>100%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A2 Ratings</th>
<th>Breakdown of CCM Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>0</td>
</tr>
<tr>
<td>Very Good</td>
<td>0</td>
</tr>
<tr>
<td>Good</td>
<td>6</td>
</tr>
<tr>
<td>Fair</td>
<td>18</td>
</tr>
<tr>
<td>Poor</td>
<td>4</td>
</tr>
</tbody>
</table>

Considering just the checkpoints for the A2 accuracy standard, the following criteria would be overlooked if only a grading system such as the HEAL grading system would be used to assess study quality:

- Communicate with the full range of stakeholders to develop concepts and terms needed to assess and judge the program within its cultural context
- Document the methodological steps taken to protect validity during data storage
- Involve clients, sponsors, and other stakeholders sufficiently to ensure that the scope and depth of interpretations are aligned with their needs and widely understood

### A3 Reliable Information

In Table 10, the ratings for the CCM studies on the A3 accuracy standard are very low. None of the CCM studies had a rating of Good or higher on the A3 standard. Less than 10% of the CCM studies rated affirmatively in five of the six items on the A3 checklist. Regarding
items 1 through 4, neither written research reports nor written evaluation reports generally report reliability estimates, especially if the instruments are fairly well known to the report’s expected audience.

Table 10  Summary of A3 Checkpoints for Selected CCM Studies

<table>
<thead>
<tr>
<th>A3 Checkpoints</th>
<th>Percent of CCM Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Determine, justify, and report the needed types of reliability—e.g., test-retest, findings from parallel groups, or ratings by multiple observers—and the acceptable levels of reliability</td>
<td>7%</td>
</tr>
<tr>
<td>2. In the process of examining, strengthening, and reporting reliability, account for situations where assessments are or may be differentially reliable due to varying characteristics of persons and groups in the evaluation’s context</td>
<td>0%</td>
</tr>
<tr>
<td>3. Assure that the evaluation team includes or has access to expertise needed to investigate the applicable types of reliability</td>
<td>0%</td>
</tr>
<tr>
<td>4. Describe the procedures used to achieve consistency</td>
<td>7%</td>
</tr>
<tr>
<td>5. Provide appropriate reliability estimates for key information summaries, including descriptions of programs, program components, contexts, and outcomes</td>
<td>25%</td>
</tr>
<tr>
<td>6. Examine and discuss the consistency of scoring, categorization, and coding and between different sets of information, e.g., assessments by different observers</td>
<td>7%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A3 Ratings</th>
<th>Breakdown of CCM Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>0</td>
</tr>
<tr>
<td>Very Good</td>
<td>0</td>
</tr>
<tr>
<td>Good</td>
<td>0</td>
</tr>
<tr>
<td>Fair</td>
<td>5</td>
</tr>
<tr>
<td>Poor</td>
<td>23</td>
</tr>
</tbody>
</table>

Considering just the checkpoints for the A3 accuracy standard, the following criteria would be overlooked if only a grading system such as the HEAL grading system would be used to assess study quality:

- Account for situations where assessments are or may be differentially reliable due to varying characteristics of persons and groups in the evaluation’s context
- Provide appropriate reliability estimates for key information summaries, including descriptions of programs, program components, contexts, and outcomes
**A4 Explicit Program and Context Descriptions.** In Table 11, the ratings for the CCM studies on the A4 accuracy standard are high. Of the 28 CCM studies, 26 studies had a Good rating or higher on A4 standard. Greater than 50% of the CCM studies rated affirmatively in all items on the A4 checklist. Items 1 and 4 on the A4 checklist each have multiple concepts represented. As mentioned above, there is a bias towards affirmative ratings when rating items with multiple concepts or aspects of the program represented.

Table 11 Summary of A4 Checkpoints for Selected CCM Studies

<table>
<thead>
<tr>
<th>A4 Checkpoints</th>
<th>Percent of CCM Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Describe all important aspects of the program—e.g., goals, design, intended and actual recipients, components and subcomponents, staff and resources, procedures, and activities—and how these evolved over time</td>
<td>100%</td>
</tr>
<tr>
<td>2. Describe how people in the program’s general area experienced and perceived the program’s existence, importance, and quality</td>
<td>57%</td>
</tr>
<tr>
<td>3. Identify any model or theory that program staff invoked to structure and carry out the program</td>
<td>100%</td>
</tr>
<tr>
<td>4. Define, analyze, and characterize contextual influences that appeared to significantly influence the program and that might be of interest to potential adopters, including the context’s technical, social, political, organizational, and economic features</td>
<td>89%</td>
</tr>
<tr>
<td>5. Identify any other programs, projects, or factors in the context that may affect the evaluated program’s operations and accomplishments</td>
<td>86%</td>
</tr>
<tr>
<td>6. As appropriate, report how the program’s context is similar to or different from contexts where the program is expected to or reasonably might be adopted</td>
<td>75%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A4 Ratings</th>
<th>Breakdown of CCM Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>10</td>
</tr>
<tr>
<td>Very Good</td>
<td>12</td>
</tr>
<tr>
<td>Good</td>
<td>4</td>
</tr>
<tr>
<td>Fair</td>
<td>2</td>
</tr>
<tr>
<td>Poor</td>
<td>0</td>
</tr>
</tbody>
</table>
Considering just the checkpoints for the A4 accuracy standard, the following criteria would be overlooked if only a grading system like the HEAL grading system would be used to assess study quality:

- Describe how people in the program’s general area experienced and perceived the program’s existence, importance, and quality
- Define, analyze, and characterize contextual influences, including the context’s technical, social, political, organizational, and economic features
- As appropriate, report how the program’s context is similar to or different from contexts where the program is expected to or reasonably might be adopted

A5 Information Management. In Table 12, the ratings for the CCM studies on the A5 accuracy standard are low. None of the CCM studies had a Good rating or higher on the A5 standard. Less than 50% of the CCM studies rated affirmatively in all but the first item on the A5 checklist. Regarding items 2 through 6, written reports (whether evaluation reports or research reports) do not usually detail the management of information as prescribed in the items on the A5 checklist.

Table 12 Summary of A5 Checkpoints for Selected CCM Studies

<table>
<thead>
<tr>
<th>A5 Checkpoints</th>
<th>Percent of CCM Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Select information sources and procedures that are most likely to meet the evaluation’s needs for accuracy and be respected by the evaluation’s client group</td>
<td>100%</td>
</tr>
<tr>
<td>2. Ensure that the collection of information is systematic, replicable, adequately free of mistakes, and well documented</td>
<td>4%</td>
</tr>
<tr>
<td>3. Establish and implement protocols for quality control of the collection, validation, storage, and retrieval of evaluation information</td>
<td>11%</td>
</tr>
<tr>
<td>4. Document and maintain both the original and processed versions of obtained information</td>
<td>0%</td>
</tr>
<tr>
<td>5. Retain the original and analyzed forms of information as long as authorized users need it</td>
<td>0%</td>
</tr>
<tr>
<td>6. Store the evaluative information in ways that prevent direct and indirect alterations, distortions, destruction, or decay</td>
<td>0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A5 Ratings</th>
<th>Breakdown of CCM Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>0</td>
</tr>
<tr>
<td>Very Good</td>
<td>0</td>
</tr>
<tr>
<td>Good</td>
<td>0</td>
</tr>
<tr>
<td>Fair</td>
<td>4</td>
</tr>
<tr>
<td>Poor</td>
<td>24</td>
</tr>
</tbody>
</table>
Considering just the checkpoints for the A5 accuracy standard, the following criteria would be overlooked if only a grading system such as the HEAL grading system would be used to assess study quality.

- Select information sources and procedures that are most likely to be respected by the evaluation’s client group
- Document and maintain both the original and processed versions of obtained information
- Store the evaluative information in ways that prevent direct and indirect alterations, distortions, destruction, or decay

*A6 Sound Designs and Analyses.* In Table 13, the ratings for the CCM studies on the A6 accuracy standard are high. Only 4 CCM studies had a rating less than Good on the A6 standard. On only one item on the A6 checklist did less than 50% of the CCM studies rated affirmatively. Regarding this item, these are activities that are rarely detailed on written reports, either evaluation or research.
Table 13  Summary of A6 Checkpoints for Selected CCM Studies

<table>
<thead>
<tr>
<th>A6 Checkpoints</th>
<th>Percent of CCM Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Create or select a logical framework that provides a sound basis for studying the subject program, answering the evaluation’s questions, and judging the program and its components</td>
<td>68%</td>
</tr>
<tr>
<td>2. Plan to access pertinent information sources and to collect a sufficient breadth and depth of relevant, high quality quantitative and qualitative information in order to answer the evaluation’s questions and judge the program’s value</td>
<td>100%</td>
</tr>
<tr>
<td>3. Delineate the many specific details required to collect, analyze, and report the needed information</td>
<td>86%</td>
</tr>
<tr>
<td>4. Develop specific plans for analyzing obtained information, including clarifying needed assumptions, checking and correcting data and information, aggregating data, and checking for statistical significance of observed changes or differences in program recipients’ performance</td>
<td>100%</td>
</tr>
<tr>
<td>5. Buttress the conceptual framework and technical evaluation design with concrete plans for staffing, funding, scheduling, documenting, and metaevaluating the evaluation work</td>
<td>0%</td>
</tr>
<tr>
<td>6. Plan specific procedures to avert and check for threats to reaching defensible conclusions, including analysis of factors of contextual complexity, examination of the sufficiency and validity of obtained information, checking on the plausibility of assumptions underlying the evaluation design, and assessment of the plausibility of alternative interpretations and conclusions</td>
<td>82%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A6 Ratings</th>
<th>Breakdown of CCM Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>0</td>
</tr>
<tr>
<td>Very Good</td>
<td>16</td>
</tr>
<tr>
<td>Good</td>
<td>8</td>
</tr>
<tr>
<td>Fair</td>
<td>4</td>
</tr>
<tr>
<td>Poor</td>
<td>0</td>
</tr>
</tbody>
</table>

Considering just the checkpoints for the A6 accuracy standard, the following criteria would be overlooked if only a grading system such as the HEAL grading system would be used to assess study quality..

- Plan to access pertinent information sources and to collect a sufficient breadth and depth of relevant, high quality quantitative and qualitative information in order to answer the evaluation’s questions and judge the program’s value
- Analyze factors of contextual complexity and assess the plausibility of alternative interpretations and conclusions
**A7 Explicit Evaluation Reasoning.** In Table 14, the ratings for the CCM studies on the A7 accuracy standard are middle of the road. Most of the CCM studies are rated either Good or Fair on the A7 standard. Less than 50% of the CCM studies rated affirmatively in three of the items on the A7 checklist.

Regarding items 3 and 4, the assessment of stakeholder values is not a routine activity when conducting a research study. Regarding item 5, alternative conclusions may be mentioned in research reports but are not usually dealt with in depth.

Table 14 Summary of A7 Checkpoints for Selected CCM Studies

<table>
<thead>
<tr>
<th>A7 Explicit Evaluation Reasoning</th>
<th>Percent of CCM Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Clearly describe all the assumptions, criteria, and evidence that provided the basis for judgments and conclusions</td>
<td>86%</td>
</tr>
<tr>
<td>2. In making reasoning explicit, begin with the most important questions, then, as feasible, address all other key questions, e.g., those related to description, improvement, causal attributions, accountability, and costs related to effectiveness or benefits</td>
<td>100%</td>
</tr>
<tr>
<td>3. Document the evaluation’s chain of reasoning, including the values invoked so that stakeholders who might embrace different values can assess the evaluation’s judgments and conclusions</td>
<td>21%</td>
</tr>
<tr>
<td>4. Examine and report how the evaluation’s judgments and conclusions are or are not consistent with the possibly varying value orientations and positions of different stakeholders</td>
<td>11%</td>
</tr>
<tr>
<td>5. Identify, evaluate, and report the relative defensibility of alternative conclusions that might have been reached based on the obtained evidence</td>
<td>39%</td>
</tr>
<tr>
<td>6. Assess and acknowledge limitations of the reasoning that led to the evaluation’s judgments and conclusions</td>
<td>93%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A7 Ratings</th>
<th>Breakdown of CCM Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>1</td>
</tr>
<tr>
<td>Very Good</td>
<td>2</td>
</tr>
<tr>
<td>Good</td>
<td>13</td>
</tr>
<tr>
<td>Fair</td>
<td>10</td>
</tr>
<tr>
<td>Poor</td>
<td>2</td>
</tr>
</tbody>
</table>

Considering just the checkpoints for the A7 accuracy standard, the following criteria would be overlooked if only a grading system like the HEAL grading system would be used to assess study quality.

- In making reasoning explicit, also address costs related to effectiveness or benefits
- Document the evaluation’s chain of reasoning, including the values invoked so that stakeholders who might embrace different values can assess the evaluation’s judgments and conclusions
• Examine and report how the evaluation’s judgments and conclusions are or are not consistent with the possibly varying value orientations and positions of different stakeholders

A8 Communication and Reporting. In Table 15, the ratings for the CCM studies on the A8 accuracy standard are very low. All of the CCM studies were rated Poor on the A8 standard and less than 50% of the CCM studies rated affirmatively in all of the items on the A8 checklist. Regarding all these items, these are activities that rarely are included in research or evaluation reports. Also, research studies do not usually require multiple reporting mechanisms. Considering this accuracy standard, there appears to be no advantage over existing grading systems when the primary focus is assessing study quality so checkpoints for the A8 accuracy standard that would be overlooked are not listed here.

Table 15 Summary of A8 Checkpoints for Selected CCM Studies

<table>
<thead>
<tr>
<th>A8 Checkpoints</th>
<th>Percent of CCM Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Reach a formal agreement that the evaluator will retain editorial authority over reports</td>
<td>0%</td>
</tr>
<tr>
<td>2. Reach a formal agreement defining right-to-know audiences and guaranteeing appropriate levels of openness and transparency in releasing and disseminating evaluation findings</td>
<td>0%</td>
</tr>
<tr>
<td>3. Schedule formal and informal reporting in consideration of user needs, including follow-up assistance for applying findings</td>
<td>0%</td>
</tr>
<tr>
<td>4. Employ multiple reporting mechanisms, e.g., slides, dramatizations, photographs, PowerPoint®, focus groups, printed reports, oral presentations, telephone conversations, and memos</td>
<td>11%</td>
</tr>
<tr>
<td>5. Provide safeguards, such as stakeholder reviews of draft reports and translations into language of users, to assure that formal evaluation reports are correct, relevant, and understood by representatives of all segments of the evaluation’s audience</td>
<td>0%</td>
</tr>
<tr>
<td>6. Consistently check and correct draft reports to assure they are impartial, objective, free from bias, responsive to contracted evaluation questions, accurate, free of ambiguity, understood by key stakeholders, and edited for clarity</td>
<td>4%</td>
</tr>
</tbody>
</table>

A8 Ratings | Breakdown of CCM Studies
--- | ---
Excellent | 0
Very Good | 0
Good | 0
Fair | 0
Poor | 28
Complementary approaches? In Table 16, there is a comparison between the breakdown of CCM studies based on accuracy metaevaluation ratings and the breakdown of CCM studies based on HEAL system grades. It is of interest that the CCM studies mostly fared poor to fair (accuracy metaevaluation ratings) or mostly a D grade (HEAL grading system). Observing this similarity in study quality outcome may lead the conclusion that the two methods for assessing study quality may be correlated somehow. But, as evident in Table 4.2 above, there is no significant relationship between the two methods at any level.

Table 16 Comparisons of Metaevaluation Ratings and HEAL Grades for Selected CCM Studies

<table>
<thead>
<tr>
<th>Accuracy Metaevaluation Ratings</th>
<th>Breakdown of CCM Studies</th>
<th>HEAL system Grades</th>
<th>Breakdown of CCM Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>0</td>
<td>A to A-</td>
<td>4</td>
</tr>
<tr>
<td>Very Good</td>
<td>0</td>
<td>B to B-</td>
<td>4</td>
</tr>
<tr>
<td>Good</td>
<td>5</td>
<td>C to C-</td>
<td>2</td>
</tr>
<tr>
<td>Fair</td>
<td>20</td>
<td>D</td>
<td>18</td>
</tr>
<tr>
<td>Poor</td>
<td>3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

One possibility is that the two instruments may be complementary rather than substitutionary. That is, perhaps it is better to use both the accuracy metaevaluation ratings and the HEAL grades to determine study quality rather than using one instrument over the other, or one instrument instead of the other. Certainly, there are many criteria listed above that are addressed by metaevaluation checkpoints but are not addressed by existing grading systems. If this is so, then metaevaluation may indeed be a useful adjunct to a grading system that encourages and holds accountable healthcare studies to the highest possible quality.
CHAPTER V

CONCLUSION

Summary of Dissertation Findings

The purpose of this dissertation was to explore the use of metaevaluation to assess healthcare study quality. The study procedure consisted of evaluating selected chronic care model (CCM) studies using the Program Evaluations Metaevaluation Checklist (Stufflebeam, 2011b) and processing the results to answer two questions: (1) what is the assessed quality of the selected studies by metaevaluation using the accuracy standards of program evaluation (Yarbrough et al., 2011), and (2) is there a relationship between study quality assessment by metaevaluation and study quality assessment by the HEAL grading system (Gugiu, et al, 2013)?

The answer to the first question, according to the dissertation results, is that the assessed quality of the selected studies by metaevaluation was mostly low. There were no studies that are rated Excellent or Very Good, only five (5) studies that are rated Good, and the rest of the studies were rated either Fair (20) or Poor (3) (see Table 4.1).

The answer to the second question, according to the dissertation results, is that there does not seem to be a relationship between the study quality of the selected studies as rated by metaevaluation and the study quality of the same selected studies as graded by the HEAL grading system. Neither the overall accuracy standard rating nor any of the individual accuracy standard ratings had a significant correlation with the HEAL grades for the selected CCM studies (see Table 4.2).

As stated in Chapter 4, it may be possible that the two instruments are complementary rather than substitutionary. If this is so, then metaevaluation may indeed be a useful adjunct to a grading system that encourages and holds accountable healthcare studies to the highest possible quality.
Limitations to Dissertation Findings

There are several limitations to the findings in this dissertation. The first limitation is that the selected CCM studies are research studies. When a research study is evaluated using a metaevaluation checklist, the study will not rate highly because research studies do not gather and analyze as much information as evaluations, especially information about the context of the study and about the multiple perspectives of stakeholders.

A second limitation is that even evaluation reports do not have enough information to conduct a metaevaluation. According to Stufflebeam (2001), the metaevaluator often must collect additional needed information other than the information gathered from the evaluation report. To reach valid conclusions, metaevaluators need to be able to access all available information and collect any further information if needed (Stufflebeam, 2001). However, since these studies are older studies, it is not feasible to gather any more information than what is available in the reports on the studies.

A third limitation is that several checkpoints in the metaevaluation checklist contain multiple concepts. For example, the first item in the A1 checklist includes validity, reliability, comprehensiveness, and context sensitivity. These are all very different concepts. When rating a particular study with an item containing multiple concepts, the rater needs to determine whether or not to respond affirmatively if at least one aspect but not all aspects are true about the study. Decision rules were created and decided upon to minimize confusion in this regard (see Chapter 3).

Implications for Implementation Science

The importance of implementation science is highlighted and magnified by the burden of chronic illness and the inadequacy of current healthcare delivery system to address it. Chronic illnesses are the leading cause of morbidity and mortality in this country, affecting almost half of the national population. The majority of healthcare resources, as much as 70 percent of all healthcare expenditures, are now devoted to the treatment of chronic illness (DeVol & Bedroussian, 2007).

Yet, the healthcare system is structured primarily for providing acute care. Unlike acute care, chronic care is a collaborative process which both the practitioner and the patient must understand. The collaboration involved in caring for the chronically ill requires that the healthcare system be more
centralized. To a great extent, patients are being passed from practitioner to practitioner without a medical home to coordinate the process and without the formation of vital therapeutic relationships between patient and practitioner (CQHCA, 2001). Consequently, almost half of all patients with chronic illness are not receiving evidence-based care (McGlynn et al., 2003).

This chronic illness burden/broken healthcare system predicament led to an increased resolve toward the end of the 1990s to improve healthcare quality. The definition of healthcare quality, developed by the Institute of Medicine (IOM), is “the degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge” (Lohr, 1990, p. 10). In 1998, a watershed year for the assessment of healthcare quality, an IOM roundtable concluded that quality of care can be measured with sufficient precision to make judgments and take needed actions to bring about improvement (Chassin & Galvin, 1998). That same year, however, an Advisory Commission conducted a review of academic literature on the quality of healthcare in the United States and concluded that healthcare quality in the United States was not being systematically measured or monitored (Schuster et al., 1998).

Finally, in 2001, the CQHCA’s report “Crossing the Quality Chasm,” concluded that quality problems are everywhere in the country, affecting many patients. These three reports created a cognitive dissonance in the healthcare community which fueled the demand for greater accountability in the dissemination, implementation, and diffusion of healthcare, thus renewing an interest in implementation science.

Since implementation science (IS) seeks to “understand the behavior of healthcare professionals and other stakeholders as a key variable in the sustainable uptake, adoption, and implementation of evidence-based interventions” (Madon et al., 2007, p. 1728), IS may be instrumental in not only measuring and monitoring healthcare quality, but improving it as well. As noted earlier, Grimshaw and Eccles (2004) found that establishing an empirically tested theoretical base for healthcare professional and organizational behavior would facilitate more effective interventions. When the translation of research into clinical practice is conceptualized as proceeding from awareness to acceptance to adoption,
it is clear that a necessary step in the improvement of healthcare quality requires a detailed understanding of how adoption takes place (Green & Seifert, 2005).

Implementation science requires a systematic assessment of the effect of political, professional, economic, social, and organizational factors in the adoption of new healthcare practices, plus the consideration of multiple stakeholders with multiple perspectives (Madon et al., 2007). Understanding this complexity requires more methodology than what efficacy and effectiveness research have to offer. Evaluation methods are needed to integrate and synthesize contextual and external validity evidences that can aid both local decision makers and policy-making bodies. If evaluators work in partnership with relevant decision makers and target audiences, it will be easier for practitioners and policy makers to judge program relevance (Glasgow & Emmons, 2007).

Contribution to Evaluation Practice

As stated before, the purpose of this dissertation is to explore the use of metaevaluation to assess the quality of healthcare studies. Judging from the results of the dissertation, it seems that accuracy metaevaluation ratings can play a very useful role in assessing the quality of healthcare studies.

In the CQHCA's (2001) report “Crossing the Quality Chasm,” two of the four underlying reasons that were cited for the healthcare quality chasm in the United States were the increase in chronic conditions and a poorly organized healthcare delivery system. Assessing the dissemination, implementation, and diffusion of evidence-based medicine that should be occurring at the interface between these two nationwide healthcare dilemmas is the business of implementation science. And because implementation science demands multiple methodologies to assess not only process and outcome but also multiple contexts from the perspective of multiple stakeholders, this has major implications for evaluation practice.

When comparing the accuracy metaevaluation checklist with the HEAL grading system, it is clear that evaluations are intended to assess much more than a quantitative research study. In Chapter 4, many checkpoints taken from the metaevaluation checklist addressed criteria not normally addressed by research studies. Yet, many of these criteria are well-suited to the multifaceted demands of implementation science.
In Chapter 2 of this dissertation, four implementation science frameworks were mentioned that have implications for the use of metaevaluation: RE-AIM, QUERI, PARiHS, and CFIR. The RE-AIM (reach, effectiveness, adoption, implementation, and maintenance) model, developed by Glasgow et al. (1999), is intended to aid the planning, conduct, evaluation, and reporting of studies that are designed to translate research into practice. The Department of Veterans Affairs Quality Enhancement Research Initiative (QUERI) integrates formative evaluation into its implementation program (Stetler et al., 2006). QUERI, begun in 1998, is an improvement initiative that focuses on implementation of empirically based practices and on the evaluation and refinement of these implementation efforts. In 2004, the Royal College of Nursing (RCN) Institute published a PARiHS framework (Promoting Action on Research Implementation in Health Services) to facilitate implementation science (Kitson et al., 2008). This framework uses a two-stage process whereby evaluation is used to measure activities and outcomes and context, and then the results of the evaluation is used to determine the most appropriate facilitation method for the program. The purpose of the PARiHS framework is to provide a map to enable others to make sense of the complexity of implementation, and the elements that require attention if implementation is more likely to be successful (Kitson et al., 2008). Finally, in 2009, a research group from the University of Michigan offered the Consolidated Framework For Implementation Research (CFIR) to promote implementation theory development and verification about what works where and why across multiple contexts. The CFIR comprises five major domains that support the exploration of essential factors which may be encountered during implementation through formative evaluations (Damschroder et al., 2009). All four of these frameworks include formative evaluation as an element of their approaches to implementation science. All four of these frameworks would therefore benefit from the use of accuracy metaevaluation ratings as a means for assessing healthcare quality within the frameworks. It seems likely, then, that a metaevaluation checklist using accuracy standards could play an important role in closing the quality chasm of healthcare, including the healthcare of patients with chronic illness.

Given the urgency of the CQHCA’s (2001) report “Crossing the Quality Chasm,” evaluators of healthcare programs ought to place a priority on evaluating the dissemination, implementation, and diffusion of CCM programs. Of particular concern is the prevalence of the health insurance-based disease
management model over the physician-based chronic care model, especially if it is true that neither option is more cost effective than the other. The distinguishing factor between the two models would then be the quality of care received by the patient and the outcomes of that care. Distinguishing between these models for chronic illness care would be a role ideally suited for evaluators, and perhaps a metaevaluation rating tool would a viable approach to assessing quality differences between the two models.

Need for Further Study

There are several directions that research on evaluation can take from this dissertation. One direction is to conduct more studies where healthcare studies are regarded as healthcare program evaluations and are assessed accordingly using accuracy metaevaluation ratings. Completed healthcare studies can be summatively metaevaluated and ongoing healthcare studies can be formatively metaevaluated. A metaevaluation approach to healthcare studies, especially those studies that evaluate the dissemination, implementation, and diffusion of healthcare programs, will accordingly be held to higher standards for healthcare study quality, thereby increasing healthcare quality itself.

A second direction that research on evaluation can take from this dissertation is to investigate the advantages of combining existing grading systems such as the GRADE system or the HEAL system with accuracy metaevaluation ratings to increase the ability to detect differences in study quality. Perhaps adding accuracy metaevaluation ratings to the study quality mix would only add more noise to the study quality conversation, or perhaps its addition may actually effect an increase not only in healthcare study quality, but an increase in healthcare quality as well. This research on evaluation direction would likely contribute significantly to the advancement of implementation science.

A third direction that research on evaluation can take from this dissertation is to explore modifications of existing metaevaluation checklists, including the Program Evaluations Metaevaluation Checklist, to improve the application, reliability, and validity of the checklists. As stated before, reliability is a serious concern when applying the Program Evaluations Metaevaluation Checklist (Wingate, 2009). Because the reliability of the Checklist affects the validity of the Checklist, it may not adequately assess
evaluation study quality. Modifications to metaevaluation checklists may include changes to the design (e.g. checkpoints) of the checklist or changes to the execution (e.g. rater management) of the checklist.

Since there was somewhat greater reliability to the Checklist when rating according to the accuracy standards (Wingate, 2009), limiting metaevaluation to accuracy standards will provide some increase in reliability. Eliminating checkpoints with multiple criteria will also improve reliability of metaevaluation checklists. Checkpoints with multiple criteria necessitate more detailed decision rules which increase the likelihood that those rules would not be followed perfectly. Finally, the scoring of ratings could be improved by not entering the selection of “N/A” or “???” into the calculation of the overall score. For example, calculation of the mean should exclude these choices from the denominator of the equation to prevent a neutral response from becoming a negative response.

As stated before, there are also changes in the execution of a metaevaluation checklist that can be taken to increase the reliability of the checklist, including (1) increasing documentation and transparency in the process, (2) creating and agreeing on decision rules for determining the extent to which a standard was met, (3) using more than one rater, and (4) calibrating the raters to increase inter-rater reliability (Wingate, 2009). Research on evaluation can be used to compare different checklist designs and execution procedures to increase the reliability and validity of the checklist approach to metaevaluation.

If these recommendations for further study are heeded, research on evaluation may contribute to the sustainability of the evidence-based movement in healthcare and the advancement of implementation science in chronic illness care. More specifically, metaevaluation may become one of the best means of attaining and sustaining high-quality healthcare studies.


INTRODUCTION


REFERENCES


# APPENDIX A

## HIERARCHY OF EVIDENCE AND APPRAISAL OF LIMITATIONS (HEAL) GRADING SYSTEM

<table>
<thead>
<tr>
<th>Study Design</th>
<th>Grade</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomized control trial (RCT)</td>
<td>A</td>
<td>A well-conducted RCT free of bias or confounding factors in which the randomization process is able to ensure the equivalence of the treatment and control groups beyond acceptable statistical chance</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>An RCT with a flawed randomization process but which properly employed an equivalent controlled trial (ECT) method for ensuring the equivalence of the treatment and control groups beyond a reasonable doubt</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>An RCT with a flawed randomization process, which cannot ensure the equivalence of the treatment and control groups beyond a reasonable doubt. Alternatively, a cohort RCT that failed to show the population from which the samples were drawn remained stable over time</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>An RCT with one or more of the following fatal flaws: The treatment group received a very low dosage or poor treatment fidelity; significant crossover existed between treatment and control subjects; the follow-up period was insufficient to detect change; the follow-up period (either timing or length of time) was significantly different between the treatment and control arms; the sample size was significantly lower than the recommendations from an a priori power analysis; a significant selection bias or differential attrition rates occurred between the treatment and control groups that produced baseline group differences; or significant contamination of baseline outcome measures occurred</td>
</tr>
<tr>
<td>Pseudo-RCT, matched comparison, stratified random sampling, paired comparison, regression discontinuity</td>
<td>B</td>
<td>A well-conducted ECT that demonstrated statistical equivalence of the treatment and control groups on key baseline variables or the pre-post regression mode</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>An ECT with demonstrated baseline differences on key variables or that did not employ an adequate number of covariates, pairing, or strata variables to sufficiently remove reasonable doubt regarding the statistical equivalence of the treatment and control groups on key baseline variables or a regression discontinuity design that failed to demonstrate the equivalence of the pre–post regression model. Alternatively, a cohort ECT that did not demonstrate the population from which the samples were drawn remained stable over time</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>An ECT with one or more of the fatal flaws listed under RCT Grade D (continued)</td>
</tr>
<tr>
<td>Cohort study, case–control</td>
<td>C</td>
<td>A controlled study that did not adequately establish the equivalence of the treatment and control groups beyond a reasonable doubt (i.e., a nonequivalent controlled trial [NECT])</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>An NECT with one or more of the fatal flaws listed under RCT Grade D</td>
</tr>
<tr>
<td>Before–after, case series</td>
<td>D</td>
<td>Any study that did not employ a controlled comparison between two or more groups (i.e., an uncontrolled trial [UT]), including RCT, ECT, and NECT studies that did not employ a true comparison group (i.e., compared two treatments against each other but not against a treatment as usual group)</td>
</tr>
<tr>
<td>Type of Limitation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Serious limitations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Controlled study without a true control group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Low dosage/poor treatment fidelity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Treatment crossover/no intent-to-treat</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Insufficient follow-up period</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Differential follow-up period</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Low statistical power</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Selection bias</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Differential attrition between groups</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Contaminated baseline measure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Unstable cohort sample</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Design-Specific Limitations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Biased allocation in an RCT (e.g., poor blinding or allocation concealment)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Flawed randomization in a RCT (e.g., poor sequence generation)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Insufficient number of matching variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- No tests for statistical equivalence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Non-equivalence or evidence of significant difference/differences</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minor limitations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Lengthy data collection period</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Constantly changing intervention</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Comparison groups are not mutually exclusive</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Used post-baseline covariates in analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Potential crossover effect</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Poor generalizability of results (low participation or high attrition)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Adapted from Gugiu & Gugiu, 2010
Date: November 11, 2013

To: Chris Coryn, Principal Investigator
    Jan Fields, Student Investigator for dissertation

From: Amy Naugle, Ph.D., Chair

Re: Approval not needed for HSIRB Project Number 13-11-11

This letter will serve as confirmation that your project titled “A Metaevaluation of Evaluations of Health Care Programs” has been reviewed by the Human Subjects Institutional Review Board (HSIRB). Based on that review, the HSIRB has determined that approval is not required for you to conduct this project because you are analyzing programs and not collecting personal identifiable (private) information about individuals.

Thank you for your concerns about protecting the rights and welfare of human subjects.

A copy of your protocol and a copy of this letter will be maintained in the HSIRB files.